IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 9, 2021, accepted May 23, 2021, date of publication May 27, 2021, date of current version June 11, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3084121

FAWOS: Fairness-Aware Oversampling Algorithm Based on Distributions of Sensitive Attributes

TERESA SALAZAR[1], MIRIAM SEOANE SANTOS[1], HELDER ARAÚJO[2], (Member, IEEE), AND PEDRO HENRIQUES ABREU[1], (Member, IEEE)

[1]Centre for Informatics and Systems, Department of Informatics Engineering, University of Coimbra, 3030-790 Coimbra, Portugal
[2]Department of Electrical and Computer Engineering, University of Coimbra, 3004-531 Coimbra, Portugal

Corresponding author: Teresa Salazar (tmsalazar@ dei.uc.pt)

This work was supported by Fundação para a Ciência e Tecnologia (FCT) under the Project UIDB/00048/2020–ISR-Coimbra.

ABSTRACT With the increased use of machine learning algorithms to make decisions which impact people's lives, it is of extreme importance to ensure that predictions do not prejudice subgroups of the population with respect to sensitive attributes such as race or gender. Discrimination occurs when the probability of a positive outcome changes across privileged and unprivileged groups defined by the sensitive attributes. It has been shown that this bias can be originated from imbalanced data contexts where one of the classes contains a much smaller number of instances than the other classes. It is also important to identify the nature of the imbalanced data, including the characteristics of the minority classes' distribution. This paper presents FAWOS: a Fairness-Aware oversampling algorithm which aims to attenuate unfair treatment by handling sensitive attributes' imbalance. We categorize different types of datapoints according to their local neighbourhood with respect to the sensitive attributes, identifying which are more difficult to learn by the classifiers. In order to balance the dataset, FAWOS oversamples the training data by creating new synthetic datapoints using the different types of datapoints identified. We test the impact of FAWOS on different learning classifiers and analyze which can better handle sensitive attribute imbalance. Empirically, we observe that this algorithm can effectively increase the fairness results of the classifiers while not neglecting the classification performance. Source code can be found at: https://github.com/teresalazar13/FAWOS

INDEX TERMS Classification bias, fairness, imbalanced data, K-nearest neighborhood, oversampling

I. INTRODUCTION

The growing use of automated decision-making to make decisions in domains with high societal impact resulted in the need of developing Fairness-Aware methodologies which ensure that predictions are discrimination free. This bias occurs when decisions are made on the basis of sensitive attributes (e.g. race), containing one or more privileged and unprivileged groups (e.g. White and Black), where the privileged groups tend to receive more positive outcomes by the algorithm.

There have been studies which show that the reason for discrimination of unprivileged groups can be related to Imbalanced Data contexts, where one of the classes contains a much smaller number of instances than the other classes [1], [2]. Furthermore, it has been shown that the degradation of performance in imbalanced datasets can be linked to data distribution factors such as the presence of many minority datapoints inside the majority class regions [3]. In particularly, in [4] the authors demonstrate that most of the data difficulty factors can be approximated by analyzing the local neighbourhood of minority datapoints.

In this work we propose FAWOS: a novel Fairness-Aware algorithm which attenuates discrimination by balancing the dataset with respect to multiple sensitive attributes. FAWOS categorizes each point in the dataset according to the local neighbourhood as in [4]. However, we extend this work to the fairness domain, considering multiple sensitive attributes. FAWOS classifies each point as: *Safe*, *Borderline*, *Rare*, or *Outlier*, where last three types correspond to unsafe datapoints that are more difficult to learn [4].

Afterwards, FAWOS performs an oversampling technique to create new synthetic datapoints which belong to positive unprivileged groups in order to balance the dataset.

The associate editor coordinating the review of this manuscript and approving it for publication was Senthil Kumar.

In addition, we take into account the typology label and try to create new datapoints around the regions which contain datapoints that are more difficult to learn.

The proposed approach is evaluated on two real-world datasets which are commonly used in the literature [5], [6] for fairness-related studies: Ricci dataset [7] and German Credit dataset [8]. The Ricci dataset contains data on firefighter promotion exams as part of the Ricci v. DeStafano court case [7], with race being considered a sensitive attribute. The German Credit dataset [8] classifies people as good or bad credit risks, containing two sensitive attributes: gender and age.

We analyze the impact of FAWOS on a broad range of classifiers, namely: Support Vector Machines (SVM) [9], Gaussian Naive-Bayes (GNB) [10], Decision Trees (DT) [11], Logistic Regression (LR) [12], and K-Nearest-Neighbours (KNN) [13]. We measure their performance and fairness by checking whether privileged and unprivileged groups are treated equally by the classifiers. The main objective is that the classifiers' predictions of positive outcomes are similar for both privileged and unprivileged groups. Therefore, this work aims to address the following research question:

> *Is it possible to increase the fairness of a classifier without degrading the classification performance by oversampling datapoints according to the distribution of sensitive attributes?*

We conclude that FAWOS can effectively increase the fairness results while maintaining the classifiers' performance. Regarding the different classifiers' results, we observe that KNN and GNB seem to be more affected by imbalanced datasets. Furthermore, we analyzed the impact of oversampling different types of datapoints (*Safe*, *Borderline*, *Rare*) in different proportions and show that it is possible to tune FAWOS's hyperparameters depending on the datasets sensitive attributes' distributions to achieve better results.

The main contributions of FAWOS are the following:

- The identification of different types of datapoints according to the local characteristics on the basis of multiple sensitive attributes which can be binary or multi-valued.
- The ability of attenuating class imbalance by oversampling the dataset through the creation of new datapoints which belong to unprivileged groups and have a positive outcome.
- The possibility of adjusting FAWOS's hyper-parameters to control the impact of oversampling different types of datapoints and the ratio of the number of samples in the unprivileged group over the number of samples in the privileged group after oversampling.
- The analysis of the impact of FAWOS on multiple classifiers in terms of Fairness and classification performance.
- The improvement of Fairness results without compromising classification performance.

The remainder of this work is structured as follows: Section 2 provides some background and literature review on the subject, Section 3 describes the proposed approach, Section 4 presents the experimental setup, Section 5 provides the results and discussion, and Section 6 presents the conclusions and possible future research.

II. RELATED WORK

This section presents the related work to FAWOS, which belongs to the group of Fairness-Aware algorithms, although imbalanced data techniques can have similar concepts to these methodologies. While imbalance data algorithms aim to ensure that the classifiers are not biased towards the majority group, fairness methodologies such as FAWOS have the objective of certifying that the probability of a positive outcome does not change across privileged and unprivileged groups.

A. IMBALANCED DATA

There are three main approaches to tackle problems of imbalanced data: preprocessing methods, algorithm-level methods, and hybrid methods. While preprocessing methods aim to modify the data to balance distributions or remove difficult samples, algorithm-level methods have the objective of modifying existing classifiers and alleviating the bias towards majority class. Hybrid-methods combine the advantages of the two methods [14].

Since FAWOS is a preprocessing method, our focus is on these types of algorithms. Preprocessing methods can generate new datapoints belonging to the minority class (oversampling [15], [16]) or remove datapoints from the majority class (undersampling [17], [18]). Random undersampling is a technique that randomly removes datapoints from the majority class. Similarly, random oversampling aims to randomly duplicate datapoints belonging to the minority class, which can lead to overfitting [19]. To solve this issue, SMOTE (synthetic minority oversampling technique) [20] synthesizes new examples for the minority class by interpolation using the k-nearest minority class neighbours. Oversampling methodologies are typically more used in the literature than undersampling since they are capable of balancing class distributions without discarding potentially important majority examples [21].

Furthermore, in the work in [4], the authors show that classifiers learnt from imbalanced data may be deteriorated by the presence of data difficulty factors. They propose a method for their identification by analyzing the local neighbourhood of minority datapoints and consider four types of minority class examples: *Safe*, *Borderline*, *Rare* and *Outlier*. The last three types correspond to unsafe datapoints that are more difficult to learn.

Similarly to the work in [22], FAWOS applies the algorithm in [4] to identify the different types of examples and use this information to oversample the dataset by creating new datapoints using SMOTE's interpolation technique. While their algorithm experiments with different configurations that determine whether or not datapoints from a specific type should be used to create new datapoints, FAWOS introduces

hyperparameters for each of the different types to control the probability of datapoints belonging to each type to be used for oversampling. Furthermore, FAWOS presents an additional hyperparameter to control the amount of new points to be generated. Finally, while their work uses minority class datapoints to oversample the dataset, FAWOS considers the sensitive attributes and target class imbalance when oversampling the dataset.

B. FAIRNESS

Fairness definitions can be divided into two main categories: group fairness and individual fairness [23]. Group fairness states that different groups are treated equally by optimizing metrics such as Disparate Impact [5], [24]. On the other hand, individual fairness aims to give similar predictions to similar individuals. In this work, we focus on optimizing group fairness to attenuate classification bias by ensuring that unprivileged and privileged groups receive positive outcomes at equal rates.

Broadly, Fairness-Aware algorithms have been categorized into three different groups, according to the stage in which they are performed: preprocessing [25]–[27], in-processing [2], [28] and post-processing [29], [30]. The algorithm proposed in this work falls into the preprocessing category which aims to modify the training data so that an algorithm does not discriminate unprivileged groups. The advantage of these algorithms is that they are flexible since they are independent of the classification algorithm used afterwards.

Several Fairness-Aware preprocessing techniques have been proposed in the literature, including: resampling [26], [31], [32], relabelling [26], [33], transformation/perturbation [5], [34], and adversarial learning [27], [35]. The work that most relates to ours is the resampling work in [26], where the authors balance the dataset with respect to sensitive attributes in order to attenuate classification bias. However, their oversampling technique is achieved through duplication of datapoints, making them highly focused in a small area, which can lead to overfitting [19].

In addition, many of the current Fairness-Aware approaches (including the work in [26]) are unable to formally handle multiple sensitive directly in the algorithm [36]. FAWOS can handle multiple sensitive attributes combined by integrating such information on neighbourhood analysis to provide a more powerful fair methodology.

Moreover, most of the approaches evaluate fairness on a small number of classifiers. Hence, in this work we analyze the impact of FAWOS on a broad range of learning classifiers and indicate which can benefit more from Fairness-Aware methodologies.

III. FAWOS

In this section, we describe our proposed Fairness-Aware algorithm: FAWOS. We assume that each training dataset, D, contains:

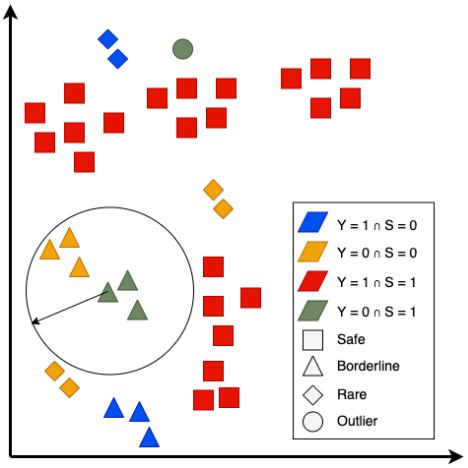

FIGURE 1. Diagram of typology and sensitive labels. The datapoint in the center of the neighbourhood region is *Borderline* since it has 2 datapoints (green) of the same sensitive and target classes against 3 (yellow) of different sensitive attribute but same target class.

- S - the set of sensitive attributes (e.g. gender and race) with size M where each S_i (e.g. race) contains privileged attributes (e.g. White, male) represented as 1 and unprivileged attributes (e.g. Black, Hispanic, female) represented as 0
- CS - the set of combination of sensitive attributes where each combination contains at least one unprivileged attribute (e.g. Black male)
- Y - a target class where 1 is the positive class and 0 is the negative class (e.g. receiving credit or not)
- $\hat{Y}$ - the predicted class where 1 is the positive class and 0 is the negative class

We hypothesize that by transforming the ratio between the positive unprivileged (PU) and the negative datapoints (NU) to be the same as the ratio between the positive privileged (PP) and the negative (NP) datapoints we can prevent unfair treatment. Hence, FAWOS's objective is to satisfy this condition:

$$\frac{P(Y = 1 \wedge S = 1)}{P(Y = 0 \wedge S = 1)} \approx \frac{P(Y = 1 \wedge S = 0)}{P(Y = 0 \wedge S = 0)}. \quad (1)$$

We show that this condition is not satisfied in D and balance the dataset by creating new synthetic datapoints which belong to $Y = 1 \wedge S = 0$. Since there can be multiple sensitive attributes, we generate all combinations of sensitive attributes' values, CS, where each combination contains at least one unprivileged attribute (e.g. Black male).

$$CS = S_0 \times .. \times S_M, \text{ where } S_i \in \{0, 1\}, \exists S_i : S_i = 0. \quad (2)$$

For each combination of sensitive attributes, CS_i, we calculate the number of synthetic points to generate, N, which belong to the positive class ($Y = 1$) and at least one sensitive unprivileged attribute, as follows:

$$PP = D(Y = 1 \wedge S = 1) \quad (3)$$

$$NP = D(Y = 0 \wedge S = 1) \quad (4)$$

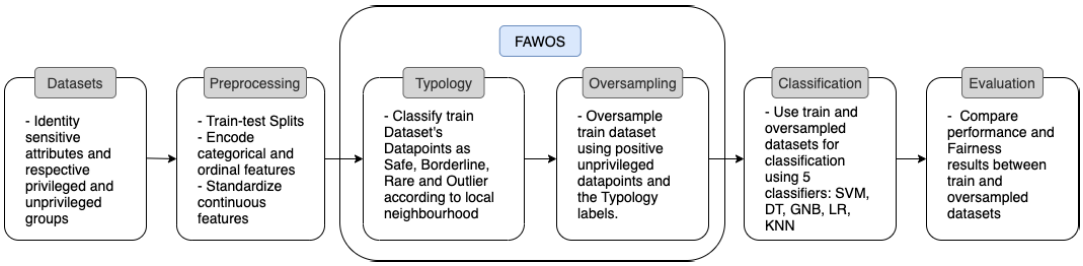

FIGURE 2. Diagram of the experimental setup.

$$PU = D(Y = 1 \wedge S = CS_i) \tag{5}$$

$$NU = D(Y = 0 \wedge S = CS_i) \tag{6}$$

$$N = \alpha * \left(\frac{|PP| * |NU|}{|NP|} - |PU| \right), \tag{7}$$

where α is the oversampling factor which is a configurable parameter used to control the number of points to generate. As a consequence, in case $\alpha = 1$, the ratio will be the same; in case $\alpha < 1$, the ratio of unprivileged attributes will be smaller, and in case $\alpha > 1$, the ratio of unprivileged attributes will be higher than the ratio of privileged attributes.

After calculating the value of N, we consider the typology of points that belong to PU. We categorize these points according to the classes' distribution in the local neighbourhood of each datapoint. This technique is based on the work in [4], which analyses the class labels of each point's k-nearest neighbours using a distance function and calculates the proportion of neighbours from the same class against neighbours from the opposite class. The difference in this work is that we do not only consider the target classes of the neighbourhood, but we also consider the sensitive attributes values. Hence, a point can only be considered a neighbour of another if they have the same value of the target class and the same values of all the sensitive attributes.

With regard to the neighbourhood, we consider $k = 5$ as in the paper in the original paper [4], where the authors state that smaller values might result in difficulties distinguishing the types of datapoints and higher values contradict their assumption of the locality of the method. Hence, the proportion from neighbours from the same classes compared to the opposite classes can vary from 5:0 (all neighbours belong to the same target class and sensitive attributes as the analyzed datapoint) to 0:5. Depending on this proportion, we assign the labels to datapoints follows:

- 5:0 or 4:1 - *Safe* datapoint.
- 3:2 or 2:3 – *Borderline* datapoint.
- 1:4 – *Rare* datapoint, in case its neighbour has the proportion of neighbours of 0:5 or 1:4. In addition, in case it is 1:4, the neighbour must be the considered datapoint. Otherwise, it is considered to be a *Borderline* datapoint.
- 0:5 – *Outlier* datapoint.

We use the Heterogeneous Euclidean Overlap Metric (HEOM) [37] as the distance metric which can handle both continuous and nominal attributes.

After labeling the datapoints, FAWOS generates N random points that belong to PU. To create a new point, FAWOS first selects a point, P, by performing a weighted random selection of a datapoint belonging to PU, where each point has a certain probability (weight) of being chosen depending on its typology label. These probabilities are defined as S_w, B_w, R_w and O_w, being hyper-parameters of FAWOS that can be configured and tuned to provide optimal results.

After selecting P, FAWOS creates a new synthetic datapoint by interpolation such as SMOTE [20], using P and one random neighbour of P. Note that since *Outlier* datapoints have no neighbours, then the value of O_w always has to be set to 0.

Algorithm 1 presents the pseudo-code of FAWOS.

IV. EXPERIMENTAL SETUP

In this section, we present the datasets, classifiers and evaluation metrics used in our experiments. The summary of the experimental setup is described in Figure 2.

A. DATA COLLECTION

Our experiments are performed on two fairness-related datasets which contain sensitive attributes:

a: RICCI DATASET

This dataset is part of the Ricci v. DeStafano court case [7], containing 118 entries which consider if firefighters should receive a promotion. The promotion was given to firefighters if they achieved a minimum combined score of 70 on certain exams. Within its features, it contains the race, which is considered to be a sensitive attribute with White (W) being the privileged group and Black (B) and Hispanic (H) the unprivileged groups.

b: GERMAN CREDIT DATASET

This dataset contains 1000 credit records which consider individuals as having good or bad credit risk [8]. It has 20 features with the sensitive attributes being the gender and age. The sensitive attribute age was converted into a categorical feature by considering the value of Adult (A) (when the age is equal or more than 25 years old) and Youth (Y). This conversion was based on the work in [25] which proves that this provided the most discriminatory effects. Adult is considered to be the privileged group and Youth is considered

Algorithm 1 FAWOS

procedure OVERSAMPLE(D)
  $PP \leftarrow D(Y = 1 \wedge S = 1)$
  $NP \leftarrow D(Y = 0 \wedge S = 1)$
  $CS \leftarrow S_0 \times .. \times S_M$, where $S_i \in \{0, 1\}, \exists S_i : S_i = 0$
  for each $CS_i \in CS$ **do**
    $PU \leftarrow D(Y = 1 \wedge S = CS_i)$
    $NU \leftarrow D(Y = 0 \wedge S = CS_i)$
    $weights \leftarrow []$
    for each $PU_i \in PU$ **do**
      $weight \leftarrow GetTypologyWeight(PU_i)$
      $weights.insert(weight)$
    $N \leftarrow \alpha * \left(\frac{|PP|*|NU|}{|NP|} - |PU|\right)$
    for N points **do**
      $P \leftarrow RandomWeightedSelect(PU, weights)$
      $neighbour \leftarrow random.select(5Closest(P))$
      $newPoint \leftarrow Interpolate(P, neighbour)$
      $D.insert(newPoint)$
  return D ▷ The Oversampled Dataset

procedure GET TYPOLOGY WEIGHT(P)
  $N \leftarrow NeighboursOfSameTypes(P)$
  if $|N| = 5$ **or** $|N| = 4$ **then**
    $weight \leftarrow S_w$
  else if $|N| = 3$ **or** $|N| = 2$ **then**
    $weight \leftarrow B_w$
  else if $|N| = 1$ **then**
    $NN \leftarrow NeighboursOfSameTypes(N_0)$
    if $|NN| = 0$ **or** ($|NN| = 1$ **and** $NN_0 = P$) **then**
      $weight \leftarrow R_w$
    else
      $weight \leftarrow B_w$
  else
    $weight \leftarrow 0$
  return weight

procedure NEIGHBOURS OF SAME TYPES(P)
  $N \leftarrow 5Closest(P)$
  $neighbours \leftarrow []$
  for each $N_i \in N$ **do**
    if $N_i^Y = P^Y$ **and** $N_i^S = P^S$ **then**
      $neighbours.insert(N_i)$
  return $neighbours$

procedure 5CLOSEST(P)
  return The 5 closest nearest neighbours of P

procedure RANDOM WEIGHTED SELECT(PS, WS)
  return An element from PS with probabilities/weights of each element being selected.

procedure INTERPOLATE(P, N)
  return A new point by Interpolation of P and N

to be the unprivileged group. In addition, the gender feature was generated from the personal status feature since it was not directly included in the dataset. In the gender feature, Male (M) is the privileged group and Female (F) the unprivileged group.

Each dataset is divided into 70% and 30% for training and testing. We report the average performance results of running 10 different training-test splits. Both datasets are tested on the same test dataset.

The datasets' features were categorized into continuous, ordinal or categorical. The continuous features were standardized using scikit-learn's StandardScaler [38]. Ordinal features were converted into ordinal integers following a predefined order and categorical features were label encoded.

B. CLASSIFICATION

We wanted to analyze the performance and fairness results of a broad range of classifiers when applied to fairness-related datasets. In addition, we wanted to investigate which classifiers benefit more from FAWOS. As such, we test the performance and fairness on three different datasets:

- the train dataset
- the oversampled dataset generated through a simple random fair oversampling (RFO) algorithm presented in Algorithm 2. This algorithm was developed in the scope of this work.
- the oversampled dataset generated through FAWOS.

The first two datasets are used as the baseline for this study.

Algorithm 2 Random Fair Oversamplor (RFO)

procedure OVER SAMPLE(D)
  $PP \leftarrow D(Y = 1 \wedge S = 1)$
  $NP \leftarrow D(Y = 0 \wedge S = 1)$
  $CS \leftarrow S_0 \times .. \times S_M$, where $S_i \in \{0, 1\}, \exists S_i : S_i = 0$
  for each $CS_i \in CS$ **do**
    $PU \leftarrow D(Y = 1 \wedge S = CS_i)$
    $NU \leftarrow D(Y = 0 \wedge S = CS_i)$
    $N \leftarrow \left(\frac{|PP|*|NU|}{|NP|} - |PU|\right)$
    for N points **do**
      $P \leftarrow RandomSelect(PU)$
      $D.insert(P)$
  return D ▷ The Oversampled Dataset

For classification we used scikit-learn's classifiers [38], namely: Decision Trees (DT) [11], Logistic Regression (LR) [12], Support Vector Machines (SVM) [9], K-Nearest-Neighbours (KNN) [13], and Gaussian Naive-Bayes (GNB) [10]. All classifiers were tuning using grid search on the hyperparameters presented in Table 1.

C. EVALUATION

For performance evaluation, the objective is to check if our oversampling technique can increase the fairness results while not neglecting the classifiers' performances. For measuring the performance, the standard (uniform) accuracy

TABLE 1. Experimental setup of classifiers' hyperparameters.

Classifier	Hyperparameters
DT	min-samples-split: { 0.1, 0.325, 0.55, 0.775, 1 } min-samples-leaf: { 0.1, 0.2, 0.3, 0.4, 0.5 } criterion: {'gini', 'entropy'}
LR	C: logspace(-4, 4, 20) penalty: { 'l1', 'l2' } solver: { 'newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga' }
SVM	C: { 0.1, 1, 10, 100 } gamma: { 'scale', 'auto' } kernel: { 'rbf', 'poly', 'sigmoid' }
KNN	n-neighbours: { 3, 5, 11, 19 } weights: { 'uniform', 'distance' } metric: { 'euclidean', 'manhattan' }
GNB	var-smoothing: logspace(0, -9, 100)

(ACC) is used, given by:

$$ACC = P[\hat{Y} = y \mid Y = y]. \tag{8}$$

With regard to fairness metrics, a common used metric to evaluate group fairness is the Disparate Impact (*DI*) [5], [24], given by:

$$DI = \frac{P[\hat{Y} = 1 \mid S = 0]}{P[\hat{Y} = 1 \mid S = 1]}. \tag{9}$$

This metric compares the proportion of individuals that were predicted to receive a positive output for two groups: an unprivileged group and a privileged group. We propose an adaptation of DI, Adapted Disparate Impact (*ADI*), given by:

$$ADI = \begin{cases} DI & \text{if } DI \leq 1 \\ \frac{1}{DI}, & \text{otherwise.} \end{cases} \tag{10}$$

When the value of *DI* is bigger than 1, it means that the unprivileged class became the privileged class. As such, values of 0.5 (where the proportion of unprivileged individuals is twice as small) and 2 (the proportion is twice as large) of *DI*, would have the same value of *ADI*: 0.5. This is useful since we are expecting values smaller than 1 for *DI* using the train dataset and values around 1 for *DI* using the oversampling dataset. In addition, in the cases where there is more than one sensitive attribute being considered, we calculate the average value of *ADI* of all the sensitive attributes.

However, *DI* may never be aligned with a perfect predictor $\hat{Y} = Y$ [39]. There are other fairness metrics which consider an algorithm to be fair if it is independent of the sensitive attribute while conditioned on Y. For instance, Average Absolute Odds Difference (*AOD*) [40] represents the average of absolute difference in *FPR* and *TPR* for unprivileged and privileged groups, given by:

$$AOD = \frac{|FPR_{S=0} - FPR_{S=1}| + |TPR_{S=0} - TPR_{S=1}|}{2} \tag{11}$$

where $FPR_{S=s}$ and $TPR_{S=s}$ are given by:

$$FPR_{S=s} = P[\hat{Y} = 1 \mid S = s, Y = 0] \tag{12}$$

$$TPR_{S=s} = P[\hat{Y} = 1 \mid S = s, Y = 1] \tag{13}$$

TABLE 2. Typology of sensitive attributes distribution values in the ricci dataset.

	S %	B %	R %	O %	Total %	Ratio +/-
W + W -	27 11	5 11	0 0	0 1	32 23	1.39
B + B -	0 10	10 6	0 0	0 0	10 16	0.63
H + H -	0 7	4 6	2 0	0 0	6 13	0.46

TABLE 3. Typology of sensitive attributes distribution values in the german credit dataset.

	S %	B %	R %	O %	Total %	Ratio +/-
A M + A M -	33 1	14 10	0 3	0 3	47 17	2.75
A F + A F -	3 0	10 2	1 1	1 2	15 5	3
Y M + Y M -	0 0	3 1	0 0	1 1	4 2	2
Y F + Y F -	1 0	4 3	0 0	0 0	5 3	1.67

TABLE 4. Experimental setup.

Variables	Settings
Datasets	Ricci German Credit
Classifiers (and Hyperparameters from Table 1)	SVM, GNB, DT, LR, KNN
Typology Weights (S_w, B_w, R_w)	$0, 0.4, 0.6$ $0, 0.5, 0.5$ $0.33, 0.33, 0.33$ $0, 0.6, 0.4$
Oversampling Factor (α)	Varied from 0.6 to 1.4

A value of 0 indicates equality of odds. Furthermore, in the cases where there is more than one sensitive attribute being considered, we calculate the average value of *AOD* of all the sensitive attributes.

V. RESULTS AND DISCUSSION

In this section we present the experimental results from following the setup described in the previous section.

The experiments can be divided into two main parts. In the first part, we analyze the typology and sensitive classes' distributions in the Ricci and German Credit datasets before oversampling. In the second part, we analyze the Fairness and performance results of FAWOS and the baseline. In addition, we investigate which are the most robust configurations of Typology Weights (S_w, B_w, R_w) and Oversampling Factor (α) providing the best results for each classifier in particular.

A. DISTRIBUTION OF TYPOLOGIES AND SENSITIVE ATTRIBUTES

In this section, we analyze the sensitive attributes' distributions of the Ricci and German Credit datasets before

TABLE 5. Performance classification and fairness results for SVM, DT, GNB, LR, KNN on the ricci dataset.

		Baseline	RFO	$S_w = 0$ $B_w = 0.4$ $R_w = 0.6$	$S_w = 0$ $B_w = 0.5$ $R_w = 0.5$	$S_w = 0.33$ $B_w = 0.33$ $R_w = 0.33$	$S_w = 0$ $B_w = 0.6$ $R_w = 0.4$	p-value
SVM	α			0.6	0.6	0.6	0.9	
	ACC	0.94	**0.96**	0.94	0.95	0.95	0.95	
	ADI	0.31	0.34	**0.40**	0.35	0.35	0.33	
	AOD	0.09	**0.03**	0.05	0.06	0.06	0.05	0.086
DT	α			0.9	0.7	1.0	0.7	
	ACC	**0.87**	0.82	**0.87**	0.85	0.83	0.85	
	ADI	0.33	0.43	0.41	0.44	**0.47**	**0.47**	
	AOD	0.20	0.22	**0.14**	0.15	0.16	0.16	0.048
GNB	α			1.0	1.0	1.0	1.2	
	ACC	0.86	0.85	0.84	**0.87**	**0.87**	0.86	
	ADI	0.18	**0.47**	0.39	0.40	0.40	0.42	
	AOD	0.31	0.15	0.18	**0.14**	**0.14**	0.17	0.45
LR	α			0.6	0.6	1.3	0.7	
	ACC	**0.99**	0.98	**0.99**	0.98	**0.99**	**0.99**	
	ADI	0.34	0.33	**0.35**	0.34	0.33	**0.35**	
	AOD	**0.01**	0.02	0.02	0.02	0.02	**0.01**	-
KNN	α			1.4	1.4	0.7	0.7	
	ACC	**0.89**	0.88	**0.89**	0.88	**0.89**	**0.89**	
	ADI	0.14	**0.37**	0.31	0.34	0.28	0.26	
	AOD	0.35	0.25	0.24	0.20	**0.18**	0.19	0.180

oversampling. To this end, we investigate whether or not the dataset is imbalanced by calculating the ratio presented in Equation 1. In addition, we explore the percentage of datapoints which belong to each typology label.

Tables 2 and 3 present the distribution of typology labels for each sensitive attribute combination in the Ricci and German Credit datasets before the Oversampling stage, respectively.

Regarding the Ricci dataset, it can be observed that the privileged attribute, White, is the most represented in total, containing a total of 55% of the datapoints. In addition, almost all of its positive points are labeled as *Safe*, which suggests that they are easier to be learned by the models. With regard to the Black and Hispanic unprivileged attributes, it can be observed that there are more negative datapoints than positive, contrary to the White privileged attribute. In addition, there are also some Hispanic *Rare* datapoints within the positive target class, which means that they should be more difficult to learn.

Looking at Table 3, we can observe that the majority of points belong to the combination of privileged attributes Adult Male positive, where almost all the datapoints belong to the *Safe* and *Borderline* labels. However, in this dataset, the differences of positive/negative ratios between the combinations of privileged and unprivileged attributes are not as noticeable as in the Ricci dataset. Furthermore, one particular fact that stands out is the fact that the ratio of Adult Female is bigger than the ratio of Adult Male, whereas the ratio of Young Female is smaller than the ratio of Young Male. This means that considering combinations of several sensitive attribute values is important since in this case it is clear that the most discriminated group is the Young Female group.

B. FAIRNESS OF THE CLASSIFIERS

In the second part of the experiments, we compare the performance and fairness results between the train dataset (baseline) and the oversampled datasets (RFO and FAWOS). The summary of the experimental setup can be observed in Table 4. For each dataset and classifier in the table, we report the optimal Oversampling Factor, α, for each of the Typology Weights' configurations presented, which provide the best results in terms of *AOD* when compared to the baseline. In case of a tie, we report the result with the best values of *ADI* and *ACC*. In addition, if there are multiple α values providing similar results, we report the results of the lowest α. The values of α were varied from 0.6 to 1.4. These threshold values were obtained through experimentation and values outside this range did not yield better results. We run each configuration 10 times, and report the average results of *ACC*, *ADI* and *AOD*.

Looking at the results, it can be observed that FAWOS can reduce unfair treatment of unprivileged groups as the values of *AOD* are closer to 0 and the values of *ADI* are closer to 1, when compared to the baseline and RFO in almost all the setups. In addition, we can also state that the values of *ACC* remain closer to the values of the baseline, which means that FAWOS does not degrade the classification performance when increasing the fairness.

With regard to the classifiers' results in the Ricci dataset, it can be observed that the classifiers which seem to be most affected in terms of fairness (*ADI* and *AOD*) by imbalanced data are KNN, GNB and DT. Furthermore, the classifiers which seem to benefit more from FAWOS in terms of fairness are KNN, GNB and DT. KNN uses the k-closest training datapoints in the feature space to make the classification and since

TABLE 6. Performance classification and fairness results for SVM, DT, GNB, LR, KNN on the german credit dataset.

		Baseline	RFO	$S_w = 0$ $B_w = 0.4$ $R_w = 0.6$	$S_w = 0$ $B_w = 0.5$ $R_w = 0.5$	$S_w = 0.33$ $B_w = 0.33$ $R_w = 0.33$	$S_w = 0$ $B_w = 0.6$ $R_w = 0.4$	p-value
SVM	α			1.0	0.9	1.4	0.9	
	ACC	0.73	0.73	0.73	**0.74**	0.73	**0.74**	
	ADI	0.89	0.90	0.95	0.95	0.93	**0.96**	
	AOD	0.09	0.07	**0.05**	0.06	**0.05**	**0.05**	0.035
DT	α			1.3	1.1	1.2	1.4	
	ACC	**0.69**	**0.69**	**0.69**	**0.69**	**0.69**	**0.69**	
	ADI	0.93	0.96	**0.97**	0.96	0.96	**0.97**	
	AOD	0.07	0.03	**0.02**	0.03	0.03	**0.02**	0.234
GNB	α			1.2	0.9	0.7	0.7	
	ACC	0.70	0.70	**0.71**	**0.71**	**0.71**	**0.71**	
	ADI	0.91	**0.97**	**0.97**	**0.97**	**0.97**	**0.97**	
	AOD	0.11	**0.04**	0.05	0.05	0.05	0.05	0.153
LR	α			0.9	0.8	0.8	0.8	
	ACC	**0.72**	0.71	0.71	0.71	**0.72**	0.70	
	ADI	0.92	0.97	**0.98**	**0.98**	0.97	**0.98**	
	AOD	0.08	0.05	**0.04**	0.05	0.05	**0.04**	0.030
KNN	α			1.2	0.8	1.4	1.4	
	ACC	**0.71**	0.70	**0.71**	**0.71**	**0.71**	0.70	
	ADI	0.88	0.91	0.93	**0.94**	**0.94**	0.93	
	AOD	0.10	0.09	**0.07**	**0.07**	0.08	0.08	0.036

FAWOS uses the k-nearest neighbours to create new synthetic datapoints then it is clear that KNN achieves better results. GNB is a probabilistic classifier with strong independence assumptions between the features. Hence, creating new positive unprivileged synthetic datapoints can help the algorithm increase its fairness. With regard to Decision Trees, although its fairness results increased substantially, its accuracy seems to have degraded. SVM fairness results also improved. With regard to the LR classifier, almost no improvements were registered. Concerning the typology weights and α values configurations, it is difficult to reach conclusions since it is not possible to find a common pattern for all classifiers. This might have to do with the fact that the Ricci dataset contains very few *Rare* datapoints.

Regarding the German Credit dataset, the first conclusion is that the classifiers in the baseline were able to provide good fairness results. This might have to do with the fact that this dataset contained more datapoints. Nevertheless, the fairness results of *AOD* achieved with FAWOS are almost optimal, reaching values of 0.02 for DT, 0.04 for LR, 0.05 for GNB and SVM and 0.07 for KNN. Concerning the typology weights, it can be observed there is not much difference in the results. This probably has to do with the fact that the positive unprivileged groups in this dataset contain almost no *Rare* and *Safe* datapoints, with the majority being *Borderline*.

Finally, we performed Kolmogorov-Smirnov test of normality followed by Student's one-sided t-tests with significance level of $\alpha = 0.05$ to verify the statistical significance of FAWOS values of *AOD* compared to the best baseline (train or RFO). Looking at the p-values presented in the tables, it can be observed that in the seven times FAWOS outperformed both baselines, four were statistically significant. In addition, the two times RFO outperformed FAWOS were not statistical significant.

VI. CONCLUSION AND FUTURE WORK

We proposed FAWOS: a Fairness-Aware oversampling algorithm based on the distributions of multiple sensitive attributes' datapoints. To the best of our knowledge, this is the first work which proposes to reduce bias in machine learning by considering the typology presented in [4] adapted to Fairness.

Moreover, we evaluated the fairness performance of different classifiers on fairness-related datasets, and analyzed their improvements on the oversampled datasets. Furthermore, we observed which typology weights and oversampling factor configurations achieved the best results for each dataset. Revisiting the question presented at the start of this work:

> *Is it possible to increase the fairness of a classifier without degrading the classification performance by oversampling datapoints according to the distribution of sensitive attributes?*

We conclude that FAWOS can effectively increase the fairness results of ADI while maintaining the performance in terms of accuracy. We evaluated the performance of different classifiers (SVM, DT, GNB, LR, KNN) without applying FAWOS and conclude that KNN and GNB seem to be more affected by imbalanced datasets, presenting low fairness results. However, FAWOS is able to improve the fairness of all classifiers, in particular for KNN, GNB and DT.

In addition, by experimenting with different typology weights configurations, we concluded that some of them

resulted in better values of fairness, depending on the datasets' distributions.

There is, however, a limitation to our approach: for extremely large datasets (e.g. Adult Income [41]), our algorithm becomes infeasible when calculating the distances between all datapoints. As such, we propose to applying approximation distance function to FAWOS. In addition, future work comprises the following:

- Investigating other values of k with regard to the local neighbourhood.
- Exploring other approximation distance functions for calculating the distances between all datapoints such as the Heterogeneous Value Difference Metric (HVDM) [37].

We hope that this work inspires the community to develop more Fairness-Aware algorithms and consider sensitive attributes' data distribution patterns for attenuating classification bias.

REFERENCES

[1] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proc. 1st Conf. Fairness, Accountability Transparency*, vol. 81, S. A. Friedler and C. Wilson, Eds., New York, NY, USA, Feb. 2018, pp. 77–91.

[2] V. Iosifidis and E. Ntoutsi, "AdaFair," in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manage.*, Nov. 2019, pp. 781–790.

[3] K. Napierała, J. Stefanowski, and S. Wilk, "Learning from imbalanced data in presence of noisy and borderline examples," in *Proc. 7th Int. Conf. Rough Sets Current Trends Comput. (RSCTC)*. Berlin, Germany: Springer-Verlag, 2010, pp. 158–167.

[4] K. Napierala and J. Stefanowski, "Types of minority class examples and their influence on learning classifiers from imbalanced data," *J. Intell. Inf. Syst.*, vol. 46, no. 3, pp. 563–597, Jun. 2016.

[5] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, Aug. 2015, pp. 259–268.

[6] F. Calmon, D. Wei, B. Vinzamuri, K. Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2017, pp. 1–10.

[7] *Ricci v. DeStefano. 557 U.S. 557, 174*, Supreme Court of the United States, Washington, DC, USA, 2009.

[8] D. Dua and C. Graf, *German Credit Data Set—UCI Machine Learning Repository*. Irvine, CA, USA: Univ. of California, Irvine, School of Information and Computer Sciences, 2007.

[9] M. A. Hearst, S. T. Dumais, E. Osman, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intell. Syst. Appl.*, vol. 13, no. 4, pp. 18–28, Jul./Aug. 2008.

[10] R. Duda and P. Hart, *Pattern Classification and Scene Analysis*. Hoboken, NJ, USA: Wiley, 1973.

[11] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.

[12] J. S. Cramer, "The origins of logistic regression," Tinbergen Inst. Discuss. Papers, Amsterdam, The Netherlands, Tech. Rep. TI 2002 - 119/4, Jan. 2002.

[13] E. Fix and J. L. Hodges, "Discriminatory analysis. Nonparametric discrimination: Consistency properties," *Int. Stat. Rev.*, vol. 57, no. 3, pp. 238–247, 1989.

[14] N. Rout, D. Mishra, and M. K. Mallick, "Handling imbalanced data: A survey," in *Proc. Int. Proc. Adv. Soft Comput., Intell. Syst. Appl.*, M. S. Reddy, K. Viswanath, and K. M. S. Prasad, Eds. Singapore: Springer, 2018, pp. 431–443.

[15] X. Tao, Q. Li, W. Guo, C. Ren, Q. He, R. Liu, and J. Zou, "Adaptive weighted over-sampling for imbalanced datasets based on density peaks clustering with heuristic filtering," *Inf. Sci.*, vol. 519, pp. 43–73, May 2020.

[16] X. Tao, Q. Li, C. Ren, W. Guo, C. Li, Q. He, R. Liu, and J. Zou, "Real-value negative selection over-sampling for imbalanced data set learning," *Expert Syst. Appl.*, vol. 129, pp. 118–134, Sep. 2019.

[17] C.-F. Tsai, W.-C. Lin, Y.-H. Hu, and G.-T. Yao, "Under-sampling class imbalanced datasets by combining clustering analysis and instance selection," *Inf. Sci.*, vol. 477, pp. 47–54, Mar. 2019.

[18] A. D. Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi, "Calibrating probability with undersampling for unbalanced classification," in *Proc. IEEE Symp. Ser. Comput. Intell.*, Dec. 2015, pp. 159–166.

[19] M. Koziarski, B. Krawczyk, and M. Woźniak, "Radial-based approach to imbalanced data oversampling," in *Hybrid Artificial Intelligent Systems*, F. J. M. de Pisón, R. Urraca, H. Quintián, and E. Corchado, Eds. Cham, Switzerland: Springer, 2017, pp. 318–327.

[20] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[21] M. S. Santos, J. P. Soares, P. H. Abreu, H. Araujo, and J. Santos, "Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches [Research frontier]," *IEEE Comput. Intell. Mag.*, vol. 13, no. 4, pp. 59–76, Nov. 2018.

[22] J. A. Sáez, B. Krawczyk, and M. Woźniak, "Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets," *Pattern Recognit.*, vol. 57, pp. 164–178, Sep. 2016.

[23] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," 2019, *arXiv:1908.09635*. [Online]. Available: http://arxiv.org/abs/1908.09635

[24] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," 2015, *arXiv:1507.05259*. [Online]. Available: http://arxiv.org/abs/1507.05259

[25] F. Kamiran and T. Calders, "Classifying without discriminating," in *Proc. 2nd Int. Conf. Comput., Control Commun.*, Feb. 2009, pp. 1–6.

[26] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowl. Inf. Syst.*, vol. 33, no. 1, pp. 1–33, 2012.

[27] R. Feng, Y. Yang, Y. Lyu, C. Tan, Y. Sun, and C. Wang, "Learning fair representations via an adversarial framework," 2019, *arXiv:1904.13341*. [Online]. Available: http://arxiv.org/abs/1904.13341

[28] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proc. AIES*, New York, NY, USA, pp. 335–340, 2018.

[29] P. Awasthi, M. Kleindessner, and H. J. Morgenstern, "Equalized odds postprocessing under imperfect group information," in *Proc. AISTATS*, 2020, pp. 1770–1780.

[30] A. Mishler, E. H. Kennedy, and A. Chouldechova, "Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds," in *Proc. ACM Conf. Fairness, Accountability, Transparency*, New York, NY, USA, Mar. 2021, pp. 386–400.

[31] L. E. Celis, A. Deshpande, T. Kathuria, and K. N. Vishnoi, "How to be fair and diverse?" 2016, *arXiv:1610.07183*. [Online]. Available: http://arxiv.org/abs/1610.07183

[32] S. Verma, M. Ernst, and R. Just, "Removing biased data to improve fairness and accuracy," 2021, *arXiv:2102.03054*. [Online]. Available: http://arxiv.org/abs/2102.03054

[33] T. Calders and S. Verwer, "Three naive bayes approaches for discrimination-free classification," *Data Mining Knowl. Discovery*, vol. 21, no. 2, pp. 277–292, Sep. 2010.

[34] H. Wang, B. Ustun, and F. P. Calmon, "Repairing without retraining: Avoiding disparate impact with counterfactual distributions," 2019, *arXiv:1901.10501*. [Online]. Available: http://arxiv.org/abs/1901.10501

[35] S. Chiappa, "Path-specific counterfactual fairness," in *Proc. AAAI Conf. Artif. Intell.*, Jul. 2019, vol. 33, no. 1, pp. 7801–7808.

[36] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth, "A comparative study of fairness-enhancing interventions in machine learning," in *Proc. Conf. Fairness, Accountability, Transparency*, New York, NY, USA, Jan. 2019, pp. 329–338.

[37] D. R. Wilson and T. R. Martinez, "Improved heterogeneous distance functions," *J. Artif. Intell. Res.*, vol. 6, no. 1, pp. 1–34, Jan. 1997.

[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

[39] S. Liu and L. N. Vicente, "Accuracy and fairness trade-offs in machine learning: A stochastic multi-objective approach," 2020, *arXiv:2008.01132*. [Online]. Available: http://arxiv.org/abs/2008.01132
[40] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," 2016, *arXiv:1610.02413*. [Online]. Available: http://arxiv.org/abs/1610.02413
[41] A. Asuncion and D. J. Newman, *Adult Data Set—UCI Machine Learning Repository*. Berkeley, CA, USA: Univ. of California, School of Information and Computer Sciences, 2007.

TERESA SALAZAR received the B.S. degree in informatics engineering from the University of Coimbra, in 2018, and the M.S. degree in informatics from The University of Edinburgh, in 2019, with specialization in machine learning and natural language processing. She is currently pursuing the Ph.D. degree in information science and technology with the University of Coimbra.

Since 2020, she has been a member of the Centre for Informatics and Systems, University of Coimbra. Her research interests include fairness and imbalanced data, natural language processing, and information retrieval.

MIRIAM SEOANE SANTOS received the master's degree in biomedical engineering from the University of Coimbra, in 2014, where she is currently pursuing the Ph.D. degree in information science and technology.

She is a member of the Centre for Informatics and Systems, University of Coimbra. Her research interests include pattern recognition problems, imbalanced and missing data, and personalized medicine in oncology.

HELDER ARAÚJO (Member, IEEE) is currently a Professor at the Department of Electrical and Computer Engineering, University of Coimbra. He has also worked on non-central camera models, including aspects related to pose estimation and their applications. He has started work on the development of vision systems applied to medical endoscopy with focus on capsule endoscopy. His research interests include computer vision applied to robotics, robot navigation, and visual servoing.

PEDRO HENRIQUES ABREU (Member, IEEE) received the degree in informatics engineering and the Ph.D. degree in informatics engineering from the University of Porto, in 2006 and 2011, respectively.

He is currently an Assistant Professor with the Department of Informatics Engineering, University of Coimbra, and a Full Member of the Cognitive and Media System Group, Centre for Informatics and Systems, University of Coimbra. He has authored more than 80 publications in highly rated *JCR* journals and international conferences. His research interests include medical informatics and personal healthcare systems applied to oncology.