FAPFID: A Fairness-Aware Approach for Protected Features and Imbalanced Data

Ginel Dorleon, Imen Megdiche, Nathalie Bricon-Souf, Olivier Teste

FAPFID: A Fairness-aware Approach for Protected Features and Imbalanced Data

Ginel Dorleon[1][0000−0003−2343−4445], Imen Megdiche[1][0000−0002−1331−8662], Nathalie Bricon-Souf[1][0000−0003−2150−0998], and Olivier Teste[1][0000−0003−0338−9886]

[1]Toulouse Institute for Computer Science Research (IRIT), France
firstname.lastname@irit.fr

Abstract. The use of automated decision-making based on machine learning algorithms has raised concerns about potential discrimination against minority group defined by protected features such as gender, race, etc. Particularly, in some areas with high social impact such as justice, job search or healthcare, it has been observed that using protected feature in machine learning algorithms can lead to unfair decisions that favor one group (privileged) over another group (unprivileged). In order to improve fairness in decision-making with regard to protected features, many machine learning approaches focus either on discarding the protected features or maintaining an overall accuracy performance for both unprivileged and privileged groups. However, we notice that these approaches have limited efficiency in the case where the protected features are useful for the learning model or when dealing with imbalanced data. To overcome this limitation when dealing with such issues, we propose in this work FAPFID, a fairness-aware strategy based on the use of balanced and stable clusters. To do this, we divide our input data into stable clusters (subgroups) while ensuring that privileged and unprivileged groups are fairly represented in each cluster. Experiments on three real-world and biased datasets demonstrated that our proposed method outperforms state-of-the-art fairness-aware methods under comparison in terms of performance and fairness scores.

Keywords: Decision Systems · Bias · Fairness · Machine Learning · AI

1 Introduction

Nowadays, machine learning-based decision support systems have become increasingly automated while assisting human judgment with largely data-driven decisions. Since these systems are data-driven, they can be applied in a wide variety of applications such as transportation [30], recruitment or employment screening [32], healthcare [13], finance [1] and many more. However, concerns have been raised [42] that machine learning algorithms may lead to decisions against certain groups defined by sensitives or protected features such as gender, race, religion. In areas with high social impacts such as justice, risk assessment, online purchase and delivery, loan application, there are already many

cases [41,29,17,37] where discriminatory decisions have already been made against minority or unprivileged group with harmful consequences. Basically, two majors problems were identified [33,7] as the main cause of the unfairness in automated decision-making: the uncontrolled use of protected/sensitive features and the used of imbalanced datasets [23]. Protected or sensitive features, according to [12], are features that are of particular importance either for social, ethical or legal reasons when making decisions. According to [7], a dataset suffers from class imbalance when there is significant or extreme disproportion between the number of examples of each class in the dataset. By class in the dataset we mean, in the context of supervised machine learning and with a classification task in particular, the label or output we want to predict based on a set of inputs values. Based on a protected feature such as *gender*, a privileged group (male for example) would be more likely to receive an advantageous treatment than the unprivileged group (here, female for example). Such a behavior is not only undesirable but may have serious impact on the unprivileged group [34].

To this end, many machine learning approaches have proposed to help improving fairness in decision-making systems in areas where automated decision-making based on machine learning algorithms are used. Some of the proposed machine learning approaches [11] for fairness improvement with regard to protected features tend to remove them prior the learning model in order to obtain a fair outcome. However, while this strategy may work, we found that it is limited and can lead to a significant performance loss in the case where protected features are relevant for the learning task. Some other approaches [5,18] to improve fairness also tend to focus on maintaining an overall accuracy for both privileged and unprivileged. Again, we noticed that this strategy may not always work when using data that suffer from class imbalance. It has been proved [43,16] that overall accuracy is not always a good performance indicator when using imbalanced dataset since it tends to favor the majority group over the minority. Since most of fairness-related datasets suffer from class imbalance, addressing fairness with regards to protected features in machine learning algorithm also requires addressing the issue of imbalanced dataset.

Thus, in our work, we focused on these two issues, the use of protected features and class imbalance, that directly impact performance and fairness of machine learning algorithms. To this end, we propose FAPFID: A Fairness-aware Approach for Protected Features and Imbalanced Data. Our method allows to handle protected feature and class imbalance while ensuring an efficient and fair model for decision-making involving machine learning algorithms. Using the input dataset, our method creates a set of balanced and stable clusters while ensuring that both privileged and unprivileged groups are fairly represented in each cluster. Then an ensemble learning model is built upon the aggregated balanced and stable clusters which allow to obtain a cumulative and fair model.

Our contributions in this work can be summarized as follows:

- We define a cumulative-fairness approach for dealing with protected features in decision support, it is tested on a binary classification task using an ensemble learning strategy.

- The proposed approach is based on stable and balanced clusters, thus we propose a clustering stability algorithm to this end.
- Our method takes into consideration protected features and class imbalance while making fair decisions, so that the balanced-accuracy score remains high.
- Our method to achieve fairness is based on Equalized of Odds as fairness metric and it being tested on three real-world dataset suitable for fairness study and is easily adaptable to any social decision problems with regards to protected features and class imbalance.

The rest of this paper is organized as follows: in section 2, we summarize the different existing methods to tackle the issues identified with their limitations. In section 3, we introduce some basics concepts and definitions. We present our new approach in section 4. The experimental results are described and analyzed in section 5. Conclusions and future work are presented in section 6.

2 Related Work

Many existing work have proposed various machine learning methods to deal with fairness issues related to the use of protected features and imbalanced data [2]. Here we look at those existing methods under these two categories and we also look at what previous work has defined in terms of fairness metrics.

Many definitions of fairness [39,14] have been proposed over the recent years. Most of the recent proposed methods use fairness definitions such as demographic parity [40,36,22]. This fairness metric suggests that a predictor is unbiased if the prediction ($\hat{y}$) is independent of the protected feature such that positive prediction rate between the two subgroups are the same. Other proposed methods have instead used other fairness metric such as equalized odds [15,31,27]. Unlike demographic parity, this fairness metric instead suggests that the true positive rate and the false positive rate will be the same for both unprivileged and privileged groups. However while each of these definition has merit, there is no consensus on which one is consequently the best, and this issue is beyond the scope of this article. Our goal is not to address the relative virtues of these definitions of fairness, but rather to assess the strength of the evidence presented by a set of subgroup that a model is unfair to a certain group based on a given metric and the best possible trade-off between fairness and performance

For proposed methods that deal with fairness related to protected features, we notice several approaches [25,24]. Particularly, we notice the work in [11,9] where authors introduced naive approaches consisting of removing completely all protected features of the dataset to ensure fairness. However, we notice that these approaches may not solve the problem because there may be redundant features or even proxies to the protected [38]. As underlined by [42], some features known as proxies such as zip code, for example, can reveal the predominant race of a residential area. Thus, this can still lead to racial discrimination in a decision making problem such as loan application despite the fact that zip code appears to be a non-protected feature. We also notice the work of [19]

where authors introduced a framework that combines pre-processing balancing strategy with post-processing decision boundary adjustment in order to deal with fairness related to protected features and class imbalance. In the pre-processing strategy, they created local subgroups where they performed random under-sampling technique to guarantee equitable representation between minority and majority groups. While this strategy may work on large datasets with thousands of instances, we notice that it suffers from a performance loss when used on a restricted dataset.

Given the limitations of the above approaches, there is a need for more in-depth research to overcome these limitations. Thus, we propose FAPFID, a new fairness-aware strategy that allows the obtaining of an efficient and fair models with regards to protected features and imbalanced data. We would like to recall here, as part of our approach, a given model is said to be "fair", or "equitable", if its results are independent of one or more given features, in particular those considered to be protected [28,21].

3 Basic Concepts and Definitions

In this paper, we consider an input dataset $S = \left(X_{m,n}, Y_{1,n}\right)$ that consists of n observations and m features. Let f be a learning model and its performance score $f[S]$ which will be used to predict a binary output $\hat{y} \in \{0, 1\}$. Each sample $x_i \in X_{m,n}$ is associated to a protected feature P, for simplicity we consider that P is binary: $P \in \{P_0, P_1\}$. We consider P_0 to be an unprivileged group and P_1 a privileged group. For instance, P='gender' could be the protected attribute with P_0 ='female', the unprivileged group, and P_1='male' the privileged one. Likewise, we consider $\hat{y} = 1$ to be the preferred outcome, assuming it represents the more desirable of the two possible outcomes.

Suppose for some samples we know the ground truth; i.e., the true value $y \in \{0, 1\}$. Note that these outcomes may be statistically different between different groups, either because the differences are real, or because the model is somewhat biased. Depending on the situation, we may want our estimate $\hat{y}$ to take these differences into account or to compensate them.

Choice of Fairness Metric: In this work, we have used Equalized Odds (EqOd) as fairness metric since it is widely used and adopted by recent state-of-the-art method and other methods. EqOd measures the difference of true classified examples between privileged and unprivileged group in all classes [3]. That being said, prediction $\hat{y}$ is conditionally independent of the protected feature P, given the true value y : $Pr(\hat{y}|y, P) = Pr(\hat{y}|y)$. This means that the true positive rate and the false positive rate will be the same between the privileged and unprivileged groups. To compute the difference between classified instances of the two groups, EqOd is defined as follow:

$$EqOd = Pr(\hat{y} = 1|P_1, y = y_i) - Pr(\hat{y} = 1|P_0, y = y_i), y_i \in \{0, 1\} \qquad (1)$$

A fair value for this metric is between $[-0.1, 0.1]$. The ideal value of this metric is 0. A value < 0 implies higher benefit for the privileged group and a value > 0 implies higher benefit for the unprivileged group.

Ensemble Learning Choice: We will use an ensemble learning strategy to help obtaining a final model. Ensemble learning helps improving machine learning results by combining several intermediate models. This approach allows the production of better predictive performance compared to a single model. For our ensemble learning strategy, we will use Bagging. Also known as bootstrap aggregating, Bagging is the aggregation of multiple versions of a predicted model. Each model is trained individually upon a subset, and combined using a majority voting process. Thus, we believe using an ensemble learning is an efficient technique to tackle imbalanced ratio towards protected feature as it divides the learning problem into multiple sub-problems and then combines their solutions (local models) into an final model. Intuitively, we found it easier to tackle the problem related to fairness in the subset with locals models rather than in a single and global model.

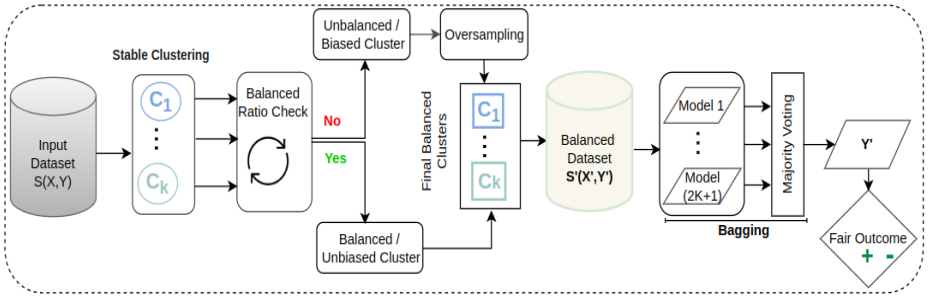

Fig. 1. FAPFID: the proposed approach with different steps

4 Proposed Approach

We introduce in this section our approach, shorten as FAPFID, to achieve fairness as illustrated in Fig. 1. It works as follows: first the input data is divided into K stable clusters by a clustering strategy; then we ensure that obtained clusters are balanced with respect to the protected feature in each cluster. In the case where some clusters are imbalanced, we apply an oversampling technique, SMOTE [6]. Then a final set of balanced clusters is constructed. The final ensemble is then divided into bags where we apply an ensemble learning strategy, Bagging. A learner is trained on each bag and then a final model is obtained by majority vote. Below we describe each step.

4.1 Stable Clustering

In this step, we use a strategy to ensure that the number of clusters that we obtain are stable, i-e optimal. For this, we define a stability strategy to strengthen our clustering solution.

Why stable clusters ? Obtaining stable clusters is useful to maintain a great performance hence ensuring a reliable fairness. A stable clustering guarantees a better homogeneity within clusters and ensure that the instances are really in their respective clusters [8]. Thus, we establish a clustering stability strategy based on K-means to avoid that wrongly clustered instances impact the balancing strategy that we later perform. In order to guarantee the obtaining of stable clusters, we define a statistical setup. Our stability strategy aims to provide information on the variation of instances for different values of k between two clusterings solutions of sub-samples of the same dataset. Thus, for each value of k, we seek to obtain a stability rate by looking at the percentage of instances, points or pairs of points on which the two clusterings agree or disagree. The value of k whose instances variation percentage between the two clusterings is closer to zero will be the one that guarantees the best stability, and therefore the optimal value of k to choose.

Stability Strategy. Here we define our clustering stability approach which is based on K-means. The generic clustering algorithm receives as input a dataset $S = \left(X_{m,n}, Y_{1,n}\right)$ and an additional parameter K. It then assigns clusters to all samples of S. The dataset S is assumed to consist of n samples 1, . . ., x_n that have been drawn independently from a probability distribution T on some space X.
Assume we agree on a way to compute distances $d(C, C')$ between clusterings C and C'. Then, for a fixed probability distribution T, a fixed number K of clusters and a fixed sample size n, the stability of the clustering algorithm is defined as the expected distance between two clusterings $C_K(S_z), C_K(S'_z)$ on different samples S_z, S'_z of size z, that is:

$$C_{stab}(K, z) = d(C_K(S_z), C_K(S'_z)) \tag{2}$$

Algorithm 1 below shows how we performed the stability analysis.

In line (8), since the two clusterings are defined on the same samples, then it is straightforward to compute a distance score between these clusterings using any of the well-known clustering distances such as the Rand index, Jaccard index, Hamming distance, Variation of Information distance [26]. All these distances estimate, in some way or the other, the percentage of points or pairs of points on which the two clusterings C_z and $C_{z'}$ agree or disagree. In our experiments, we have used the Jaccard Index Distance [35].
Jaccard Index Similarity. The Jaccard similarity is a measure of how close two clusters, $C_z, C_{z'}$ are. The closer the clusters are, the higher the Jaccard

Algorithm 1: Clustering stability algorithm

Input: a set **S** of samples, a clustering algorithm **A**, k_{max} clusters and z_{max} samples.
Output: Optimal value of **K**

Begin
for $k = 2$ ... k_{max}:
    Generate z_{max} subsamples S_z ($z = 1$, ..., z_{max}) of S
    for z = 1 ... z_{max}:
        Split S_z into k clusters C_z using A
    end for
    for z, $z' = 1$ ... z_{max} :
        Compute pairwise distance $d(C_z, C_{z'})$ using Jaccard index distance(4)
        Compute stability as the mean distance between clustering C_z and $C_{z'}$ as: $C_{stab}(k, z_{max}) = \frac{1}{z_{max}^2} \sum_{z,z'=1}^{z_{max}} d\,(C_z, C_{z'})$
        Choose the parameter K with highest C_{stab}
    end for
end for
End

similarity. We can associate an actual distance measure to it, which is called the Jaccard distance. The Jaccard similarity of two clusters C_z and $C_{z'}$ is given by:

$$SIM(C_z, C_{z'}) = \frac{C_z \cap C_{z'}}{C_z \cup C_{z'}} \tag{3}$$

The Jaccard distance $d(C_z, C_{z'})$ is then given by (4) and, it equals 1 minus the ratio of the sizes of the intersection and the union of the clusters C_z and $C_{z'}$.

$$d(C_z, C_{z'}) = 1 - SIM(C_z, C_{z'}) \tag{4}$$

4.2 Balanced Check Ratio

The main goal here is to divide the clusters into balanced and imbalanced clusters. We compute the ratio rp (5) between privileged and unprivileged instances for each cluster:

$$rp = \frac{privileged}{unprivileged} \tag{5}$$

Clusters with ratio $rp \neq 1$ are considered to be biased thus are sent to the oversampling stage to be oversampled using SMOTE [6]. We qualify these clusters as biased by the fact that the ratio $rp \neq 1$ reflects the presence of the demographic bias between privileged and non-privileged instances (group imbalance). We apply the SMOTE strategy in a different way of what have being used. In the original paper where SMOTE has been introduced [6], it is applied globally to the minority class. However, SMOTE in our approach is only applied to protected features label, that means our clusters are balanced towards the

unprivileged and privileged group and not the class label. Once the imbalanced clusters are oversampled, we construct a set of final balanced clusters that are therefore aggregated into a final set from which bags will be created to train different classification models.

4.3 Bagging

Estimating the number of bags b must be sufficient to construct enough learners, since we consider each bag as a sample of the training data. To ensure that all the clustered instances are at least in one of the bags, we estimate the number of bags b as: $b = 2K + 1$, K is the number of stable clusters obtained in 4.1 with Algorithm 1. Since we will consider a classification task, the final model will be chosen by a majority voting strategy.

4.4 Proposed Method

Using the basic concepts that we previously defined in section 3, the algorithm defined below takes as inputs a clustering algorithm A, a set of samples S, K number of clusters, privileged group P_1, unprivileged group P_0 and a base classifier G. We start by initializing an empty set of balanced clusters M (2) which later will contains the final balanced clusters as explained in 4.2. Then split S into K clusters (the value of K is known for each dataset with algorithm 1) using A to obtain C_i, $i = 1...K$ (3). For each C_i cluster, we compute the imbalance ratio between privileged group P_1 and unprivileged group (P_0) of clusters C_i. If the computed ratio is equal to 1, we add the current cluster C_i to M (4-6), that means this cluster is balanced toward privileged and unprivileged group. However, if the computed ratio is not equal to 1, we oversample the current cluster C_i using SMOTE [6] to obtain a balanced cluster C_i^{bal}. We add this balanced cluster C_i^{bal} to the final set M then we start over using a different value of i(7-11).

Once we have used all the values of K and obtain our final list of balanced cluster M, we create a balanced dataset X' from M (12). Then, we create b, $(b = 2 * K + 1)$, number of bags from X'. For each bag X'_j extracted from X', we train a model using the base classifier G (14-16). The final output ensemble model E is obtained by a majority vote over G_j (17).
After obtaining the final ensemble model E, we then compute the performance scores based on accuracy and balanced-accuracy, and we compute the fairness score using Equalized of odds (EqOd).

5 Experiment and Results

In this section we detail on the experimental approach, our goal, the learning parameter, the dataset used, baseline and results.

Algorithm 2: Pseudo-code of the proposed method

Input: a clustering algorithm **A**, **S** samples, **K** number of clusters, privileged group P_1 & P_0, a base classifier **G**
Output: Ensemble Model **E**

Begin
M← { } *//final set of balanced cluster*
Split **S** into **K** clusters C_i, $i = 1...K$ using **A**
for $i = 1...K$:
    if rp $C_i(P_1)/C_i(P_0) = 1$:
        M ← M ∪ $\{C_i\}$
    else
        C_i^{bal} ← SMOTE(C_i)
        M ← M ∪ $\{C_i^{bal}\}$
    end if
end for
Create X' from M
for $j = 1...2K + 1$:
    Extract boostrap sample X'_j from X'
    Fit $G_j(X'_j)$
end for
Output **E** : ensemble model of G_j
End

5.1 Goal

We carried out an experimental approach with the goals of i) comparing our method FAPFID to existing methods of fairness [20,19,5] and ii) assessing the impacts of the imbalance ratio between P_0 and P_1 on the performance of FAPFID (section 5.7). In particular, for the first goal, the comparison was made based on two criterion: performance and fairness score. For performance score, we have used Accuracy and Balanced-Accuray. Accuracy summarizes the performance of the classification task by dividing the total correct prediction over the total prediction made by the model. It is the number of correctly predicted samples out of all the samples. However, since all of the three datasets used are highly imbalanced, we also use Balanced-accuracy [4] in order to shade more lights on our model's evaluation on imbalanced datasets compared to the Accuracy. It is the arithmetic mean of the true positive rate for each class.

5.2 Learning, evaluation and parameters

To evaluate and compare the proposed method to existing methods, we proceeded to a learning task by considering a binary classification problem over the three datasets that we described below in section 5.3. For this binary classification problem, Decision Tree is used as base classifier. This choice is made in order to be consistent with the evaluation protocol for concurrent methods. For training and testing, first we use the classic train-test split strategy with a

70%-30% respectively then use k-fold validation on the train set, $2K + 1$ folds in total with K the number of clusters obtained for the used dataset. The folds are made by preserving the percentage of samples for each class.

5.3 Datasets

To evaluate our method, we carried out experiments using three well-known and real-world datasets [10]. They each contain known protected features, which allowed us to evaluate our method on appropriate cases. These datasets were chosen on the basis of the differences and the characteristics, i.e, number of instances, dimensionality and class imbalance. These datasets also provide an interesting benchmark, which is tough, for fairness evaluation as most of recent proposed fairness approaches in the literature have used them. Moreover, they facilitated our comparison with other competitors.

- **Adult census income** dataset [10] contains census data from the U.S whose task it to predict whether someone's income exceeds "50K/yr". After removing duplicate instances and instances with missing values, we ended up with $n = 45,175$ instances. Like our competitors, $P = gender$ was considered as protected feature with $P_0 = female$ and $P_1 = male$. Ratio between unprivileged and privileged instances is 2.23 and 3.53 between classes.
- **Bank dataset [10]** is related to direct marketing campaigns of a Portuguese banking institution with $n = 40,004$ instances. The task is to determine whether a person subscribes to the product (bank term deposit). As target class we consider people who subscribed to a term deposit. Again like our competitors, we consider $P = maritalstatus$ as protected feature with $P_0 = married$ and $P_1 = unmarried$. The dataset suffers from severe class imbalance with global ratio between unprivileged and privileged instances of 2.13. Imbalance ratio between classes is 7.57.
- **KDD census dataset [10]** is basically the same with Adult census, however the target field of this data, was drawn from the "total person income" field rather than the "adjusted gross income" and, therefore, behave differently than the original ADULT target field. This dataset is very skewed, the global ratio between unprivileged and privileged instances is 1.09 . $P = gender$ was considered as protected feature with $P_0 = female$ and $P_1 = male$ like in the other methods used for comparison. This is a very skew dataset in terms of class imbalance, the ratio between classes is 15.11. More details on these datasets are given in Table 1.

5.4 Experimental Baseline

We compare our approach to three other recent state-of-the-art proposed methods tackling the problem of imbalance and protected attributes with the aim of improving fairness. The three other approaches used for comparison are:

Table 1. Experimental Datasets used and characteristics. For each dataset, n instances: number of instances of each dataset, m Features: number of features, P Feature: protected feature, P Ratio: ratio between privileged (P_1) and unprivileged (P_0) group of the protected feature, Class Ratio: ratio between class label of the dataset

	Adult Income	Bank	KDD Adult
n Instances	45175	40004	299285
m Features	14	16	41
P Feature	Gender	Marital S.	Gender
Privileged	Male	Unmarried	Male
Unprivileged	Female	Married	Female
P Ratio	2.23	2.13	1.09
Class Ratio	3.53	7.57	15.11
Majority Label	1	1	1

- *AdaFair* [20]: This method is a fairness-aware boosting approach that adapts AdaBoost to fairness by changing the data distribution at each round based on the notion of cumulative fairness.
- *Fairness Aware Ensemble (FAE)* [19]: This strategy is fairness aware classification that combines pre-processing balancing strategies with post-processing decision boundary adjustment. They use a bagging approach to create sub-datasets while handling the imbalance by an undersampling strategy.
- *SMOTEBoost* [5]: This is an extension of AdaBoost for imbalanced data where new synthetic instances of the minority class are created using SMOTE [6] at each boosting round to compensate the imbalance. This strategy does not tackle the fairness problem, however we used its performance score to evaluate fairness and see if by addressing only the imbalance between classes, the fairness problem can be resolved.

5.5 Results Analysis

We present in the Tables 2, 3, 4 and 5 the results obtained with the different methods. For every dataset, first, we present the result for our stability algorithm that allows us to select the K numbers of stable clusters to use prior our learning strategy. Secondly, for predictive performance, we report on Balanced Accuracy (Bal. Acc.) and Accuracy, for fairness, we report Equalized of Odds (EqOd).

Cluster Stability. In Table 2 below, we report the results for our stability algorithm, the value of K and the stability rate for each dataset. For Adult Income dataset, the best and stable value for K is 4 with a stability rate of 93%. This means, among all of possible values for K, we tried 12 values, $K = 4$ is the one that allowed us to obtain more consistent and stable clusters. For the other datasets, the respective stability rates are 90 and 93%

Table 2. Cluster stability

	Adult Income	Bank	KDD Adult
K Value	4	5	4
Stability Rate	93%	90%	93%

Adult Income Dataset. Performance results with the different approaches for this dataset is presented in Table 3. For predictive performance, we can see that three methods, our proposed one, AdaFair and FAE achieve the same and highest performance score of 83% for Accuracy. However, like we stated above, Accuracy is not as good when we are dealing with imbalanced data. Since this dataset suffers from class imbalance, Balanced-Accuracy is the metric that will tell us how good our model is in terms of performance score. For Balanced-Accuracy, our proposed method outperforms our competitors with a score of 83% as the highest, then FAE and SMOTEBoost both with 81%. We notice that our proposed method performance score is the same for Balanced-Accuracy and Accuracy, this is meaningful since it highlights our strategy of balancing with regards to the protected features in each subgroup prior training the classifier.
For fairness score, we see clearly that our proposed method has surpassed the other three methods used for comparison. Our proposed method has the lowest Equalized Odds score, 0.05 (the lower the better for EqOd) following by AdaFair with 0.08. In short, the proposed method outperforms our competitors on this dataset in terms of Balanced-Accuracy and Fairness score.

Table 3. Adult Income: Predictive and Fairness performance, the best results are in bold.

Score	FAPFID	AdaFair	FAE	SMOTEBoost
Bal. Acc.	**0.83**	0.78	0.81	0.81
Accuracy	**0.83**	**0.83**	**0.83**	0.80
EqOd	**0.05**	0.08	0.15	0.47

Bank Dataset. Performance results with the different approaches for this dataset is presented in Table 4. For predictive performance, our proposed method and SMOTEBoost achieve the same and highest performance score of 90% on Accuracy. However, since we are dealing with imbalanced data, we look at Balanced-Accuracy instead. For this, our proposed method achieves the highest score for Balanced-Accuracy, 88% following by the others with a Balanced-Accuracy score under 79%.
For fairness score, our proposed method has surpassed the other three methods used for comparison since it has the lowest Equalized Odds score, 0.06 following

by FAE and SMOTEBoost with -0.12 and 0.12 respectively, which based on the definitions of Equalized Odds are not fair at all . Again, the proposed method outperforms our competitors on this dataset in terms of Balanced-Accuracy and Fairness score.

Table 4. Bank Dataset: Predictive and Fairness performance, the best results are in bold.

Score	FAPFID	AdaFair	FAE	SMOTEBoost
Bal. Acc.	**0.88**	0.77	0.78	0.74
Accuracy	**0.90**	0.87	0.83	**0.90**
EqOd	**0.06**	0.27	-0.12	0.12

KDD Adult Dataset. Performance results with the different approaches for this dataset is presented in Table 5. For predictive performance, we can see that FAE achieves the highest performance score of 95% for Accuracy following by SMOTEBoost, 94% then the proposed method, 92%. However, our proposed method has the highest Balanced-Accuracy score, 91% which is the one we look at if since this dataset is highly imbalanced. Despite the fact that FAE has the highest Accuracy score, it fails to provide a great Balanced-Accuracy score, it achieves the lowest score of 66%. That means, since this dataset is highly imbalanced, FAE has a higher predictive rate for one group at the expense of the other. FAPFID, our proposed approach instead, has a better fairness score, 0.01 which is the lowest here on this dataset.
In brief, FAPFID outperforms our competitors on this dataset in terms of Balanced-Accuracy and Fairness score.

Table 5. KDD Adult: Predictive and Fairness performance, the best results are in bold.

Score	FAPFID	AdaFair	FAE	SMOTEBoost
Bal. Acc.	**0.91**	0.84	0.66	0.76
Accuracy	0.92	0.86	**0.95**	0.94
EqOd	**0.01**	0.07	0.27	0.36

5.6 Discussion

The results on these three datasets show that our method performs well. Compared to the three other fairness-aware methods for dealing with protected feature and data imbalance, we clearly see that our a method has a higher score for

Balanced-Accuracy and the lowest score for fairness evaluation. Even in the case where other methods achieve a higher or equal value for Accuracy, our method still outperforms them in terms of Balanced-accuracy and fairness core (EqOd). This is very interesting for handling social decisions problems guarantying a fair outcome for different groups.

In general, on these 3 datasets, we get satisfactory results and we have maintained a good level of performance (Balanced-accuracy) and the best fairness score (the lowest) in terms of Equalized of Odds.

5.7 Effects of Imbalance Ratio

The second goal of our experiments is to evaluate the effects of different imbalance ratios on the performance. Our method is able to achieve efficient and reliable results on the benchmarks datasets above. However, in this section, we investigate the effects of imbalance ratio between privileged and unprivileged group for a given dataset. The goal is to observe the evolution of performance scores of the proposed method with regards to different imbalance ratio. Thus, for a given dataset, we create 10 sub-samples where we maintain a fixed imbalance ratio between privileged and unprivileged group, then we report the balanced accuracy for these 10 sub-samples using box-plot.

Basically we proceed as follow: we consider a ratio of 40/60 between unprivileged (P_0) and privileged(P_1) group and create 10 sub-samples, i-e, each sub-sample is created with 40% of (P_0) and 60% of (P_1). We repeated this by varying the ratio such that we obtain different imbalanced ratios between privileged and unprivileged group. The different ratio that we used are: 30/70, 20/80, 10/90 and 1/99.

We report on Fig. 2 the results obtained with Adult Income dataset for performance using Balanced-Accuracy.
As we can see, there is a huge difference between performance scores for different ratio of imbalance. For an imbalance ratio of at least 20% (for P_0), our method still maintains a great averaged Balanced-Accuracy score of 80% at least. With an imbalanced ratio of 10/90, our method suffers from a decreasing in terms of Balanced-Accuracy. We also tested on an extreme case of imbalance ratio between P_0 and P_1 :1/99 where we observed a performance loss. This is because there are not enough P_0 in the cluster so the oversampling method used, SMOTE, can not generate as many meaningful samples for the under-represented group P_0.

We also report on Fig. 3 the results obtained with Adult Income dataset for fairness using Equalized of Odds. For an imbalanced ratio between 20/80 and 40/60, we get satisfactory results in terms of fairness score with an average score under 0.1 which acceptable for Equalized Odds. However, starting at 10/90 to lower, our method has limited ability to maintain a high level of fairness on this dataset due to the limitations of the oversampling method used and the lack of data for the under-represented group. A limitation that we will later overcome

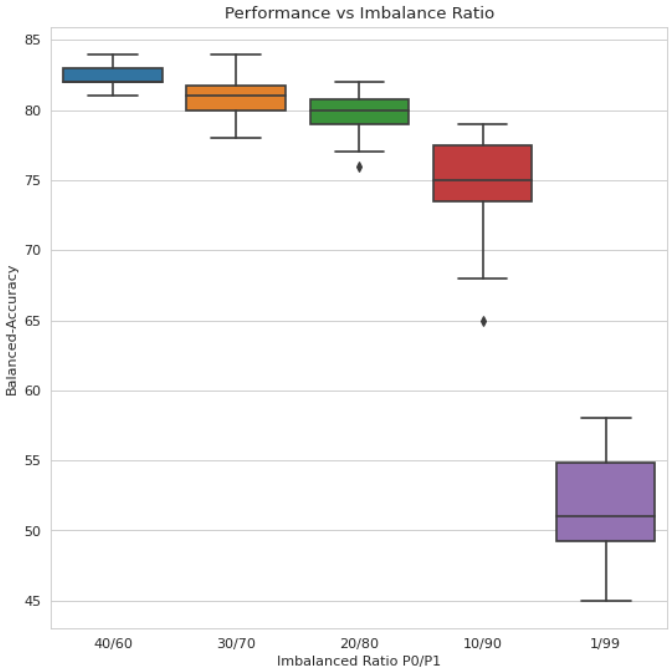

Fig. 2. Effects of Imbalanced Ratio on Balanced-Accuracy

in our future work.

6 CONCLUSION

In this article, we have proposed a fairness-aware ensemble learning method based on balanced and stables clusters. The proposed method achieves fairness with regards to protected features and class imbalance while maintaining a great performance score.
To do this, we divide the inputs dataset into stable clusters and ensure that privileged and unprivileged groups are fairly represented in each clusters. To obtain stable clusters, we introduce a stability clustering approach that helps maintaining a better homogeneity within clusters. To ensure that privileged and unprivileged instances are fairly represented in each cluster, we have used a novel strategy where we compute a balanced ratio rate within cluster and apply SMOTE only on clusters where the balanced ratio is $\neq 1$.
The performance of our method was experimentally evaluated on three well known biased datasets. Compared to three recent state-of-the-art fairness-aware methods, we obtain satisfactory results and the proposed approach outperforms our competitors in terms of performance (Balanced-Accuracy) and fairness

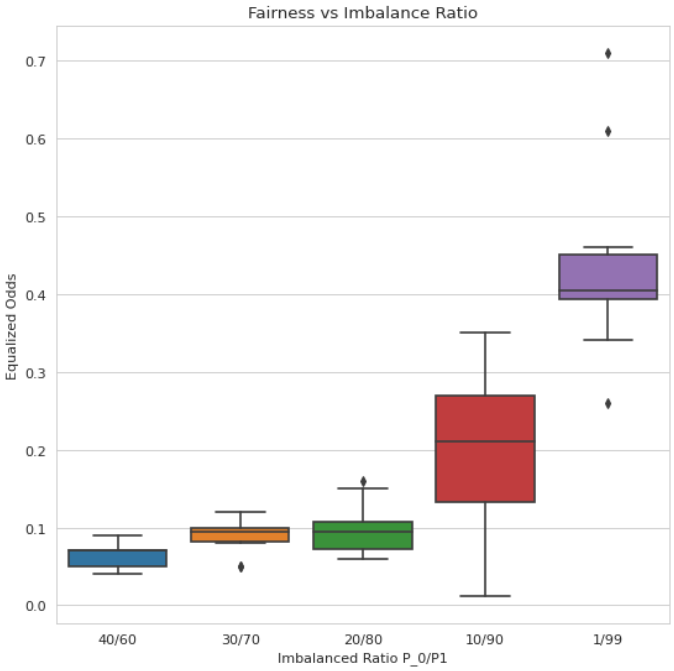

Fig. 3. Effects of Imbalanced Ratio on Fairness

(EqOd) scores. The comparative results obtained show our method's effectiveness in boosting fairness while maintaining a high level of performance.

For our future work, we will look forward to generalise our approach on datasets that are not part of this benchmark and improve our model's performance in dealing with datasets that suffer from a high (10/90) imbalance ratio.

Source Code: The full source code including data of our experiments is available on GitHub under request.

References

1. Amarasinghe, T., Aponso, A., Krishnarajah, N.: Critical analysis of machine learning based approaches for fraud detection in financial transactions. In: Proceedings of the 2018 International Conference on Machine Learning Technologies. p. 12–17. ICMLT '18, Association for Computing Machinery, New York, NY, USA (2018)
2. del Barrio, E., Gordaliza, P., Loubes, J.M.: Review of mathematical frameworks for fairness in machine learning (2020). https://doi.org/10.48550/ARXIV.2005.13755, `https://arxiv.org/abs/2005.13755`
3. Bellamy, R.K.E., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P.K., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K.N.,

Richards, J.T., Saha, D., Sattigeri, P., Singh, M., Varshney, K.R., Zhang, Y.: Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. ArXiv **abs/1810.01943** (2018)
4. Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M.: The balanced accuracy and its posterior distribution. In: 2010 20th international conference on pattern recognition. pp. 3121–3124. IEEE (2010)
5. Chawla, N., Lazarevic, A., Hall, L.O., Bowyer, K.: Smoteboost: Improving prediction of the minority class in boosting. In: PKDD (2003)
6. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: Synthetic minority over-sampling technique **16**(1), 321–357 (jun 2002)
7. Chawla, N.V., Japkowicz, N., Kotcz, A.: Special issue on learning from imbalanced data sets. ACM SIGKDD explorations newsletter **6**(1), 1–6 (2004)
8. Dash, M., Liu, H.: Feature selection for clustering. In: Pacific-Asia Conference on knowledge discovery and data mining. pp. 110–121. Springer (2000)
9. 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)
10. Dua, D., Graff, C.: UCI machine learning repository (2017)
11. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. p. 214–226. ITCS '12, Association for Computing Machinery, New York, NY, USA (2012)
12. Fang, B., Jiang, M., Cheng, P., Shen, J., Fang, Y.: Achieving outcome fairness in machine learning models for social decision problems. In: Bessiere, C. (ed.) Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020. pp. 444–450. ijcai.org (2020)
13. Farahani, B., Barzegari, M., Aliee, F.S.: Towards collaborative machine learning driven healthcare internet of things. In: Proceedings of the International Conference on Omni-Layer Intelligent Systems. p. 134–140. COINS '19, Association for Computing Machinery, New York, NY, USA (2019)
14. Garg, P., Villasenor, J., Foggo, V.: Fairness metrics: A comparative analysis. In: 2020 IEEE International Conference on Big Data (Big Data). pp. 3662–3666. IEEE (2020)
15. Ghassami, A., Khodadadian, S., Kiyavash, N.: Fairness in supervised learning: An information theoretic approach. In: 2018 IEEE International Symposium on Information Theory (ISIT). pp. 176–180. IEEE (2018)
16. Gu, Q., Zhu, L., Cai, Z.: Evaluation measures of the classification performance of imbalanced data sets. In: International symposium on intelligence computation and applications. pp. 461–471. Springer (2009)
17. Hamilton, M.: The sexist algorithm. Behavioral sciences & the law **37**(2), 145–157 (2019)
18. Hu, S., Liang, Y., Ma, L.T., He, Y.: Msmote: Improving classification performance when training data is imbalanced. 2009 Second International Workshop on Computer Science and Engineering **2**, 13–17 (2009)
19. Iosifidis, V., Fetahu, B., Ntoutsi, E.: Fae: A fairness-aware ensemble framework. 2019 IEEE International Conference on Big Data (Big Data) pp. 108–110 (2019)
20. Iosifidis, V., Ntoutsi, E.: Adafair: Cumulative fairness adaptive boosting. p. 781–790. CIKM '19, Association for Computing Machinery, New York, NY, USA (2019)

21. Ji, D., Smyth, P., Steyvers, M.: Can i trust my fairness metric? assessing fairness with unlabeled data and bayesian inference. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 18600–18612. Curran Associates, Inc. (2020)
22. Jiang, Z., Han, X., Fan, C., Yang, F., Mostafavi, A., Hu, X.: Generalized demographic parity for group fairness. In: International Conference on Learning Representations (2021)
23. Kotsiantis, S., Kanellopoulos, D., Pintelas, P., et al.: Handling imbalanced datasets: A review. GESTS international transactions on computer science and engineering **30**(1), 25–36 (2006)
24. Lahoti, P., Beutel, A., Chen, J., Lee, K., Prost, F., Thain, N., Wang, X., Chi, E.: Fairness without demographics through adversarially reweighted learning. Advances in neural information processing systems **33**, 728–740 (2020)
25. Martínez, N., Bertrán, M., Papadaki, A., Rodrigues, M.R.D., Sapiro, G.: Blind pareto fairness and subgroup robustness. In: ICML (2021)
26. Meilă, M.: Comparing clusterings by the variation of information. In: COLT (2003)
27. Mishler, A., Kennedy, E.H., Chouldechova, A.: Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. pp. 386–400 (2021)
28. Oneto, L., Chiappa, S.: Fairness in machine learning. Recent Trends in Learning From Data pp. 155–196 (2020)
29. O'Reilly-Shah, V.N., Gentry, K.R., Walters, A.M., Zivot, J., Anderson, C.T., Tighe, P.J.: Bias and ethical considerations in machine learning and the automation of perioperative risk assessment. BJA: British Journal of Anaesthesia **125**(6), 843 (2020)
30. Paparrizos, I., Cambazoglu, B.B., Gionis, A.: Machine learned job recommendation. In: Proceedings of the Fifth ACM Conference on Recommender Systems. p. 325–328. RecSys '11, Association for Computing Machinery, New York, NY, USA (2011)
31. Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., Weinberger, K.Q.: On fairness and calibration. Advances in neural information processing systems **30** (2017)
32. Qin, Z.T., Tang, J.: Deep reinforcement learning with applications in transportation. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining. p. 3201–3202. KDD '19, Association for Computing Machinery, New York, NY, USA (2019)
33. Ristanoski, G., Liu, W., Bailey, J.: Discrimination aware classification for imbalanced datasets. In: Proceedings of the 22nd ACM international conference on Information & Knowledge Management. pp. 1529–1532 (2013)
34. Romei, A., Ruggieri, S.: A multidisciplinary survey on discrimination analysis. The Knowledge Engineering Review **29**, 582 – 638 (2013)
35. Shameem, M.U.S., Ferdous, R.: An efficient k-means algorithm integrated with jaccard distance measure for document clustering. 2009 First Asian Himalayas International Conference on Internet pp. 1–6 (2009)
36. Singh, A., Joachims, T.: Fairness of exposure in rankings. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 2219–2228 (2018)
37. Sweeney, L.: Discrimination in online ad delivery (2013)
38. Teste, O.: Feature selection under fairness and performance constraints. In: Big Data Analytics and Knowledge Discovery: 24th International Conference, DaWaK

2022, Vienna, Austria, August 22–24, 2022, Proceedings. vol. 13428, p. 125. Springer Nature (2022)
39. Verma, S., Rubin, J.: Fairness definitions explained. In: Proceedings of the International Workshop on Software Fairness. p. 1–7. FairWare '18, Association for Computing Machinery, New York, NY, USA (2018)
40. Wadsworth, C., Vera, F., Piech, C.: Achieving fairness through adversarial learning: an application to recidivism prediction. arXiv preprint arXiv:1807.00199 (2018)
41. Washington, A.L.: How to argue with an algorithm: Lessons from the compas-propublica debate. Colo. Tech. LJ **17**, 131 (2018)
42. Yeom, S., Datta, A., Fredrikson, M.: Hunting for discriminatory proxies in linear regression models. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. p. 4573–4583. NIPS'18, Curran Associates Inc., Red Hook, NY, USA (2018)
43. Zhuang, L., Dai, H.: Reducing performance bias for unbalanced text mining. In: Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06). pp. 770–774 (2006)