2019 IEEE International Conference on Big Data (Big Data)

FAE: A Fairness-Aware Ensemble Framework

Vasileios Iosifidis
Leibniz University Hanover
L3S Research Center
Hanover, Germany
iosifidis@L3S.de

Besnik Fetahu
Leibniz University Hanover
L3S Research Center
Hanover, Germany
fetahu@L3S.de

Eirini Ntoutsi
Leibniz University Hanover
L3S Research Center
Hanover, Germany
ntoutsi@L3S.de

***Abstract*—Automated decision making based on big data and machine learning (ML) algorithms can result in discriminatory decisions against certain protected groups defined upon personal data like gender, race, sexual orientation etc. Such algorithms designed to discover patterns in big data might not only pick up any encoded societal biases in the training data, but even worse, they might reinforce such biases resulting in more severe discrimination. The majority of thus far proposed fairness-aware machine learning approaches focus solely on the pre-, in- or post-processing steps of the machine learning process, that is, input data, learning algorithms or derived models, respectively. However, the fairness problem cannot be isolated to a single step of the ML process. Rather, discrimination is often a result of complex interactions between big data and algorithms, and therefore, a more holistic approach is required.**

The proposed `FAE` (*Fairness-Aware Ensemble*) framework combines fairness-related interventions at both pre- and post-processing steps of the data analysis process. In the pre-processing step, we tackle the problems of under-representation of the protected group (*group imbalance*) and of *class-imbalance* by generating balanced training samples. In the post-processing step, we tackle the problem of *class overlapping* by shifting the decision boundary in the direction of fairness.

***Index Terms*—fairness-aware classification, class imbalance, group imbalance, class overlap, ensemble learning.**

I. Introduction

Machine Learning powered by big data offers incredible opportunities for effective decision making and automation. However, several recent incidents have raised concerns about the implications of such systems in terms of fairness [1]. Amazon's models, to name but one example, that decide which regions of a city are eligible for the prime service, excluded predominantly black ZIP codes in several US cities, like Bronx [2]. According to Amazon, the *protected attribute* race was not used as a predictor. Nonetheless, there might exist *proxy-attributes* to race which lead to discriminatory decisions. Protected attributes and proxies are not the only causes of the problem [3]. Training data often reflect *societal biases* and are not representative of the population (*sample bias*). Moreover, *system bias* might lead into generation of biased data which result into biased models that further reinforce such discriminatory policies, like in predictive policing [4].

Despite extensive research work in the area of fairness-aware learning, most of the approaches isolate the problem and its solutions to a single step of the ML process, namely, input data, algorithms or resulting models. While, we share the view on the importance of working on the main source of bias, i.e., the training data as pointed out by recent work, e.g., [5], [6], we believe that this in itself is insufficient, and that in- and post-processing adjustments are necessary to deal with discrimination.

To this end, we propose the Fairness-Aware Ensemble (`FAE`) framework, a holistic approach that combines pre- and post-processing fairness-enhancing interventions to deal with different bias factors and real-world data complexities, namely group imbalance, class imbalance and class overlap. At pre-processing, we learn an ensemble of ensembles through a combination of bagging and boosting; the bags are carefully selected via stratified cluster sampling to ensure a balanced group- and class-representation, whereas boosting on each bag forces the classifier to focus on the hard-to-classify examples. At post-processing, the decision boundary of the learner is shifted so that the target fairness criterion is fulfilled. Our experiments show that such a joint consideration ensures better fairness- and predictive-performance.

II. Related Work

Pre-processing methods aim to tackle discrimination by "correcting" the training data to eliminate any biases. Bias can be inherited from the input data, e.g., there might exist proxies to sensitive attributes, or under-represented groups or biased class labels. Among the most popular methods in this category are class-label swapping, instance re-weighting, sampling, and instance transformation [6]–[8]. *In-processing* methods modify the learning algorithm to eliminate discriminatory behavior. These interventions are typically learner-specific [5], [9]–[12]. For instance, Zafar et al. [5] add fairness-related constraints in the objective function of a logistic regression model to account for fairness. *Post-processing* methods try to modify the model's predictions or decision boundary in order to ensure fairness [10], [13], [14]. Kamiran et al. [10] propose a fair decision tree learner that combines a fairness-aware splitting criterion with post-processing leaf-relabeling. Fish et al. [13] adjust the decision boundary of a boosting model based on the confidence scores of the misclassified instances. Finally, *class-imbalance* methods aim to deal with skewed class distributions. Over the years, many methods have been proposed such as over-sampling [15], under-sampling [16], synthetic data generation like SMOTE [17] and boosting [18].

978-1-7281-0858-2/19/$31.00 ©2019 IEEE
978-1-7281-0858-2/19/$31.00 ©2019 IEEE

III. BASIC CONCEPTS

We consider binary classification with $A = \{A_1, \ldots, A_n\}$ being the attribute space and $Y = \{y^+, y^-\}$ the class attribute. Let $dom(A_i)$ be the domain of A_i, and y^+ is the *target class*, for example, "receive a benefit". Let $SA \in A$ be a *protected attribute* with $dom(SA) = \{s, \bar{s}\}$; s is the discriminated group (referred to as *protected group*), and $\bar{s}$ is the non-discriminated group (referred to as *non-protected group*). For instance, $SA =$'*gender*' could be the protected attribute with $s =$'*female*' being the protected group and $\bar{s} =$'*male*' the non-protected. By combining sensitive attribute SA and class Y values, we define four sub-groups: s^-, s^+,$\bar{s}^-$,$\bar{s}^+$; e.g., s^- denotes the protected negative group, $\bar{s}^+$ denotes the non-protected positive group etc. We assume the following learning challenges: *class imbalance*, that is $|s^+|+|\bar{s}^+| \ll |s^-|+|\bar{s}^-|$; *group imbalance*, that is $|s^+| + |s^-| \ll |\bar{s}+| + |\bar{s}^-|$ as well as class overlap, i.e, the positive class y^+ overlaps with the negative class y^-.

The goal of a fairness-aware classifier is to learn a function $f(\cdot) : dom(A_i) \times \cdots \times dom(A_n) \rightarrow Y$, s.t. $f(\cdot)$ can generalize well to unseen instances and does not discriminate against the protected group for the target class y^+.

Discrimination measure: We adopt the *equal opportunity measure* (EQOP) [19] that compares the probability of being predicted as positive while belonging to the positive class (TPR) between protected s and non-protected $\bar{s}$ groups:

$$EQOP : P\Big(f(d) = y^+|\bar{s}^+\Big) - P\Big(f(d) = y^+|s^+\Big) \qquad (1)$$

$EQOP \in [-1, 1]$: a value close to 0 means *fair outcomes*, and is desirable, whereas a value close to 1 indicates *discriminatory* behavior towards the protected group. A value close to -1 indicates *reverse discrimination* towards the non-protected group. A classifier $f(\cdot)$ is said to *not discriminate* if: $\mid EQOP \mid \leq \epsilon$. The user-defined threshold ϵ controls how much discrepancy between the two groups is tolerated.

Predictive performance measure: The vast majority of existing works minimize the standard error rate, e.g., [5]–[7], [10], [13], which is not useful in case of *class-imbalance* as it mainly reflects the performance of the model in the majority class. Moreover, EQOP measure, (c.f., Equation1) which relies on the TPR difference, is oblivious to the problem of class imbalance. As an extreme case, if a classifier totally rejects the minority (positive) class and correctly classifies the majority (negative) class then, based on EQOP, the classifier is both fair (in terms of EQOP) and accurate (in terms of error rate). Recent methods fall in this pitfall and their low reported discrimination scores are mainly due to low TPR values (c.f., Section VI). Hence, we use balanced accuracy [20]:

$$B.ACC = \frac{1}{2} \cdot (\frac{TP}{TP + FN} + \frac{TN}{TN + FP}) = \frac{(TPR + TNR)}{2} \qquad (2)$$

Our approach resembles the EasyEnsemble approach [18], which we adapt for group as well as class imbalance. Specifically, we combine *bagging* and *boosting*; thus, the final model is an ensemble of ensembles. *Bagging* reduces model variance by generating multiple models from bootstrap samples drawn from the training data. *Boosting* reduces both (model) bias and variance by combining many weak learners, each focusing on missclassified examples from previous learners [21].

IV. FAE - A FAIRNESS-AWARE ENSEMBLE FRAMEWORK

Figure 1 shows an overview of FAE, from training (left side) to prediction of new instances (right side). FAE combines pre- and post-processing fairness-related interventions, as follows:

- **Fairness-aware ensemble learning**
  In *pre-processing*, we tackle the problems of group- and class-imbalance. In particular, we employ *bagging* to *balance* the groups in each bag by taking into account the protected positive group, and a representative sample from the other groups (Section IV-A). Afterwards, *boosting* [21] is employed on each bag, so at the end, an ensemble of ensembles is learned.
- **Fairness-aware decision boundary shift**
  In the *post-processing*, we shift the decision boundary of the learner in the direction of fairness based on a tunable parameter θ, until the $EQOP$ score satisfies the user-defined threshold ϵ (Section IV-B).
- **Selecting the shortest hypothesis** Finally, we select the optimal number of boosting models $u \in [k, 2k]$ that exhibits the best performance in terms of both fairness and balanced error (Section IV-C).

A. Fairness-aware ensemble training

In the pre-processing step, we tackle discrimination in the training data caused by group and class imbalance ensuring that the protected positive group will also be learned by the model. For that, we propose a fair and representative sample generation process. Each sample is created s.t it contains the whole *protected positive group* s^+ and a *representative equisized* sample from each of the other groups (i.e., from $s^-, \bar{s}^+, \bar{s}^-$).

Algorithm 1 shows the different steps in the ensemble's training phase. Clustering is applied in the beginning for each group $s^-, \bar{s}^+, \bar{s}^-$ (line 2). We employ *stratified sampling* to ensure a balanced representation, where the *strata* correspond to clusters[1] extracted through some clustering algorithm from the other groups $s^-, \bar{s}^+, \bar{s}^-$. The bags are created (lines 6-7) by combining s^+ and a stratified sample from the generated clusters for each group. In each bag, an AdaBoost classifier is trained (line 8) and added to the ensemble (line 9). The output model is an ensemble of ensembles E (line 12):

$$E(x) = \sum_{i=1}^{2k} \Big(\sum_{j=1}^{z} \Big(a_{i,j} h_{i,j}(x) \Big) \Big) \qquad (3)$$

where k is the number of bags (c.f., Eq. 4), z the number of boosting rounds and $a_{i,j}$ is the weight of the weak learner $h_{i,j}$ (a and h are obtained through AdaBoost).

[1]Clustering better approximates the underlying data distributions, accounting for sub-groups, and thus ensuring representative samples from each group.

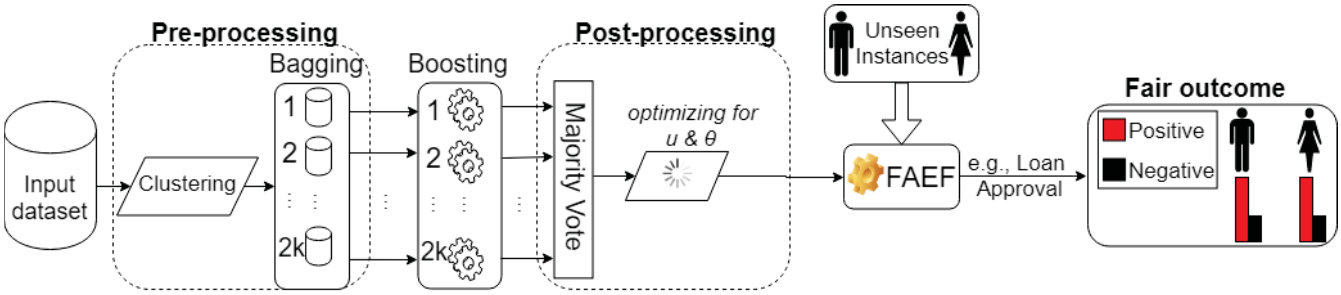

Fig. 1: An overview of our holistic pre- and post-processing FAE framework

*1) **Stratified sampling**:* The goal is to generate the different bags $s^{-'}, \bar{s}^{+'}, \bar{s}^{-'}$ from the majority groups $s^-, \bar{s}^+, \bar{s}^-$, respectively, such that: $|s^+| = |s^{-'}| = |\bar{s}^{+'}| = |\bar{s}^{-'}|$. To ensure representative samples from each group, we cluster each group (i.e., each of $s^-, \bar{s}^+, \bar{s}^-$) and use the resulting clusters for bag generation. Note that clusters are generated only once in the beginning of the training process (line 2, Algorithm 1) and re-used afterwards.

*2) **Estimating the initial number of bags**:* The number of bags k must be sufficient to overcome the drawback of potential loss of useful information due to under-sampling (i.e., each bag is a sample of the training data). We overcome this drawback by estimating the number of bags k s.t. we insure that the clustered instances are at least in one of the bags. We calculate the number of bags k as following:

$$k = \left\lceil \frac{max\{|s^-|, |\bar{s}^+|, |\bar{s}^-|\}}{|s^+|} \right\rceil + 1 \tag{4}$$

In other words, k provides an estimation that an instance from the most populated group will be at least in one bag, thus, avoiding the under-sampling drawback. In practice, we train the ensemble with twice the amount of bags ($2k$ bags); at the post-processing step, we select the best set of learners for the ensemble (Section IV-C).

B. Fairness-aware decision boundary tuning

Despite the pre-processing interventions, the resulting model E might not fulfill the discrimination threshold ϵ. In FAE, if EQOP $> \epsilon$, a post-processing procedure is invoked that shifts the decision boundary based on a parameter θ s.t. EQOP $\leq \epsilon$.

As we show in Section VI, by employing only the pre-processing step, the discrimination is significantly reduced. However, a post-processing step is necessary given that discrimination can stem from other factors including class overlap and the accuracy-oriented objective function of Adaboost.

Parameter tuning. For a SA (e.g. SA ='*gender*') our goal is to find the optimal threshold parameter θ_s or $\theta_{\bar{s}}$ (for the different attribute values $dom(SA) = \{s, \bar{s}\}$) to minimize $EQOP$. Furthermore, at any given time our ensemble learner E can discriminate against only one of the group s or $\bar{s}$.

Algorithm 2 shows the detailed steps for tuning the optimal θ_s and $\theta_{\bar{s}}$. To begin with, we compute the $EQOP$ score, which

Algorithm 1 Pre-processing step

Input: Training set D, target class y^+, SA, k
Output: Ensemble E

1: Extract groups $s^+, s^-, \bar{s}^+, \bar{s}^-$ based on y^+ and SA from D;
2: Generate clusterings $C_{s^-}, C_{\bar{s}^+}, C_{\bar{s}^-}$ from $s^-, \bar{s}^+, \bar{s}^-$, respectively;
3: Ensemble $E \leftarrow \{\emptyset\}$;
4: $i \leftarrow 1$;
5: **for** $i = 1 : 2k$ **do**
6: Stratified sample $s_i^{-'}, \bar{s}_i^{+'}, \bar{s}_i^{-'}$ from $C_{s^-}, C_{\bar{s}^+}, C_{\bar{s}^-}$;
7: Bag $B_i = s^+ \cup s_i^{-'} \cup \bar{s}_i^{+'} \cup \bar{s}_i^{-'}$;
8: Train an AdaBoost classifier H_i upon B_i;
9: $E \leftarrow E \cup H_i$;
10: $i \leftarrow i + 1$;
11: **end for**
12: **return** ensemble E;

represents the difference between true positive ratios between s and $\bar{s}$ (line 6). Next, we sort the misclassified instances from s^+ and $\bar{s}^+$ groups (lines 7 – 8) in a descending order (w.r.t the target class) based on their ensemble classification score from Equation 3. In case $EQOP$ score is below the discrimination threshold ϵ, then $\theta_{\bar{s}} = \theta_s = 0.5$ (lines 9 – 10). Setting the threshold parameter to 0.5 has no implication in classifying test instances in Equation 7. For $|EQOP| > \epsilon$, we distinguish between *discrimination* and *reverse discrimination* (lines 11 – 17). That is, for $EQOP > 0$ the model discriminates against instances with $SA = s$, otherwise against instances with $SA = \bar{s}$. The threshold parameter θ_s or $\theta_{\bar{s}}$ represents the $E(d)$ score of the last instance from the top_k necessary instances from MC_{s^+} or $MC_{\bar{s}^+}$ (lines 12 and 15) that need to be classified correctly to fulfill the criteria $|EQOP| \leq \epsilon$. The top_k instances needed for minimizing the discrimination are obtained as following:

$$\frac{top_k + TP_s}{TP_s + FN_s} = \frac{TP_{\bar{s}}}{TP_{\bar{s}} + FN_{\bar{s}}} \Rightarrow top_k = \left\lceil \frac{TP_{\bar{s}}(TP_s + FN_s)}{TP_{\bar{s}} + FN_{\bar{s}}} - TP_s \right\rceil \tag{5}$$

where TP and FN stand for true positive and false negative instances of protected and non-protected group respectively.

Algorithm 2 Post-processing step

Input: D, E, $s, \bar{s}$, ϵ
Output: $\theta_s, \theta_{\bar{s}}$

$\theta_s = \theta_{\bar{s}} = 0.5$
MC_{s^+}, $MC_{\bar{s}^+} \leftarrow \{\emptyset\}$
True positive rate TPR_s and $TPR_{\bar{s}}$ for s and $\bar{s}$
CC_{s^+} = #correctly classified instances in s^+
$CC_{\bar{s}^+}$ = #correctly classified instances in $\bar{s}^+$
$EQOP = TPR_{\bar{s}} - TPR_s$
Misclassified instances MC_{s^+} and $MC_{\bar{s}^+}$ for s^+ and $\bar{s}^+$
Sort MC_{s^+}, $MC_{\bar{s}^+}$ in descending order based on $E(d)$
IF $|EQOP| \leq \epsilon$ // no discrimination
    $\theta_s = \theta_{\bar{s}} = 0.5$
ELSE IF $EQOP > 0$ // discrimination
    $top_k = \frac{CC_{\bar{s}^+}}{|\bar{s}^+|}|s^+| - CC_{s^+}$
    $\theta_{\bar{s}} = MC_{s^+}[top_k]$
ELSE IF $EQOP < 0$ // reverse discrimination
    $top_k = \frac{CC_{s^+}}{|s^+|}|\bar{s}^+| - CC_{\bar{s}^+}$
    $\theta_s = MC_{\bar{s}^+}[top_k]$
ENDIF
return $\theta_s, \theta_{\bar{s}}$

C. Hypothesis selection

Out of the $2k$ learners, we select the shortest hypothesis (in terms of number of bags) that optimizes the following objective function:

$$\underset{u}{\operatorname{argmin}} \ (B.ERR_u + 2 \cdot |EQOP_u|) \tag{6}$$

where $B.ERR$ is the balanced error rate and $u \in [k, 2k]$ is a set of AdaBoost models (each AdaBoost is trained upon a different bag). The objective function is applied after the decision boundary adjustment i.e., Algorithm 2 is taking place after the pre-processing step, and afterwards the set of learners that minimize Equation 6 is selected. Since class imbalance is tackled in the pre-processing step, more emphasis is given to the ensemble's fairness in the objective function. The final model (FAE) is:

$$E(x) = \sum_{i=1}^{u} \Big(\sum_{j=1}^{z} \Big(a_{i,j} h_{i,j}(x) \Big) \Big)$$

.

D. FAE Classification

In classifying instances with *FAE*, we distinguish two cases. If $|EQOP| \leq \epsilon$, the classification is done solely through the majority voting scheme in $E(d)$ (c.f., Equation 3). This is the case, where no post-processing tuning is required, rather pre-processing interventions are adequate in fulfilling the $EQOP$ threshold. For $EQOP < 0$ and $|EQOP| > \epsilon$, our model discriminates against $SA = \bar{s}$ in the training set, hence, instances will be classified based on Equation 7.

$$f(d) = \begin{cases} y^+ & \text{if d(SA) = } \bar{s} \text{ and } E^+(d) \geq \theta_{\bar{s}} \\ \text{E(d)} & \text{otherwise.} \end{cases} \tag{7}$$

where E^+ is the probability of d assigned to y^+. Similar is the case for $EQOP > 0$ and $|EQOP| > \epsilon$; in this case, Equation 7 is altered by replacing $d(SA) = \bar{s}$ to $d(SA) = s$ and $\theta_{\bar{s}}$ to θ_s.

V. Experimental Setup

Our framework[2] has been instantiated with Logistic Regression as base learners. Each dataset is randomly split into train $(2/3)$ and test set $(1/3)$ (holdout evaluation, similar to [5]). We report on the average of 10 random splits. We set $\epsilon = 0$ as a threshold for EQOP (no discrimination). For AdaBoost, the maximum number of boosting rounds z is set to 25. We evaluate the following aspects: (i) classification performance based on balanced accuracy (B.ACC, Equation 2) and (ii) discriminative performance based on *EQOP* (Equation 1) .

A. Datasets

We evaluate our approach with two well known datasets: Adult census income and Bank. **Adult census income** dataset [22] contains demographic data from the U.S. The task is to determine if a person receives more than 50K dollars annually. We use as the target class, people who receive more than 50K per year. We remove duplicate instances and instances containing missing values which results to 45,175 instances. We consider as protected attribute $SA = Gender$ with $s = female$. **Bank** dataset [22] is related to direct marketing campaigns of a Portuguese banking institution and contains 40,004 instances. The task is to determine if a person subscribes to the product (bank term deposit). As target class we consider people who subscribed to a term deposit. We consider as $SA = maritial\ status$ with $s = married$.

B. Baselines and FAE Ablations

*1) **Baselines**:*

Shifted Decision Boundary (SDB) [13]: SDB uses a set of base classifiers in an AdaBoost classifier. Instead of majority voting (i.e., $\sum_{i=1}^{T} a_i h_i(x)$), SDB employs confidence scores (i.e., $\frac{\sum_{i=1}^{T} a_i h_i(x)}{\sum_{i=1}^{T} a_i}$) for predictions. The best threshold value for a specific protected group is established to minimize statistical parity. The shift in the boundary takes place after the training phase, thus, making it a post-processing method and suitable for comparison. To have a fair comparison, we find the best threshold estimation of SDB for EQOP, instead of statistical parity as in the original paper.

Disparate Mistreatment (DM): Zafar et al. [5] formulate the fairness problem as a set of constraints, for which they optimize a logistic regression (LR) model. They consider three sets of constrains: (i) minimize difference in FPR (false positive rate), (ii) minimize difference in FNR (false negative rate), and (iii) a combination of both. For our comparison, we employ only (ii) since $TPR = 1 - FNR$. We employ the method's default parameters.

AdaBoost: here we consider an ensemble learner (equipped with LR as a weak learner) without any pre- or post-processing

[2]https://iosifidisvasileios.github.io/Fairness-Aware-Ensemble-Framework/

	Adult Cen.		*Bank*	
Approach	**B.ACC. (%)**	**EQOP (%)**	**B.ACC. (%)**	**EQOP (%)**
AdaBoost	76.56	11.92	66.32	-6.25
EasyEnsemble	80.58	15.72	83.24	-4.52
DM	70.96	-11.83	65.69	-0.97
SDB	77.02	-2.72	66.23	-5.88
SMT	76.86	-2.99	73.26	30.58
OB (EM)	80.91	-4.31	83.10	2.21
OB (K-means)	80.92	-4.70	83.10	1.89
FAE (EM)	**81.09**	**1.52**	**83.29**	**-0.12**
FAE (K-means)	81.01	1.67	83.24	0.24

TABLE I: Evaluation results for $B.ACC.$ and $EQOP$. EQOP is in the range of [-1,1], in this case we show the percentage points. The best results are marked in boldface.

fairness-related interventions. The goal is to show the ability of these ensembles to classify under group and class imbalance and its impact on discrimination scores like EQOP.

EasyEnsemble: EasyEnsemble [18] is an ensemble that employs bagging and AdaBoost to tackle class imbalance, with LR as a weak learner. We employ EasyEnsemble to compare our approach with a method that directly tackles class imbalance. We set as number of bags to $N = 20$.

*2) **FAE Model Ablation**:*

FAE is a joint framework of pre-and post-processing interventions. We consider the following ablations, to evaluate the individual effect of the pre- and post-processing interventions:

Only Bagging (OB) is the pre-processing step in FAE (c.f. Section IV-A). We use OB to show the behavior of the ensemble that is trained upon fair and representative groups, without further tuning its decision boundary.

Simple Majority Threshold (SMT) refers to the post-processing part in FAE (c.f. Section IV-B). This method is similar to SDB [13], however, instead of using confidence scores, we use the default majority vote of an AdaBoost classifier. That is, after training, we compute the best parameter θ for a specific protected group to minimize EQOP (Algorithm 2). We use SMT to show how individual post-processing tuning affects the performance of the models.

We use EM and K-means clustering algorithms to compare the impact of clustering in the bagging step in FAE and its pre-processing step OB, which we indicate with FAE (EM) and OB (EM), and FAE (K-means) and OB (K-means), respectively. For EM, the optimal number of clusters for each group is estimated via cross validation (100 iterations) while for K-means we use the elbow metric (least squares), where the number of clusters ranges in $[2, 25]$.

VI. Evaluation Results and Discussion

We report on: (i) *classification performance* w.r.t *B.ACC* and (ii) *fairness performance* w.r.t. EQOP.

Table I shows the scores for the B.ACC metric for both datasets and approaches under comparison. Our approach FAE achieves some of the highest $B.ACC$ scores, with an average score of $\overline{B.ACC} = 82.19\%$ across all datasets for FAE (EM). Similar is the score of EasyEnsemble with $\overline{B.ACC} = 81.91\%$. Yet, in terms of EQOP EasyEnsemble produces highly discriminatory results, since it focuses solely on predictive performance.

A detailed inspection across the competing approaches reveals that the differences between non-bagging and non-ensemble approaches are highly significant. An even representation of all groups is important for classification performance. For models like AdaBoost, SMT, SDB, DM that do not account for the group imbalance, we see a huge drop in B.ACC scores. FAE (EM) has a 20% relative increase when compared against DM, and 15% relative increase against the other models.

Ensemble Learners: The case of AdaBoost shows that using solely ensemble learners is not sufficient to ensure a non-discriminatory classification. It has the second lowest performance with $\overline{B.ACC} = 71.44\%$. EasyEnsemble which focus on class imbalance has very good predictive performance with $\overline{B.ACC} = 81.91\%$; however, this is not sufficient to tackle discrimination. Same behavior can be observed for OB. This confirms our assumption, that such discriminatory behaviors are a result of other factors such as class overlap.

Bagging: Bagging ensures even representations of the different groups, thus, it enables models that achieve better B.ACC. Models that employ bagging achieve similar B.ACC scores. Comparing against other non-bagging approaches, such as AdaBoost, DM and SMT, we note a significant drop in terms of B.ACC. However, it is important to note that a high B.ACC score is not sufficient for non-discriminatory classification behavior because, discrimination is often manifested in terms of uneven probabilities for granting a benefit to different groups (c.f. Section III).

Regarding discrimination, we observe that high B.ACC scores do not necessarily correlate with low EQOP scores, that is, discrimination free classification behavior. In our choice of competitors, it is evident that such strategies are often insufficient in minimize discrimination.

From the competitors, only AdaBoost and EasyEnsemble have low EQOP scores. EasyEnsemble is particularly interesting; its B.ACC score is on average close to FAE (EM), however, it exhibits a high discrimination score with $\overline{EQOP} = 10.12\%$. This highlights that optimizing only for classification performance is subject to pitfalls of uneven distributions of groups. Whereas our models, the pre-processing stage OB, and FAE, achieve the lowest discrimination results. FAE (EM) has the lowest score with $\overline{EQOP} = 0.82\%$ with nearly an ideal EQOP score.

Contrary, for models that optimize for discrimination free classification, we note a significant decrease of EQOP scores compared to AdaBoost and EasyEnsemble. For example, DM in its optimization function minimizes for the EQOP score, leading to $\overline{EQOP} = 8.18\%$. Yet, its B.ACC score is severely impacted. This is mostly due to the fact that it learns a logistic regression model under high group imbalance.

An important comparison is between FAE and DM. FAE provides a high relative decrease of 90% in terms of EQOP. This shows, that despite the fact that DM optimizes the training objective to reduce discrimination, the impact of fair and balanced representations of all groups in training supervised models is highly important.

VII. Conclusions and Future Work

In this paper, we addressed the problem of discrimination against marginal groups in classification models caused by group imbalance, class imbalance and societal encoded biases manifested as class overlap esp. for the protected group. We presented the FAE framework, a holistic approach to fairness-aware classification that combines pre-processing balancing strategies with post-processing decision boundary adjustment. The pre-processing stage, which computes the number of bags and determines the different groups and clusters to ensure fair representation allows the models to learn representative classifiers that significantly increase the performance and at the same time reduce the discrimination. Due to the encoded societal biases (*class overlap*) in the data, even representations among groups are insufficient in addressing discrimination. Hence, we shift the decision boundary and additionally select hypotheses from the ensemble learners for nearly ideal EQOP scores. Such steps ensure that a reduction in terms of EQOP does not come at the cost of the ability of the model to correctly classify instances into their corresponding classes.

Our experiments show that discrimination free models are feasible, and for a given *feature space*, we can achieve maximal classification performance, and account for important factors like discrimination for a given target measure, e.g., EQOP. In our current version of FAE, we employ pre- and post-processing fairness-enhancing interventions. Furthermore, improvements are possible by including in-processing interventions at the algorithm level, thus targeting the whole ML process from data to algorithms and models.

Acknowledgment

This work is part of a project that has received funding from the European Unions Horizon 2020, under the Innovative Training Networks (ITN-ETN) programme Marie Skodowska-Curie grant (NoBIAS-Artificial Intelligence without Bias) agreement no. 860630. The work is also inspired by the Volkswagen Foundation project BIAS ("Bias and Discrimination in Big Data and Algorithmic Processing. Philosophical Assessments, Legal Dimensions, and Technical Solutions") within the initiative "AI and the Society of the Future"; the last author is a Project Investigator for both of them.

References

[1] U. S. E. O. of the President and J. Podesta, *Big data: Seizing opportunities, preserving values*. White House, Executive Office of the President, 2014.

[2] D. Ingold and S. Soper, "Amazon doesnt consider the race of its customers. should it," *Bloomberg, April*, 2016.

[3] T. Calders and I. Žliobaitė, "Why unbiased computational processes can lead to discriminative decision procedures," in *Discrimination and privacy in the information society*. Springer, 2013, pp. 43–57.

[4] K. Lum and W. Isaac, "To predict and serve?" *Significance*, vol. 13, no. 5, pp. 14–19, 2016.

[5] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi, "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment," in *Proceedings of the 26th International Conference on World Wide Web*, 2017, pp. 1171–1180.

[6] F. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention," in *Advances in Neural Information Processing Systems*, 2017, pp. 3992–4001.

[7] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and Information Systems*, vol. 33, no. 1, pp. 1–33, 2012.

[8] V. Iosifidis and E. Ntoutsi, "Dealing with bias via data augmentation in supervised learning scenarios," *Jo Bates Paul D. Clough Robert Jäschke*, p. 24, 2018.

[9] E. Krasanakis, E. Spyromitros-Xioufis, S. Papadopoulos, and Y. Kompatsiaris, "Adaptive sensitive reweighting to mitigate bias in fairness-aware classification," in *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, 2018, pp. 853–862.

[10] F. Kamiran, T. Calders, and M. Pechenizkiy, "Discrimination aware decision tree learning," in *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. IEEE, 2010, pp. 869–874.

[11] C. Dwork, N. Immorlica, A. T. Kalai, and M. D. Leiserson, "Decoupled classifiers for group-fair and efficient machine learning," in *Conference on Fairness, Accountability and Transparency*, 2018, pp. 119–133.

[12] V. Iosifidis and E. Ntoutsi, "AdaFair: Cumulative fairness adaptive boosting," *in Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2019.

[13] B. Fish, J. Kun, and Á. D. Lelkes, "A confidence-based approach for balancing fairness and accuracy," in *Proceedings of the 2016 SIAM International Conference on Data Mining*. SIAM, 2016, pp. 144–152.

[14] D. Pedreschi, S. Ruggieri, and F. Turini, "Measuring discrimination in socially-sensitive decision records," in *Proceedings of the 2009 SIAM International Conference on Data Mining*. SIAM, 2009, pp. 581–592.

[15] A. Estabrooks, T. Jo, and N. Japkowicz, "A multiple resampling method for learning from imbalanced data sets," *Computational intelligence*, vol. 20, no. 1, pp. 18–36, 2004.

[16] C. Drummond, R. C. Holte *et al.*, "C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling," in *Workshop on learning from imbalanced datasets II*. Citeseer, 2003, pp. 1–8.

[17] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*. IEEE, 2008, pp. 1322–1328.

[18] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 2, pp. 539–550, 2009.

[19] M. Hardt, E. Price, N. Srebro *et al.*, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems*, 2016, pp. 3315–3323.

[20] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *2010 20th International Conference on Pattern Recognition*. IEEE, 2010, pp. 3121–3124.

[21] R. E. Schapire, "A brief introduction to boosting," in *IJCAI*, vol. 99, 1999, pp. 1401–1406.

[22] K. Bache and M. Lichman, "UCI machine learning repository," 2013.