FairShap: A Fairness Framework Based Explainable Machine Learning

Xikuan Wang[1], Min Zhang[2], and Jie Li

[1] East China Normal University,
wang117687@gmail.com
[2] East China Normal University,
mzhang@sei.ecnu.edu.cn

Abstract. With the increasing application of machine learning in real-world decision-making systems, the fairness and interpretability of tasks involving humans have not yet been fully guaranteed. In order to solve the above problems, we proposes an interpretable fairness framework based on feature contributions, which aims to improve the degree of interpretability of fairness in binary classification tasks. First, the fairness contribution is explained by the importance of interpretable features and quantified by the Shapley value in Game Theory; then, groups are divided according to different protective attributes, and discrimination detection and debiasing algorithms are applied to specific groups to mitigate the bias in the original samples. The experimental results show that the proposed method significantly outperforms the existing methods in terms of interpretability and demonstrates wide applicability to different classifiers and fairness metrics.

Keywords: Machine Learning, Trustworthy AI, Fairness, Interpretability

1 Introduction

Machine Learning(ML) is increasingly used across various real-world decision-making systems, including autonomous driving[1], financial risk assessment[2], traffic control[3], medical diagnostics[4], resume recruitment[5], credit assessment[6], and criminal justice[7]. However, machine learning are susceptible to biases, which can lead to potentially discriminatory outcomes. While machine learning are fundamentally data-driven, the data itself may embody statistical or societal biases.[8] Such biases can be exacerbated during the machine learning process, particularly when sensitive attributes such as gender and race are involved, resulting in significant adverse effects on vulnerable populations.

In this paper, we propose FairShap, which aims to improve the fairness and interpretability of machine learning models. For the classification task, an interpretable fairness method is proposed, which combines group classification,

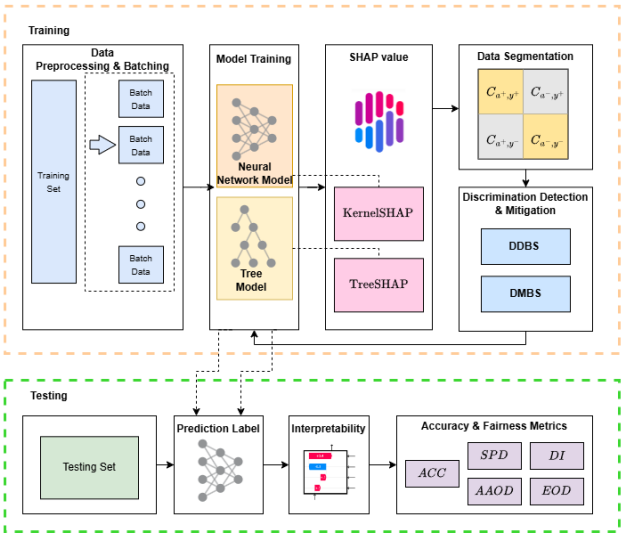


Fig. 1. FairShap Framework

discrimination detection and mitigation to process the dataset, and then combines the interpretable methods in the test and evaluation part to achieve the balance between classification fairness and accuracy.

Since there are many ways to quantify contributions in the interpretable domain[9], a model-independent and interpretable method for quantitatively calculating contributions is needed. In this paper, we will choose the Shapley value[10] to quantify contributions, which ensures that contributions can be calculated more efficiently in the following process, and achieves a balance between fairness and accuracy. While previous studies usually classify groups based on protective attributes[11], this paper uses a more detailed classification method, i.e., group classification based on the combination of protective attributes and labels, which can more accurately generate a candidate set and provide a basis for subsequent discrimination detection.

Traditional methods often use clustering techniques to recognize the fairness intuition that "similar individuals should be treated similarly"[12]. On this basis, this paper introduces the extended assumption that "similar individuals should have similar contributions", where contributions have been quantified by Shapley values. By calculating the Shapley value and filtering the samples, it is possible to identify the samples with mismatched contributions in the same group, and then label these samples as discriminatory or preferential samples.

Since the model is targeting discriminatory samples, this paper proposes an innovative discrimination mitigation algorithm. The algorithm is based on the perturbation of the protected attributes and the Shapley value for the identified

discriminated or favored samples to alleviate the imbalance problem within the dataset.

We perform experiments on three datasets for the protection attribute and compared with two other fairness methods. From the perspective of fairness and interpretability, we explore the potential problems of the model, the correctness and fairness of the model, and the degree of interpretability of the visual analytic model. Then, we interpret the experimental effects at the dataset level and analyze the effects of the experimental results on different models and different protection attributes. Finally, the results of the experiments and the framework are analyzed and discussed in the conclusion.

2 Definition

2.1 Problem Formulation

We consider general binary classification problems. The training dataset is denoted as (X, A, Y), where X represents a sample set that excludes sensitive attributes. The variable A denotes a predefined protective attribute, such as race, gender and age, or marital status, while Y denotes the corresponding label. In this study, we assume that both A and Y are binary variables.

Specifically, we denote a complete sample, inclusive of labels, as (x_i, a_i, y_i), where $x_i = (x_i^1, x_i^2, ..., x_i^n) \in Rn$. The sample labels are categorized into two classes: $y_i \in \{y^+, y^-\}$, where y^+ signifies the positive class and y^- signifies the negative class. The protective attributes are similarly classified as $a_i \in \{a^+, a^-\}$, with a^+ representing advantageous protective attributes and a^- representing disadvantageous protective attributes.

2.2 Fairness Metrics

Four principal guidelines define fairness within the context of machine learning: Statistical Parity Difference, Equal Opportunity Differences, Demographic Parity Difference, and Average Absolute Odds Difference.[13,14,15]

Statistical Parity Difference (SPD) A classifier h trained on a data distribution (X, A, Y) satisfies Statistical Parity Difference if the prediction $\hat{Y} = h(X)$ independent of the protective attribute A. The Statistical Parity Difference can be articulated as follows:

$$SPD = Pr\{\hat{Y} = y^+ | A = a^-\} - Pr\{\hat{Y} = y^+ | A = a^+\} \leq \tau$$

Here, τ serves as the threshold for the fairness constraint. Smaller values of SPD indicate fairer models.

Disparate Impact (DI) Also based on statistical parity, defined as follows:

$$DI = \frac{Pr\{\hat{Y} = y^+|A = a^-\}}{Pr\{\hat{Y} = y^+|A = a^+\}} \leq \tau$$

The closer the value of DI is to 1, the fairer the model is. The difference between DI and SPD is that SPD focuses more on absolute differences, and is suitable for scenarios where there is a need to look at the equality of outcomes between different groups. DI focuses on proportional comparisons and is used in scenarios where the degree of relative fairness is important, especially when there are large differences in group bases.

Equal Opportunity Differences (EOD) Based on equal opportunity, outputs from different groups are required to have the same True Positive Rate(TPR), defined as follows:

$$EOD = Pr\{\hat{Y} = y^+|A = a^-, Y = y^+\} - Pr\{\hat{Y} = y^+|A = a^+, Y = y^+\} \leq \tau$$

This metric is assessed by calculating the predicted difference in positive class samples between different groups, with smaller values of EOD indicating a fairer model.

Average Absolute Odds Difference (AAOD) Both True Positive Rate(TPR) and False Positive Rate(FPR) are considered, defined as follows:

$$AAOD = \frac{1}{2}(|FPR_{a^+} - FPR_{a^-}| + |TPR_{a^+} - TPR_{a^-}|) \leq \tau$$

2.3 Interpretability Assessment

The main focus of the interpretability assessment is the change of the model on the protection attribute A. By analyzing the contribution of the protection attribute A before and after the optimization of the model, the decision-making process of the model can be deeply understood.

Global Interpretability Assessment[16] is designed to provide a holistic view of the contribution of model features to the output. By visualizing the overall contribution of each feature to the model output, special attention is given to the changes in the protective attribute A before and after model optimization. This approach helps to fully understand the decision-making mechanism of the model under the influence of different features.

Comparative Interpretability Assessment[17] is the evaluation of the protective attribute A between different debiasing algorithms. It analyzes the effectiveness of different methods by comparing the distribution of protective attributes of each model. This evaluation method can not only reveal the differences in the processing of protective attributes among different algorithms, but also provide empirical evidence for the selection of the optimal debiasing strategy.

SHAP In cooperative Game Theory, Shapley values are used to distribute the benefits of cooperative outcomes. Based on this theory, the Shapley value is widely used in the field of machine learning to explain the prediction results of models, especially in interpretability assessment. The core idea is to calculate the overall contribution of each feature by considering its marginal contribution in all possible combinations.

Consider a model f that receives M features as inputs, and suppose we want to interpret the output of the model f with input x. Since there is no origin for the scale of the model output, we can only explain the difference between the observed model output and the chosen origin, which can be either the output value of a function of some arbitrary record or the average output value over a set of records D, and if the latter is used, the computation is expressed as follows:

$$\sum_{i=1}^{M} \phi_i = f(x) - \mathbb{E}_{y \sim \mathcal{D}} \left[f(y) \right]$$

Define F to be the set having M attributes, i.e., $F = \{F_1, ..., F_M\}$, and define S to be an arbitrary subset of $F \backslash \{F_i\}$, i.e. $S \subseteq F \backslash \{F_i\}$, then the Shapley value ϕ_i can be calculated according to the following equation:

$$\phi_i = \frac{1}{M} \sum_{S \subseteq F \backslash \{F_i\}} \frac{1}{\binom{M-1}{|S|}} \left(f(S \cup \{F_i\}) - f(S) \right)$$

where $f(S)$ are computed by treating S as a missing input.Thus the process of computing SHAP interpretations can be viewed as starting with S that does not contain F_i, adding F_i, and then observing the differences in the output values. For nonlinear functions, the values obtained will depend on which features are already in S, so we are interested in selecting a subset of the set size $|S|$ and then averaging the contributions of all subset sizes.

3 Dataset

ADULT dataset. The ADULT dataset is a publicly available dataset that is widely used in classification and machine learning research. It was originally derived from the 1994 U.S. Census and is used to predict whether an individual's annual income exceeds $50,000. The dataset has 14 characteristics, including gender, marital status, and nationality, with the protected attributes being gender and race. This dataset suffers from data imbalance.

COMPAS dataset. The COMPAS dataset is a public dataset used for risk assessment and is widely used to study fairness and bias in machine learning models. It is derived from defendant records in Broward County, Florida, USA, and is used to predict whether a defendant will reoffend within two years. The dataset has 12 characteristics, including the number of prior offenses, race, and age, with the protected attributes being gender and race.

DEFAULT dataset. The DEFAULT dataset is a classic dataset used to study credit risk and default prediction, and is often used in finance and credit rating

to predict whether a customer will default on a loan at some point in the future. The dataset has 25 characteristics, including age, gender, and income, with the protected attribute being gender.

4 Methodology

4.1 Candidate Set Generation

Algorithm 1 Candidate Set Generation Algorithm

Input: Dataset $\mathcal{D}(X, Y)$, Protected attributes A, Model M, Interpreter E
Output: Shapley value Φ_X, Candidate Set $\mathcal{C}_{\text{candi}}$

$\Phi_X \leftarrow \emptyset$ ▷ Init Φ_X Set
if Model M is Tree-Model **then**
  $\Phi_X \leftarrow E(\text{TreeSHAP}, M)$ ▷ Using TreeSHAP Model
else
  $\Gamma \leftarrow \text{Sample}(\min(\lambda|\mathcal{D}|, \nu))$ ▷ Using shap lib samples as background data
  $\Phi_X \leftarrow E(\text{KernelSHAP}, M, \Gamma)$ ▷ Using KernelSHAP Model
end if
$\mathcal{C}_{a^+,y^+} \leftarrow \emptyset$ ▷ Init $\mathcal{C}_{a^+,y^+}$ Set
$\mathcal{C}_{a^-,y^-} \leftarrow \emptyset$ ▷ Init $\mathcal{C}_{a^-,y^-}$ Set
for all $(x, y) \in \mathcal{D}(X, Y)$ **do**
  if $X[A] = a^+$ and $y = y^+$ **then**
    $\mathcal{C}_{a^+,y^+} \leftarrow \mathcal{C}_{a^+,y^+} \cup \{(x, y)\}$ ▷ For the dominant group, add to $\mathcal{C}_{a^+,y^+}$
  end if
  if $X[A] = a^-$ and $y = y^-$ **then**
    $\mathcal{C}_{a^-,y^-} \leftarrow \mathcal{C}_{a^-,y^-} \cup \{(x, y)\}$ ▷ For protected groups, add to $\mathcal{C}_{a^-,y^-}$
  end if
end for
$\mathcal{C}_{\text{candi}} \leftarrow \mathcal{C}_{a^+,y^+} \cup \mathcal{C}_{a^-,y^-}$
return Φ_X, $\mathcal{C}_{\text{candi}}$

The pseudo-code for the computation of the Shapley values and the population segmentation of the data is shown in Algorithm 1. The algorithm describes the specific steps of FairShap candidate set generation. The inputs include the dataset $D(X, Y)$, the protected attribute A, the model M, and the interpreter E; the outputs are the Shapley value Φ_X and the candidate set $\mathcal{C}_{\text{candi}}$.

Specific steps are as follows: 1. Initialize the candidate set: First, initialize the set of Shapley values Φ_X. 2. Select the interpreter: Use different interpretation methods according to the model type. If the model is a tree model, use TreeSHAP for interpretation; Otherwise, use KernelSHAP. Construct Candidate Sets: Iterate through each sample (x, y) in the dataset, and based on the values and labels of the protected attributes, add the eligible samples to the dominant group candidate set $\mathcal{C}_{a^+,y^+}$ or to the protected group candidate set $\mathcal{C}_{a^-,y^-}$ 4. Return results: Finally, the computed Shapley value set Φ_X and the candidate set $\mathcal{C}_{\text{candi}}$ are returned.

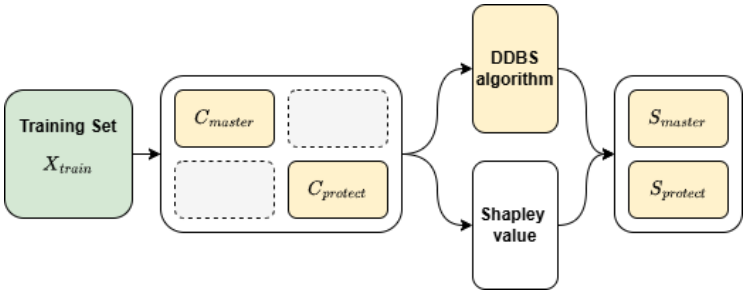


Fig. 2. The process of group segmentation to generate candidate and target sets

4.2 Discrimination Detection Based Shapley Value Algorithm

Algorithm 2 DDBS Algorithm

Input: Shapley Value Φ_X, Candidate Set $\mathcal{C}_{\text{candi}}$, Thresholds τ
Output: Dominant group set $\mathcal{S}_{\text{master}}$, Protected group set $\mathcal{S}_{\text{protect}}$

$\mathcal{S}_{\text{master}} \leftarrow \emptyset$ ▷ Init $\mathcal{S}_{\text{master}}$
$\mathcal{S}_{\text{protect}} \leftarrow \emptyset$ ▷ Init $\mathcal{S}_{\text{protect}}$
for $(x, y) \in C_{candi}$ **and** $i \in |C_{candi}|$ **do**
  $\varphi \leftarrow \Phi_{X[i]}$
  if $y = y^+$ **and** $\varphi \geq \tau$ **then**
    $\mathcal{S}_{\text{master}} \leftarrow \mathcal{S}_{\text{master}} \cup (x, y)$
  end if
  if $y = y^-$ **and** $\varphi \leq -\tau$ **then**
    $\mathcal{S}_{\text{protect}} \leftarrow \mathcal{S}_{\text{protect}} \cup (x, y)$
  end if
end for
return $\mathcal{S}_{\text{master}}, \mathcal{S}_{\text{protect}}$ ▷ Return group set

The pseudo-code for the Discrimination Detection Based Shapley Value Algorithm is shown in Algorithm 2. The algorithm describes the specific implementation process of DDBS. Based on the input Shapley values and candidate sets, it identifies and adjusts the performance of the model in different groups to achieve fairness.

Inputs include Shapley value Φ_X, which is used to evaluate the contribution of each feature to the model prediction; Candidate set $\mathcal{C}_{\text{candi}}$, a set that contains the sample features and labels to be analyzed; and Threshold τ: a threshold of importance for determining the influence of features. Outputs include Dominant group set $\mathcal{S}_{\text{master}}$, which contains the set of samples that positively influence the model decision; Protect group target set $\mathcal{S}_{\text{protect}}$, which contains the set of samples that negatively influence the model decision.

Specific steps are as follows: 1. Initialize target sets: first, the algorithm initializes two empty target sets $\mathcal{S}_{\text{master}}$ and $\mathcal{S}_{\text{protect}}$, which will be used to store the eligible samples. 2. Iterate over the candidate sets: For each sample (x, y) in the candidate set $\mathcal{C}_{\text{candi}}$, the algorithm extracts the Shapley-valued feature φ of the sample. 3. Judge and classify the samples: For each sample, if its true label y is a positive class and the Shapley-valued feature φ is greater than the threshold τ, the sample is added to the dominant group target set $\mathcal{S}_{\text{master}}$. This step ensures that a sample is included in the dominant group set only if it contributes significantly to the model predictions. Similarly, if the true label y is a negative class and the Shapley value feature φ is less than the negative threshold, that sample is added to the protected group set $\mathcal{S}_{\text{protect}}$, which ensures that the sample from the protected group reflects potential bias of the model The algorithm is then used to generate a sample of the protected group. 4. Return results: Finally, the algorithm returns the combination of the dominant group set and the protected group set $\mathcal{S}_{\text{master}}$ and $\mathcal{S}_{\text{protect}}$. this is done to further analyze and adjust the model's performance across different groups to ensure fairness.

4.3 Discrimination Mitigation Based Shapley Value Algorithm

Algorithm 3 DMBS Algorithm

Input: Dataset $\mathcal{D}(X, Y)$, Dominant group set $\mathcal{S}_{\text{master}}$, Protected group set $\mathcal{S}_{\text{protect}}$
Output: Generated Dataset $\mathcal{D}_{\text{gen}}(X_{\text{gen}}, Y_{\text{gen}})$

function ModifySensitiveAttr(C, ξ, π)
  $X_t \leftarrow \emptyset$ ▷ Init X_t
  $Y_t \leftarrow \emptyset$ ▷ Init Y_t
  for $i \in C$ **do**
    $x \leftarrow X_i$
    if $y_i = \pi$ **then**
      $x[A] \leftarrow \xi$ ▷ Modifying Protected Attributes
    end if
    $X_t \leftarrow X_t \cup \{x\}$
    $Y_t \leftarrow Y_t \cup \{y_i\}$
  end for
return X_t, Y_t
end function
$X_{a^+}, Y_{a^+} \leftarrow \text{ModifySensitiveAttr}(S_{\text{master}}, a^-, y^+)$
$X_{a^-}, Y_{a^-} \leftarrow \text{ModifySensitiveAttr}(S_{\text{protect}}, a^+, y^-)$
$X_{\text{gen}} \leftarrow X \cup X_{a^+} \cup X_{a^-}$ ▷ Construct X_{gen}
$Y_{\text{gen}} \leftarrow Y \cup Y_{a^+} \cup Y_{a^-}$ ▷ Construct Y_{gen}
$D_{\text{gen}} \leftarrow (X_{\text{gen}}, Y_{\text{gen}})$
return D_{gen} ▷ Return Generation Dataset

Algorithm 3 describes the specific implementation of DMBS. The main objective is to mitigate model bias by modifying sensitive features in order to

generate a fair dataset. Inputs include Dataset $\mathcal{D}(X, Y)$: a dataset containing the features X and labels Y; Dominant group set $\mathcal{S}_{\text{master}}$: a collection of samples that represent the dominant group; Protection group set $\mathcal{S}_{\text{protect}}$: the set of samples representing the protected group. The output consists of the generated dataset $\mathcal{D}_{\text{gen}}(X_{\text{gen}}$, modified to mitigate model bias.

Specific steps are as follows: 1. Define the function: Firstly, define the auxiliary function ModifySensitiveAttr, whose parameters are the set C, the value of the protection attribute ξ, and the label value π. The method mainly uses π to judge and modify the value of the protection attribute to ξ, in order to mitigate or eliminate the bias of the data. 2. Initialize Collections: Initialize two empty collections X_t and Y_t to store the modified features and labels. 3. Iterate over the sensitive features: for each sample i in the input set assign the sample feature to x. Check the label of the sample y_i. If the label is equal to the target label π, then modify the sensitive attributes: replace the sensitive attributes in the feature x with a specified modified value. Update the set of features X_t and the set of labels Y_t. 4. Call the function: For the dominant group set $\mathcal{S}_{\text{master}}$ and the protected group set $\mathcal{S}_{\text{protect}}$, respectively, call the ModifySensitiveAttr function to modify the sensitive attributes. For the dominant group, modify its sensitive attributes to a^- and target label y^+. For the protected group, modify its sensitive attributes to a^+ and the target label y^-. 5. Construct the generated dataset: The modified dataset is merged and a new feature set X_{gen} and label set Y_{gen} are constructed by aggregating the modifications of the dominant and protected groups. 6. Return results: Finally, the generated dataset D_{gen} is returned, which has been modified with sensitive features aimed at reducing the bias of the model on specific groups.

5 Experiments and Results

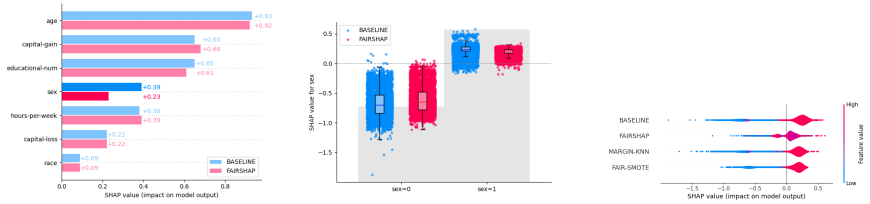
5.1 Experimental models and methods

In this paper, we will use the XGBoost and Random Forest RF models from the Tree Integration Model, both of which are based on the aspect of tree integration in calculating the Shapley values and are more advantageous in terms of computational complexity. Also we chosen MLP.

Experiment uses two comparison methods, one is the MARGIN-KNN[18], which uses KNN combined with a classification interval-based fairness algorithm, the MARGIN-KNN method looks for similar data points from the training data, and based on the principle of the maximum interval, the samples will be projected and the target set will be selected to eliminate the discrimination algorithm. The second is the FAIR-SMOTE[19], the core idea of which is an algorithm that removes bias labels and rebalances the internal distribution using SMOTE oversampling technique.

5.2 ADULT Results

Interpretability Assessment From Figure 3a, it is observed that the Shapley value of the experimental model on the protective attribute SEX is significantly



(a) Global absolute values of Shapley values before and after characterization experiments

(b) Distribution of Shapley values before and after experiments on protected attributes

(c) Comparison of Shapley value distribution of different methods

Fig. 3. ADULT Interpretability Assessment (with SEX as the protected attribute)

changed, and the global Shapley value of SEX is reduced by 39.5%. This shows that the FairShap correction significantly mitigates the gender bias of the model.

From Figure 3b, we can analyze that the distribution of outliers for gender groups is reduced and converges close to zero.

Through the comparative analysis in Figure 3c, it can be observed that in terms of SEX-related Shapley values, although both the MARGIN-KNN and FAIR-SMOTE methods improve fairness, the Shapley values generated by FairShap are more concentrated in the overall distribution, which suggests that the model is able to reflect the contribution of different groups in a more balanced way when dealing with the protective attribute of gender.

Table 1. ADULT Dataset Fairness Assessment with SEX as Protected Attribute

Model	Method	ACC	DI	SPD	EOD	AAOD
XG	BASELINE	0.8556	0.1664	0.1931	0.2503	0.1639
	MARKNN	0.8554	0.1831	0.1839	0.2248	0.1482
	FAIRSMOTE	0.8540	0.2002	0.1770	0.2085	0.1356
	FAIRSHAP	0.8551	**0.2269**	**0.1729**	**0.2063**	**0.1354**
RF	BASELINE	0.8433	0.1537	0.1577	0.2326	0.1448
	MARKNN	0.8410	0.1554	0.1479	0.2159	0.1339
	FAIRSMOTE	0.8450	0.1960	0.1505	0.1985	0.1250
	FAIRSHAP	0.8395	**0.2234**	**0.1268**	**0.1722**	**0.1042**
MLP	BASELINE	0.8057	0.3768	0.1514	0.1925	0.1262
	MARKNN	0.8056	0.3633	0.1684	0.2141	0.1428
	FAIRSMOTE	0.8052	0.2469	0.1722	0.2047	0.1374
	FAIRSHAP	0.8053	**0.4113**	**0.1322**	**0.1544**	**0.1088**

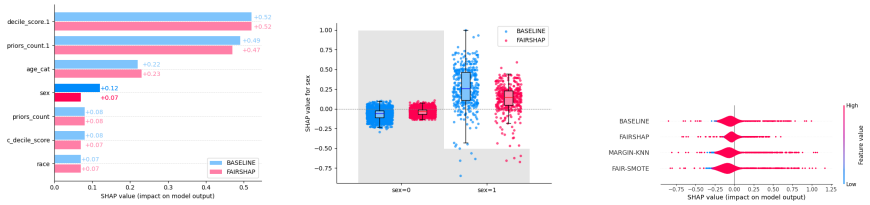
Fairness Assessment Table 1 shows the results using XGBoost, Random Forest, and MLP models on the ADULT dataset, and the related debiasing methods FairShap, MARGIN-KNN, and FAIR-SMOTE results for the ACC and the four fairness indicators DI, SPD, EOD, and AAOD.

Of the two protection attributes, SEX and RACE, the fairness metrics DI, SPD, EOD, and AAOD are all fairer in RACE than SEX. Since the model itself discriminates more deeply against SEX, the effect of FairShap on SEX is obviously better than that of RACE, so here we mainly show the results when the protected attribute is SEX.

As can be seen from Table 1, the FairShap method is better than the MARGIN-KNN method. On some metrics such as SPD and EOD, FairShap and FAIR-SMOTE perform equally well on some models, with reductions of 17.6% and 16.7% on SPD, respectively.

Overall, when dealing with the main protected attributes of the dataset, FairShap is effective in improving the fairness metrics without a significant impact on correctness, and achieves better results than the other two compared methods in most cases.

5.3 COMPAS Results



(a) Global absolute values of Shapley values and after characterization experiments

(b) Distribution of Shapley values before and after experiments on protected attributes

(c) Comparison of Shapley value distribution of different methods

Fig. 4. COMPAS Interpretability Assessment (with SEX as the protected attribute)

Interpretability Assessment Figure 4a shows significant change in Shapley value on the protection attribute SEX for the model after the experiment. The global Shapley value for SEX is reduced by 41.6%, indicating that the FairShap correction significantly mitigates the gender bias of the model.

Analyzing Figure 4b, we can see that the data converges for all gender groups and the extreme outliers are reduced.

Figure 4c shows the distribution of Shapley values on SEX for the three methods and the base method. Through the comparative analysis in Figure 4c,

it can be observed that in terms of SEX-related Shapley values, although both the MARGIN-KNN and FAIR-SMOTE methods improve fairness, the Shapley values generated by FairShap are more concentrated in the overall distribution, which suggests that the model is able to reflect the contribution of different groups in a more balanced way when dealing with the protected attribute of gender.

Table 2. COMPAS Dataset Fairness Assessment with SEX as Protected Attribute

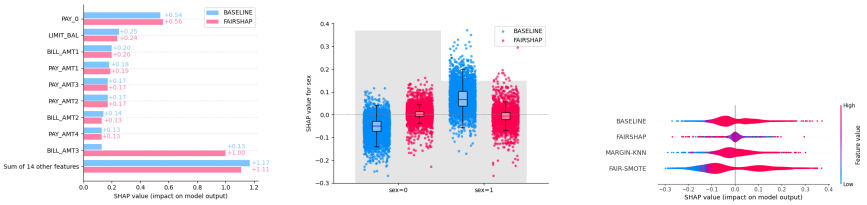
Model	Method	ACC	DI	SPD	EOD	AAOD
XG	BASELINE	0.6734	0.7342	0.2035	0.1095	0.1865
	MARKNN	0.6662	0.7320	0.2079	0.1137	0.1904
	FAIRSMOTE	0.6699	0.7009	0.2358	0.1407	0.2173
	FAIRSHAP	0.6693	**0.7894**	**0.1533**	**0.0616**	**0.1364**
RF	BASELINE	0.6896	0.7509	0.1913	0.1111	0.1642
	MARKNN	0.6901	0.7294	0.2140	0.1304	0.1884
	FAIRSMOTE	0.6933	0.7348	0.2084	0.1174	0.1844
	FAIRSHAP	0.6878	**0.7734**	**0.1684**	**0.0814**	**0.1439**
MLP	BASELINE	0.6970	0.7445	0.2163	0.1318	0.1928
	MARKNN	0.6965	0.7462	0.2133	0.1333	0.1840
	FAIRSMOTE	0.6928	0.7413	0.2113	0.1307	0.1848
	FAIRSHAP	0.6965	**0.7471**	**0.2026**	**0.1302**	**0.1713**

Fairness Assessment Table 2 shows the results using XGBoost, Random Forest, and MLP models on the COMPAS dataset, and the related debiasing methods FairShap, MARGIN-KNN, and FAIR-SMOTE results for the ACC and the four fairness indicators DI, SPD, EOD, and AAOD.

Of the two protection attributes, SEX and RACE, the fairness metrics DI, SPD, EOD, and AAOD are all fairer in RACE than SEX. Since the model itself discriminates more deeply against SEX, the effect of FairShap on SEX is obviously better than that of RACE, so here we mainly show the results when the protected attribute is SEX.

As can be seen from Table 2, FairShap reduces 25.3% on the SPD on the XGBoost model, while FAIR-SMOTE and MARGIN-KNN exacerbate discrimination on some metrics, suggesting that the FairShap method is significantly better than the other two.

Overall, as with the ADULT dataset, when dealing with the protective attributes of the main discrimination of the dataset, FairShap can be effective in improving the fairness metrics without a significant impact on correctness and achieves better results than the other two compared methods in most cases.



(a) Global absolute values of Shapley values before and after characterization experiments

(b) Distribution of Shapley values before and after experiments on protected attributes

(c) Comparison of Shapley value distribution of different methods

Fig. 5. DEFAULT Interpretability Assessment (with SEX as the protected attribute)

5.4 DEFAULT Results

Interpretability Assessment In Figure 5a, we cannot observe any specific change in SEX, mainly because its contribution is not high compared to other attributes, so it is not a major factor influencing the decision.

Figure 5b shows the global characteristic Shapley absolute value plots of SEX for the protected attribute in the DEFAULT dataset before and after FairShap's experiments.

Through the comparative analysis in Figure 5c, it can be observed that in the distribution of Shapley values of attribute SEX among different methods, FairShap is significantly better than the other two methods, in which FAIR-SMOTE exacerbates the discrimination.

Table 3. DEFAULT Dataset Fairness Assessment with SEX as Protected Attribute

Model	Method	ACC	DI	SPD	EOD	AAOD
XG	BASELINE	0.8131	0.8259	0.0254	0.0083	0.0163
	MARKNN	0.8129	0.8157	0.0285	0.0155	0.0185
	FAIRSMOTE	0.8127	0.8372	0.0223	0.0385	0.0355
	FAIRSHAP	0.8122	**0.8888**	**0.0151**	**0.0133**	**0.0131**
RF	BASELINE	0.8199	0.9063	0.0107	0.0186	0.0138
	MARKNN	0.8182	0.8989	0.0112	0.0212	0.0160
	FAIRSMOTE	0.8188	0.8604	0.0170	0.0337	0.0195
	FAIRSHAP	0.8196	**0.9605**	**0.0041**	**0.0041**	**0.0085**
MLP	BASELINE	0.7821	0.7458	0.0001	0.0012	0.0008
	MARKNN	0.7819	0.7301	0.0012	0.0010	0.0001
	FAIRSMOTE	0.7756	0.7482	0.0304	0.0029	0.0183
	FAIRSHAP	0.7813	**0.7981**	0.0012	0.0017	0.0013

Fairness Assessment Table 3 shows the results using XGBoost, Random Forest, and MLP models on the DEFAULT dataset, and the related debiasing methods FairShap, MARGIN-KNN, and FAIR-SMOTE results for the ACC and the four fairness indicators DI, SPD, EOD, and AAOD.

As can be seen from Table 3, after using the FairShap method, except for DI, which will improve slightly, the rest of the indicators are already relatively fair on BASELINE, so the effect of the three debiasing methods is not particularly significant, and the bias is aggravated in FAIR-SMOTE on the fairness indexes EOD and AAOD on the XGBoost model, which reduces fairness.

6 Conclusion

This paper focuses on improving the fairness and interpretability of machine learning models, especially in dealing with human-related decision-making tasks. With the widespread use of machine learning in real-world decision-making systems, it is important to understand and mitigate unfair decisions caused by data discrimination and algorithmic biases that have a profound impact on individual opportunities.

We propose FairShap, an interpretable fairness framework based on feature contributions, which aims to improve the fairness and interpretability of models in binary classification tasks. Based on the assumption that "the protective attributes of individuals within the same group should have similar contributions to their outcomes", and combining with the concept of Shapley value in cooperative Game Theory, the framework performs debiasing in the preprocessing stage of the model to effectively alleviate the unfairness in the dataset, and provides an explanation for the unfairness in the dataset. This can effectively alleviate the unfair phenomenon in the dataset and provide an explainable explanation.

Through the experiments on the datasets ADULT, COMPAS and DEFAULT, this paper verifies that the proposed discrimination detection algorithm DDBS and discrimination elimination algorithm DMBS effectively balance the classification fairness and accuracy, and satisfy the fairness criteria DI, SPD, EOD, AAOD. In the research process, we explore several key issues, including the interpretability effect under different protection attributes, the fairness performance on different datasets, and the comparison with other methods, and discuss the key issues with the experimental results. Finally, the experimental results show that the method outperforms existing algorithms in terms of interpretability and provides an interpretable account of fairness, and demonstrates wide applicability to different classifiers and fairness metrics.

References

1. M. Bojarski, D. D. Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao, and K. Zieba, "End to end learning for self-driving cars," 2016.

2. V. Aseervatham, C. Lex, and M. Spindler, "How do unisex rating regulations affect gender differences in insurance premiums?," *The Geneva Papers on Risk and Insurance-Issues and Practice*, vol. 41, no. 1, pp. 128–160, 2016.
3. K. D. Julian, J. Lopez, J. S. Brush, M. P. Owen, and M. J. Kochenderfer, "Policy compression for aircraft collision avoidance systems," in *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)*, pp. 1–10, IEEE, 2016.
4. Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019.
5. J. Chan and J. Wang, "Hiring preferences in online labor markets: Evidence of a female hiring bias," *Management Science*, vol. 64, no. 7, pp. 2973–2994, 2018.
6. T. Bono, K. Croxson, and A. Giles, "Algorithmic fairness in credit scoring," *Oxford Review of Economic Policy*, vol. 37, no. 3, pp. 585–617, 2021.
7. R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth, "Fairness in criminal justice risk assessments: The state of the art," *Sociological Methods & Research*, vol. 50, no. 1, pp. 3–44, 2021.
8. K. Mavrogiorgos, A. Kiourtis, A. Mavrogiorgou, A. Menychtas, and D. Kyriazis, "Bias in machine learning: A literature review," *Applied Sciences*, vol. 14, no. 19, p. 8860, 2024.
9. S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.
10. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.
11. M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," in *Artificial intelligence and statistics*, pp. 962–970, PMLR, 2017.
12. N. Anderson, S. K. Bera, S. Das, and Y. Liu, "Distributional individual fairness in clustering," *arXiv preprint arXiv:2006.12589*, 2020.
13. D. Madras, E. Creager, T. Pitassi, and R. Zemel, "Learning adversarially fair and transferable representations," in *International Conference on Machine Learning*, pp. 3384–3393, PMLR, 2018.
14. A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," *arXiv preprint arXiv:1611.01236*, 2016.
15. P. Lahoti, K. P. Gummadi, and G. Weikum, "ifair: Learning individually fair data representations for algorithmic decision making," in *2019 ieee 35th international conference on data engineering (icde)*, pp. 1334–1345, IEEE, 2019.
16. H. Hakkoum, A. Idri, and I. Abnane, "Global and local interpretability techniques of supervised machine learning black box models for numerical medical data," *Engineering Applications of Artificial Intelligence*, vol. 131, p. 107829, 2024.
17. C. S. Chan, H. Kong, and G. Liang, "A comparative study of faithfulness metrics for model interpretability methods," *arXiv preprint arXiv:2204.05514*, 2022.
18. X. Shi and Y. Li, "Discrimination sample discovery and elimination algorithm based on categorization interval in fairness machine learning," *Science China: Information Science*, vol. 50, no. 8, pp. 1255–1266, 2020.
19. J. Chakraborty, S. Majumder, and T. Menzies, "Bias in machine learning software: Why? how? what to do?," in *Proceedings of the 29th ACM joint meeting on European software engineering conference and symposium on the foundations of software engineering*, pp. 429–440, 2021.