

Towards a holistic view of bias in machine learning: bridging algorithmic fairness and imbalanced learning

Damien Dablain¹ · Bartosz Krawczyk² · Nitesh Chawla¹

Received: 1 June 2023 / Accepted: 29 February 2024

Published online: 04 April 2024

© The Author(s) 2024 [OPEN](#)

Abstract

Machine learning (ML) is playing an increasingly important role in rendering decisions that affect a broad range of groups in society. This posits the requirement of *algorithmic fairness*, which holds that automated decisions should be equitable with respect to protected features (e.g., gender, race). Training datasets can contain *both* class imbalance and protected feature bias. We postulate that, to be effective, both class and protected feature bias should be reduced—which allows for an increase in model accuracy and fairness. Our method, Fair OverSampling (FOS), uses SMOTE (Chawla in J Artif Intell Res 16:321–357, 2002) to reduce class imbalance and feature *blurring* to enhance group fairness. Because we view bias in imbalanced learning and algorithmic fairness differently, we do not attempt to balance classes *and* features; instead, we seek to *de-bias* features and *balance* the number of class instances. FOS restores numerical class balance through the creation of synthetic minority class instances and causes a classifier to pay *less attention* to protected features. Therefore, it reduces *bias* for both *classes* and *protected features*. Additionally, we take a step toward bridging the gap between fairness and imbalanced learning with a new metric, *Fair Utility*, that measures model effectiveness with respect to accuracy and fairness. Our source code and data are publicly available at <https://github.com/dd1github/Fair-Over-Sampling>.

Keywords Machine learning · Imbalanced data · Fairness · Bias · Oversampling

1 Background

Automated decision-making models are progressively being used in situations that affect humans in a broad range of areas, such as credit risk analysis [2], criminal recidivism prediction [3, 4], hiring [5] and the provision of social services [6]. For example, a bank may decide to extend credit based on whether a machine learning (ML) model predicts that an individual may default on a loan. Conversely, a judge may determine that a defendant should not be released while awaiting trial if an artificial intelligence (AI) model suggests that the defendant has a high risk of recommitting a crime. The growing prevalence of ML algorithms in decisions that affect humans is due, in part, to their perceived accuracy and ability to detect hidden patterns in data. Yet, in some cases, these models have been demonstrated to incorporate biases, such as in hiring decisions [7], face recognition [8] and even translation [9], resulting in concerns about the fairness of machine learning algorithms [10]. Despite these concerns, it is likely that the use of automated decision-making will only increase in the future, as AI becomes more wide-spread in society, government and business. Therefore, there is a growing awareness that ML algorithms should be *both* accurate and fair, which is underscored by the recently released Artificial

✉ Nitesh Chawla, nchawla@nd.edu; Damien Dablain, ddablain@nd.edu; Bartosz Krawczyk, bkrawczyk@vcu.edu | ¹Lucy Family Institute for Data & Society, Department of Computer Science and Engineering, University of Notre Dame, Notre Dame 46556, IN, USA. ²Department of Computer Science, Virginia Commonwealth University, Richmond 23824, VA, USA.



Intelligence Act—a legal framework promulgated by the European Commission that mandates non-discrimination, among other requirements, for ML models that affect individuals [11].

Although it is desirable for ML models to be both fair and accurate, there is often a *trade-off* between these two goals [3, 12, 13], such that increasing fairness comes at the cost of reduced accuracy. Accuracy and fairness are often at odds because they are influenced by *imbalanced data*. In many cases, the data used to train ML algorithms is imbalanced with respect to class and protected features, such that one class or group is under-represented with respect to another. Since most ML models learn parameters based on data, data imbalance can cause a particular class or sub-group to be over-weighted, such that preference is given to the over-represented class or group. Hashimoto et al. [14] has referred to the under-representation of a protected group in training data as *representation disparity*, such that minority groups contribute less to a ML model objective because they are under-represented in the training data, and hence model accuracy may be lower for the minority class.

Summary. The algorithmic fairness domain focuses on combating bias in decision-making originating in protected features that could affect the objectiveness of the decision. At the same time, the class imbalance domain focuses on countering bias originating from skewed class distributions, as majority classes may be preferred over minority ones during classifier training. We take a step toward bridging the gap between algorithmic discrimination and imbalanced learning by discussing the key concepts and metrics that underpin both areas. Because it is often not possible for a ML algorithm to meet multiple fairness criteria (e.g., individual and group non-discrimination) at the same time [15, 16], we focus on a single element—group fairness. We show that a common approach used in imbalanced learning—data oversampling—can be used to increase model fairness and accuracy.

Main contributions. This paper offers the following insights and contributions towards fair machine learning:

- *Bridging the Gap Between Fairness and Imbalanced Learning:* We take a step toward bridging the gap between the algorithmic fairness and imbalanced learning fields by discussing commonalities and differences in the approaches that both fields use to overcome bias in machine learning.
- *Fair Oversampling:* We propose a new data pre-processing technique, Fair Oversampling (FOS), that enhances fairness and classifier accuracy. Unlike existing fairness pre-processing methods, which seek to balance *both* class instances and protected groups, FOS numerically balances classes and de-biases protected features through feature blurring. Balancing the number of class instances improves prediction accuracy and blurring protected features improves group fairness metrics.
- *Fair Utility Metric:* We propose a new metric that combines fairness with imbalanced learning—Fair Utility—that relies on measures commonly used in both fields.

Organization The paper is organized as follows. It first discusses algorithmic fairness and its key concepts. Then, it reviews the central elements of imbalanced learning. Next, it introduces our algorithm and proposed fairness metric. Finally, the paper discusses experimental results, and the commonalities and differences between algorithmic fairness and imbalanced learning.

2 Related work

2.1 Algorithmic fairness

Discrimination can generally be defined as the prejudicial treatment of an individual based on membership in a legally protected group. Algorithmic fairness is concerned with ensuring that decisions made by machine learning models are equitable with respect to protected groups [17, 18]. Algorithmic fairness commonly invokes one of the following changes in order to produce equitable results: (1) modifications to the training data, (2) changes in the machine learning model, or (3) modifications to the decisions themselves [19].

We concentrate on supervised learning in the context of binary classification. In binary classification, the goal of algorithmic fairness is to fairly select between two actions, a_0 and a_1 (e.g., approve or decline the extension of credit in banking). In our discussion, we adopt notation used by Speicher et al. [20] to describe our algorithmic environment. Thus, a ML decision algorithm, A , can be described as a function $A : \mathbb{R} \rightarrow \{0, 1\}$ that outputs a binary decision. The machine learning algorithm, A , parameterized by θ , accepts as input training data, D , minimizes a loss function $l(\theta)$, and predicts a label (i.e., 0 or 1).

More formally, A accepts as input training data, $D = \{(x_i, y_i)\}_{i=1}^n$ with n examples, features $x_i \in X$, where $y_i \in Y$ represents the prediction or label ($Y = \{0, 1\}$) for each individual i . The features, X , can either be discrete or continuous. We partition the set of features (or attributes) into two groups, sensitive or protected features, such as gender, race or age, and unprotected features, such that $x = (x_p, x_u)$. We also assume that protected features can be further partitioned into two classes, privileged and unprivileged (i.e., $x_p = (x_{pr}, x_{up})$). For purposes of this paper, we assume that the label contained in the dataset is the correct, unbiased label.

Narayanan described at least 21 mathematical definitions of fairness that have been proposed by the fairness research community [21]. Two broad classes of algorithmic equity have gained prominence: group and individual fairness. Group fairness requires that the ML algorithm, A , produces parity for a given metric, M , for protected features, such that $M_{x_p}(A) = M_{x'_p} \forall x_p, x'_p \in X$. Individual fairness requires that similar individuals are treated similarly. It implies the presence of a similarity metric that is capable of determining if a pair of individuals are similar.

Because of the challenges in finding suitable individual fairness similarity metrics, we focus on group fairness in this paper. Corbett-Davies et al. describe three central concepts embodied by group fairness: anti-classification, classification parity and calibration [22]. Anti-classification requires that AI algorithms do not consider protected features when making decisions [9, 23, 24]. Thus, anti-classification provides that: $A(x) = A(x') \forall x$, such that $x_u = x'_u$. Classification parity (sometimes referred to as statistical parity) requires that certain measures are equal across sensitive features. Statistical parity can be expressed in a variety of ways. Under one formulation, the proportion of members in a protected group receiving a positive classification must be identical to the proportion in the population as a whole [25]. Other measures focus on the difference in positive or negative rates (instead of proportions) between sensitive groups (e.g., equal true positive rates for both male and female applicants). Classification parity has been widely used as a fairness metric in machine learning [26–28]. As discussed below, we use classification parity in our metrics. Demographic parity, or the proportion of positive decisions, means that $Pr(A(X) = 1 | x_p) = Pr(A(X) = 1)$ [29]. Whereas, parity of false positives requires that $Pr(A(X) = 1 | Y = 0, x_p) = Pr(A(X) = 1 | Y = 0)$. We also incorporate demographic parity into our metrics, although we focus on differences in true positive, false positive, and true negative rates, instead of their relative proportions.

In order to achieve group fairness in machine learning, a variety of techniques have been employed, which can be broadly separated into pre-processing, in-processing and post-processing methods. Pre-processing techniques involve manipulating the training data before it is consumed by a classification algorithm, in-processing incorporates fairness into a ML algorithm loss function, and post-processing aims to adjust the decisions of a classifier to be fair. We briefly survey below the key pre-processing techniques that are relevant to our approach.

Kamiran and Calders propose a pre-processing method, Reweighting, that creates weights for the training instances to ensure fairness [30]. They effectively divide the training set into four groups: (1) privileged group, majority class; (2) unprivileged group, majority class; (3) privileged group, minority class; and (4) unprivileged group, minority class. They then develop separate weights for each of the four groups and apply the weights to each instance. Similar to Kamiran and Calders, Li and Liu propose to reweight data samples to improve fairness by granularly modeling the influence of each training sample [31]. Feldman et al. propose a pre-processing method, Disparate Impact Remover, that modifies features to enhance group fairness while preserving rank-ordering within protected groups [29].

2.2 Imbalanced learning

Imbalanced learning is concerned with disproportions among classes. In binary classification, the number of instances of one class (the majority) outnumber the other (minority). The skewed distribution of examples in favor of the majority class can cause classifiers to be biased toward the majority because the algorithm's parameters are more heavily weighted toward more frequently occurring examples. Classifiers can achieve high accuracy by merely selecting the majority class. However, the minority class is often the more important one from the data inference perspective because it may carry more relevant information.

There are three broad approaches within imbalanced learning: data-level methods that modify the training data to balance class distributions, algorithm level methods that ameliorate bias in classifiers towards the majority class, and ensemble methods that are a combination of the first two with classifier committees.

Data-level approaches. This group of methods focus on modifying the training set by balancing the number of minority and majority class examples. Oversampling generates new minority class examples, while under-sampling removes instances from the majority class. Under-sampling can result in removal of important data from the training set and therefore is often not preferred. Simple random oversampling (ROS) merely duplicates instances of the minority class to

impose parity. SMOTE, or the Synthetic Minority Oversampling Technique, [1], is a popular oversampling method used in the imbalanced learning community. It randomly selects a nearest neighbor of a minority instance and linearly generates synthetic examples based on the original instance and a nearest neighbor. SMOTE has been adapted to enhance the importance of class borderline instances [32], define safe regions that do not sample from noisy or overlapping instances [33] and has been applied in the deep learning [34] and big data [35] contexts. Alternative approaches to SMOTE have been proposed recently that do not rely on k -nearest neighbors, instead using alternative measures such as class potential [36], Mahalanobis distance [37], or manifold approximation [38].

Algorithm-level approaches. This group of methods modify the training procedure of a classifier to make it skew-insensitive, or incorporate alternative cost functions. Cost sensitive learning, which is a form of importance sampling [39], magnifies the importance of minority examples by increasing the penalty associated with the instances. Recent examples of cost-sensitive methods that have been used in imbalanced learning include the focal loss [40], the class-balanced margin loss [41], the distribution aware margin loss [42] and the asymmetric loss [43].

Ensemble approaches. Combining multiple classifiers is considered as one of the most effective approaches in modern machine learning [44]. Ensembles find their natural application in learning from imbalanced data, as they leverage the predictive power of multiple learners. By combining base classifiers with data or algorithm-level solutions, they achieve locally specialized robustness and maintain diversity among ensemble members. Most popular solutions combine resampling with Bagging [45] or Boosting [46], use mutually complimentary cost-sensitive learners [47], or rely on dynamic selection mechanisms to tackle locally difficult decision regions [48].

2.3 Imbalance within algorithmic fairness

Several papers have discussed the relationship between fairness and imbalance in machine learning. Shui et al. examine fairness with respect to group sufficiency where sub-groups have a limited number of instances [49]. They apply their method to natural language processing. Subramanian et al. observe that protected features may be associated with class labels, which may result in stereo-typing in natural language processing [50]. They propose an approach that algorithmically reweights class instances and protected features so that the cost associated with features and classes are balanced. Yan et al. observe that traditional imbalanced learning methods, such as SMOTE, can actually increase group discrimination [51]. They use a variant, K-Means SMOTE [52], and clustering to remove class instances from the original dataset that are near decision boundaries, which improves fairness (under-sampling). Their method is called Fair Class Balancing (FCB). Ferrari and Bacciu state that class and protected feature bias are related because they are caused by data complexity, as well as class and feature imbalance [53]. They propose to modify standard cross-entropy loss with an adaptive hyper-parameter that takes into account feature and class imbalance. Iosifidis and Ntoutsi state that one of the main reasons for bias in ML models is under-represented features [54]. They use SMOTE and feature generation to balance the number of protected features in training data. Chakraborty et al. attribute bias to data imbalance and improper class labeling [55]. They use SMOTE-based interpolation to equalize the number of class and protected feature instances; and then proactively remove data from the training set that is deemed biased (under-sampling). Their method works directly on categorical data. Wang et al. develop a method, Fair Streaming, to balance sub-groups in streaming data [56]. They also design multiple pseudo models in order to develop a baseline related to the trade-off between fairness and accuracy, which we address with a simple metric, Fair Utility (discussed below). Tarzanagh et al. balance sub-groups with a tri-level optimization framework that uses local predictors [57].

Although some of the above works incorporate SMOTE to address fairness, they do so to balance the number of class and protected features; and sometimes follow this balancing step with under-sampling. The above papers generally attribute bias in algorithmic fairness to spurious associations between labels and features, or to numerical disparity in classes and features. Because several of the works attribute bias in algorithmic fairness to under-representation of protected features; they seek to numerically balance them through data augmentation, under-sampling, or equalizing costs through loss functions.

In contrast, we introduce feature blurring and add an additional pre-processing step that converts categorical data into integers, which facilitates feature interpolation. We also do not attempt to numerically balance protected features and do not use under-sampling. Instead, we de-bias features by causing a classifier *not* to rely on a protected feature for prediction—thus enhancing fairness. Separately, we balance class instances, which improves prediction accuracy.

3 Why we need to bridge algorithmic fairness and imbalanced learning?

Different views on bias. The previous sections provided general background and reviewed recent advancements in algorithmic fairness and imbalanced learning. This allows us to see the strong parallel between them, as they both deal with the problem of countering class and feature bias, however from different perspectives.

- *Bias according to algorithmic fairness.* Here, bias is seen as a lack of fairness and transparency, originating from social background and the nature of the data itself. Fairness focuses on bias based on using sensitive or protected information (e.g., race or gender) to make a decision. Fairness-aware algorithms also focus on using safe information for training classifiers and debiasing them with respect to protected features.
- *Bias according to imbalanced learning.* Class imbalance focuses on bias originating in disproportion among classes, as most machine learning algorithms will become biased towards classes with a higher number of training instances. This puts smaller, yet often more important, classes at a disadvantage. Imbalance-aware algorithms focus on either balancing class distributions or removing the bias towards majority classes from the training process.

Interaction of class and feature bias. Algorithmic fairness views the source of bias with respect to protected *features* (e.g., race, gender) of a class instance (e.g., a student denied admission to a university), while imbalanced learning views bias as arising from a numerical disproportion among class *instances* themselves. In both algorithmic fairness and imbalanced learning, bias can emerge from a machine learning model performing supervised classification (e.g., a support vector machine, logistic regression classifier). In this paper, we argue that model fairness can be enhanced by addressing both class and feature bias; however, each concern requires a different remedy. Class bias can be addressed through data augmentation by numerically balancing instances; whereas feature bias can be addressed by causing a model to discount protected features—thus forcing it to pay attention to other relevant (non-protected) features when rendering a decision.

4 Fair oversampling

Our algorithm, Fair Oversampling (FOS), is designed to improve fairness and increase classifier accuracy. When training a machine learning model to accurately predict classes, it is often necessary to equalize the *number* of training examples between classes to ensure that models based on parametric learning are able to balance weights between specific classes. If a classifier observes very few instances of a minority class, its parameters may be biased toward recognizing the dominant class. At the same time, FOS addresses fairness by debiasing protected features. It does this by effectively mixing samples between protected group members, which causes the classifier to become confused about a particular feature, thus forcing it to rely on other features for accurate prediction.

FOS modifies a training dataset D so that it can be input to the machine learning model. FOS acts on two types of independent variables (X) in the training data, protected features x_p , and unprotected features x_u . FOS incorporates SMOTE, which uses feature interpolation to create synthetic instances. SMOTE relies on features being expressed as real numbers; therefore, as a pre-processing step, we convert categorical features to integers, if they are present in D . The pseudocode for FOS is displayed in Algorithm 1.

Algorithm 1 Fair Over-Sampling

Pseudocode:
NNB = nearest neighbor with KNN algorithm
Convert categorical data to integers
 D_1 = protected feature category with fewest instances
Use SMOTE to over-sample D_1 using only NNB of D_1
 D_2 = protected feature category with most instances
Use SMOTE to over-sample D_2 using NNB of entire minority class

FOS first determines the minority and majority classes ($Y = \{0, 1\}$ or $Y = \{min, maj\}$). It then subdivides the protected features x_p into two categories - privileged and unprivileged ($x_p = (x_{pr}, x_{up})$). This categorization results in four sub-groups: privileged majority (D_{prmaj}), unprivileged majority (D_{upmaj}), privileged minority (D_{prmin}) and unprivileged minority (D_{upmin}).

The objective of FOS is to restore balance between the classes through random oversampling and nearest neighbor metrics, using the mechanics of the SMOTE algorithm. FOS numerically balances the classes, such that the number of examples (N) in the majority class (N_{maj}) equals the number of examples in the minority class (N_{min}), or $N_{maj} = N_{min}$. It determines the protected group x_p in the dataset D that requires the least number of samples to obtain equivalency (denoted as D_1), and selects the K nearest neighbors of a random sample of D_1 (e.g., the unprivileged, minority group D_{upmin}). The number of random samples selected from this group equals the number of samples required to make it equal in number to the same group in the majority class. For D_1 , the samples are drawn from a single protected sub-group.

Next, the same oversampling procedure is repeated for the protected group x_p with the larger number of samples that are required to obtain numerical equivalency which is denoted as D_2 , except that instead of drawing the nearest neighbors exclusively from the D_2 pool, they are drawn from the entire minority class. This approach reduces bias because it blurs the difference between privileged minority D_{prmin} and unprivileged minority D_{upmin} group members, since it draws a nearest neighbor from the entire minority class D_{min} , which consists of both privileged and unprivileged members. FOS balances the *number* of class instances within a training dataset. It does not balance protected feature ratios.

5 Fair utility

5.1 Background

For purposes of our experiments (see Sect. 6), group fairness metrics were selected that could be expressed as elements of a binary classification confusion matrix consisting of True Positive Rate (TPR), False Positive Rate (FPR), True Negative Rate (TNR), and False Negative Rate (FNR). Metrics were chosen that are widely used in the fairness and imbalanced learning communities: Balanced Accuracy, Average Odds Difference (AOD), Absolute Average Odds Difference (AAO), Equal Opportunity Difference (EOD), and True Negative Rate Difference (TNRD) [55, 58, 59]. AOD is the average difference in the False Positive Rate plus the True Positive Rate for privileged and unprivileged groups [60]. It can be expressed as: $\frac{1}{2}((TPR_p - TPR_{up}) + (FPR_p - FPR_{up}))$, where TPR_p is the TPR of privileged instances, TPR_{up} is the TPR of the unprivileged instances, FPR_p is the FPR of privileged instances, and FPR_{up} is the FPR of the unprivileged instances. AAO is the same as AOD, except that TPR and FPR are absolute value calculations. EOD is the difference between the True Positive Rate of privileged and unprivileged groups [60], and can be expressed as: $(TPR_p - TPR_{up}) + (FPR_p - FPR_{up})$. TNRD is $(TNR_p - TNR_{up})$, where TNR_p is the TNR of privileged instances, and TNR_{up} is the TNR of the unprivileged instances.

5.2 Proposed metric

In addition to the metrics discussed above, we propose a new metric, called *Fair Utility*. In developing this metric, we are inspired by Corbett-Davies et al. [16]. They characterize algorithmic fairness in terms of constrained optimization in the context of the COMPAS algorithm for determining whether defendants in Broward County, FL, who were awaiting trial, were too dangerous to be released. In their formulation, the objective of algorithmic fairness is to *both* maximize public safety and reduce racial disparities. We also view algorithmic fairness as a multi-objective optimization problem, where the goal is to maximize the accuracy of a classifier and reduce group inequality. We approach the optimization problem with a data pre-processing technique designed to balance class accuracy prediction with protected group equity. We are also inspired by Halevy, Norvig and Pereira, who postulated that, in machine learning, a large quantity of data is more important than a strong algorithm [61]. Our metric named *Fair Utility* can be expressed as balanced accuracy multiplied by the average of TPRD plus TNRD. More explicitly, it is: $BA \times \frac{1}{2} \times ((1 - |TPRD|) + (1 - |FPRD|))$, where BA is balanced accuracy, TPRD is $(TPR_p - TPR_{up})$, and FPRD is $(FPR_p - FPR_{up})$. Utility involves maximizing the benefit of taking an action, compared with its costs. Here, we treat accuracy as equivalent with utility, which assumes that the class label assigned by the dataset is correct and does not contain inherent bias. The objective of *Fair Utility* is to combine accuracy and fairness into a single metric by incorporating balanced accuracy (which reflects the impact of class imbalance) with two fairness metrics (true positive and true negative rates) that track whether a classifier consistently accepts or rejects protected group members.

Table 1 Description of the Datasets

Dataset	Instances	Features	Protected feature	Classes
German Credit	1000	20	Gender	Good credit; bad credit
Adult Census Income	48,842	14	Gender	Income > or < 50 K
Compas	7214	28	Gender	Recidivism; No recidivism

Table 2 Class and protected feature imbalance ratios for each dataset

Class/Feature	German credit	Adult census	Compas
Classes			
Majority	700	37,155	3963
Minority	300	11,687	3251
Ratio	2.33	3.18	1.22
Protected features			
Privileged	690	32,650	5819
Unprivileged	310	16,192	1395
Ratio	2.23	2.02	4.17
Combined class and protected features			
Privileged, majority	499	22,732	3066
Unprivileged, majority	201	14,423	897
Ratio	2.48	1.58	3.42
Privileged, minority	191	9,918	2753
Unprivileged, minority	109	1,769	498
Ratio	1.75	5.61	5.53

6 Experiments

The following experiments were designed to answer our research questions (RQ):

- RQ1: Does FOS improve both algorithmic fairness and robustness to class imbalance for popular standard classifiers?
- RQ2: What is the FOS trade-off between fairness and skew-insensitive metrics when handling varying imbalance ratio levels?
- RQ3: Does FOS reduce protected feature importance?

6.1 Datasets

Three popular datasets are selected for testing that are used by the fairness research community [62]: German Credit [63], Adult Census Income [64], and Compas Two-Year Recidivism [65]. The key statistics of each dataset are summarized in Table 1. All three datasets involve binary classification. The German Credit dataset contains data that allows a classifier to predict whether an individual should have a positive or negative credit rating. The Adult Census Income dataset predicts whether an individual earns more or less than \$50K. The Compas dataset can be used to predict whether a defendant will commit a crime within a two year period.

As we can see from Table 2, all of the datasets exhibit both class and protected attribute (gender) imbalance. Compas shows the least amount of class imbalance, with a ratio of 1.22:1, while the German Credit and Adult Census datasets have class imbalance ratios ranging from approximately 2:1 to 3:1. All three datasets show greater protected attribute imbalance than class imbalance, with the ratios ranging from approximately 2:1 to 4:1. In the minority class, the maximum protected attribute imbalance ratios are even higher, ranging from 1.75:1 in German Credit to 5.61:1 in Adult Census.

6.2 Experimental design

Experiment 1: Oversampling for standard classifiers. First, modified training data produced by FOS was used as input to two standard machine learning classifiers: SVM and Logistic Regression (LG). The performance of the models was assessed based on metrics, which are discussed below. The performance of our algorithm was compared against four benchmarks for each standard classifier: (1) a baseline (no modifications to the training dataset); (2) a popular imbalanced learning oversampling method—SMOTE [1]; and two pre-processing algorithms that are specifically designed to improve fairness—(3) Reweighting [19] and (4) Disparate Impact Remover [29]. The purpose of this experiment was to determine how FOS compared to other data pre-processing algorithms that are used in both the imbalanced learning and fairness research communities.

Experiment 2: Robustness to increasing imbalance ratios. Second, we assessed how the performance of a standard ML classifier was affected by increasing levels of class and protected group imbalance. For this test, SVM was used as the ML algorithm with varying degrees of imbalance. Instances were randomly removed from classes and protected groups to achieve the intended imbalance levels. The selected imbalance levels were: $l \in \{1, 1.2, 1.4, 1.6, 1.8, 2\}$ for German Credit; $l \in \{1, 2, 2.5, 3, 3.5, 4\}$ for Compas; and $l \in \{1, 2, 4, 6, 8, 10\}$ for Adult Census, where l represents the denominator of the fraction that reduced the number of original protected group members. The reason for different levels of imbalance by dataset is that classifiers trained with the Reweighting and Disparate Impact Remover algorithms produced unstable results for datasets with a relatively small number of examples (i.e., German Credit and Compas), such that the classifiers predicted all labels to reside in a single class, thus causing True Negatives to be zero, yielding “Nan” metrics. In contrast, both SMOTE and FOS were able to work at the $l = 10$ level for all datasets. Therefore, the imbalance ratio scaling was adjusted so that all pre-processing algorithms could be assessed for all datasets.

Experiment 3: Impact of oversampling on feature importance. Third, we considered whether FOS caused a standard classifier to change the selection of the features that it used to formulate its decision boundary.

Setup. All experiments were performed using five fold cross-validation. The reported results are averaged over the respective held out validation sets. See Tables 3 and 4.

Table 3 Results for SVM classifier

Method	Bal Acc	Avg odds	Absol avg odds	TNRD	Equal odds	Fair utility
German dataset						
Base	0.6486	0.1054	0.1054	0.1358	0.0750	0.5795
SMOTE	0.7055	0.0500	0.0531	0.0812	0.0250	0.6677
Reweigh	0.6883	0.0411	0.0492	0.0506	0.0478	0.6546
Disparate	0.6152	0.0745	0.0788	0.1071	0.0504	0.5668
FCB	0.6031	0.1199	0.1199	0.0430	0.0648	0.5301
FOS	0.7003	0.0174	0.0313	0.0166	0.0460	0.6781
Adult dataset						
Base	0.7052	0.1886	0.1886	0.0666	0.3106	0.5722
SMOTE	0.8044	0.3094	0.3094	0.2934	0.3255	0.5527
Reweigh	0.7667	0.0626	0.0639	0.1174	0.0103	0.7177
Disparate	0.7259	0.4378	0.4378	0.2687	0.6070	0.4035
FCB	0.6276	0.0318	0.0318	0.1638	0.0116	0.6076
FOS	0.7935	0.0208	0.0422	0.0247	0.0598	0.7600
Compas dataset						
Base	0.6641	0.2022	0.2022	0.1541	0.2503	0.5296
SMOTE	0.6700	0.2176	0.2176	0.1835	0.2517	0.5242
Reweigh	0.6447	0.0319	0.0353	0.0462	0.0244	0.6220
Disparate	0.6613	0.0486	0.0500	0.0593	0.0407	0.6280
FCB	0.6425	0.1136	0.1136	0.2302	0.1060	0.5692
FOS	0.6680	0.0125	0.0246	0.0267	0.0224	0.6512

Table 4 Results for Logistic Regression classifier

Method	Bal Acc	Avg odds	Absol avg odds	TNRD	Equal odds	Fair utility
German dataset						
Base	0.6469	0.0886	0.1084	0.0730	0.1438	0.5756
SMOTE	0.7158	0.0659	0.0867	0.0688	0.1046	0.6536
Reweight	0.6099	0.0679	0.0679	0.0387	0.0970	0.5683
Disparate	0.6385	0.0648	0.0793	0.0417	0.1169	0.5875
FCB	0.6343	0.1020	0.1020	0.0540	0.0542	0.5694
FOS	0.7144	0.0452	0.0563	0.0414	0.0712	0.6746
Adult dataset						
Base	0.6765	0.1960	0.1960	0.0726	0.3193	0.5439
SMOTE	0.7595	0.3238	0.3238	0.3776	0.2700	0.5136
Reweight	0.6941	0.0080	0.0113	0.0056	0.0170	0.6862
Disparate	0.6532	0.0359	0.0359	0.0156	0.0561	0.6298
FCB	0.7547	0.1498	0.1498	0.0461	0.4234	0.6417
FOS	0.7352	0.0137	0.0158	0.0162	0.0153	0.7236
Compas dataset						
Base	0.6649	0.2129	0.2129	0.2704	0.1553	0.5230
SMOTE	0.6733	0.2271	0.2271	0.2578	0.1963	0.5200
Reweight	0.6669	0.0170	0.0285	0.0327	0.0243	0.6476
Disparate	0.6675	0.0844	0.0844	0.1027	0.0661	0.6111
FCB	0.6502	0.1493	0.1493	0.1986	0.1111	0.5531
FOS	0.6674	0.0139	0.0277	0.0244	0.0309	0.6487

6.3 RQ1: DA for standard classifiers

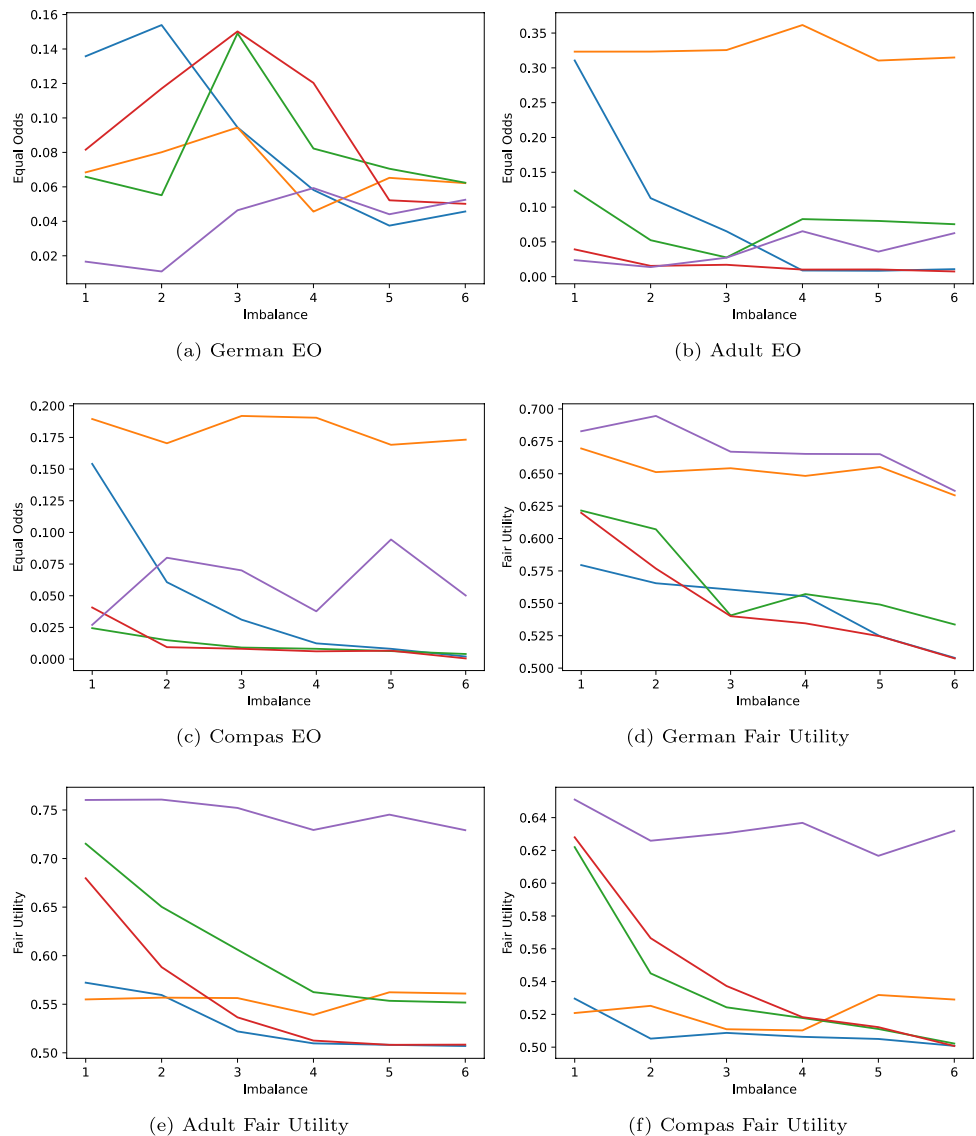
FOS displays strong performance with respect to the standard classifiers, with clear improvements in fairness, as measured by AOD, AAO, TNRD, and EOD; although it shows better results with SVM as compared to Logistic Regression. For the SVM classifier, it consistently outperforms the other algorithms in terms of average odds, absolute average odds, TNRD, and Fair Utility. See Table 3. It came in a close second to SMOTE in terms of Balanced Accuracy and clearly outperformed SMOTE in terms of the fairness metrics. Although SMOTE displays strong balanced accuracy, it often does not produce fair results with respect to protected groups.

This is likely because it balances the *class* distribution, which improves the class false positive rate; while it is not designed to improve the false positive rate with respect to specific instance features. In terms of equal odds, FOS demonstrates significant reductions in unfairness compared to the baseline, with a first place finish for Compas and second place finishes for German Credit and Adult Census.

For Logistic Regression, FOS consistently produced the top Fair Utility results, with first place finishes in terms of average odds and absolute average odds, and first place misses by less than .0057 points. It also showed significant reductions in equal odds and TNRD, when compared to baselines, with first or second place results. See Table 4.

For purposes of this experiment, FOS consistently demonstrates that it improves both accuracy and fairness over baselines. It also outperforms other fairness pre-processing algorithms on a number of measures. Thus, this experiment shows that an oversampling technique that is adopted from imbalanced learning can achieve significant improvements in group fairness measures. This also shows the close relationship between class and protected group imbalance and fairness—by jointly improving class and protected group imbalance ratios, we can affect a substantial improvement in group fairness measures. These results also indicate that it is possible to increase both balanced accuracy and fairness simultaneously (**RQ1 answered**).

Fig. 1 Impact of varying imbalance levels on Equal Odds (EO) and Fair Utility for a SVM classifier for 3 datasets. FOS shows high resilience to increasing imbalance levels



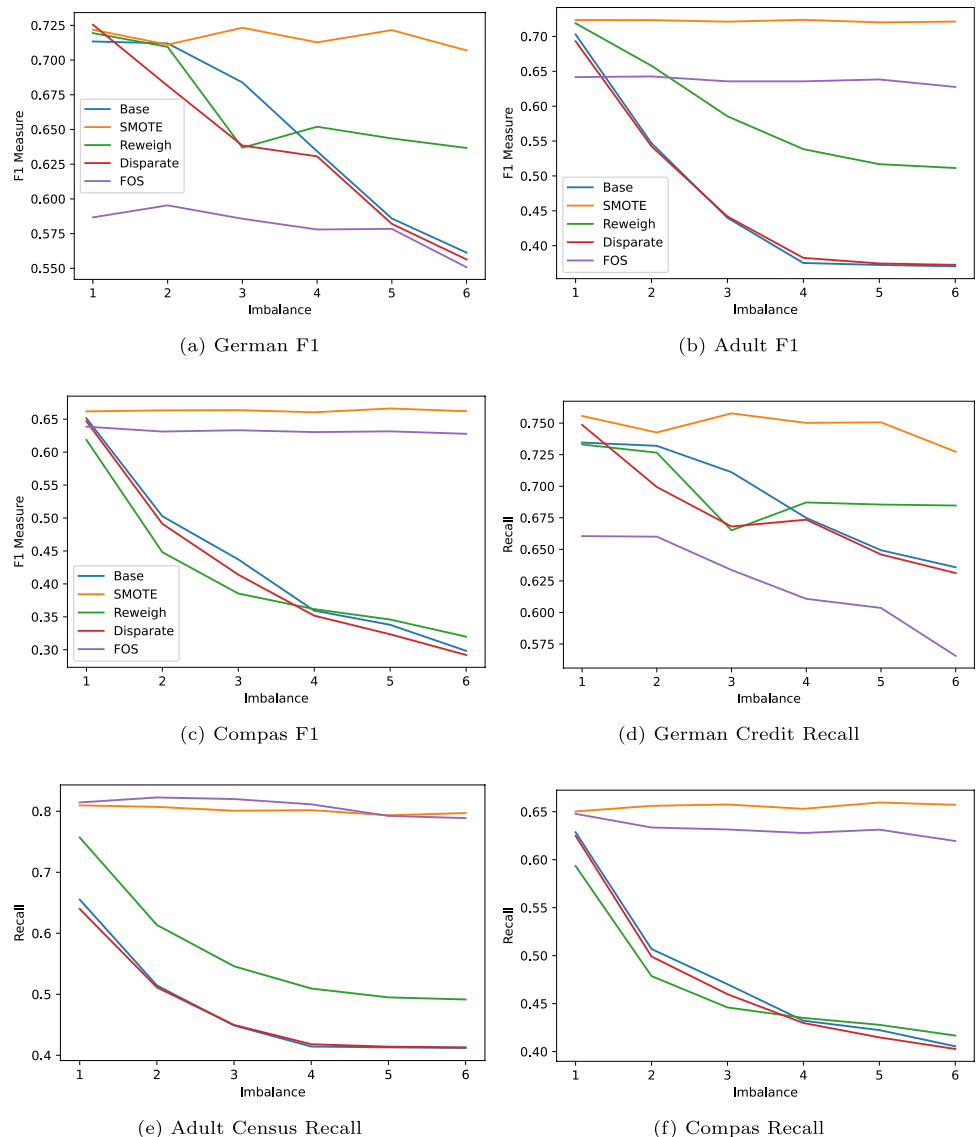
6.4 RQ2: robustness to increasing imbalance ratios

FOS performs at the top of the benchmark group in terms of Balanced Accuracy and Fair Utility under increasing levels of imbalance on all three datasets. See Fig. 1. However, at first glance, it does not outperform in terms of discrimination mitigation at higher levels of imbalance. Upon closer inspection, we believe that the reason why the baseline and other algorithms appear to be more stable at higher imbalance ratios is because their predictions focus on the true positives at the expense of true negatives. This can be seen in the Adult Census and Compas datasets, which have higher levels of imbalance. In those cases, as depicted in Fig. 2, the precision ratios increased and the recall ratios decreased for most algorithms, except for FOS and SMOTE (RQ2 answered). As discussed in the Experiments section, it should be remembered that other pre-processing techniques initially failed at imbalance levels greater than 2 and 4 on the German Credit and Compas datasets, respectively.

6.5 RQ3: impact of FOS on feature importance

Figure 3 displays the importance of each feature for logistic regression models. It compares feature importance for models trained with FOS and baseline imbalanced datasets. Feature importance is measured based on the absolute value

Fig. 2 This figure illustrates the impact of increasing imbalance levels on the F1 and recall measures. For the Compas and Adult Census datasets, which experience relatively more class and protected group imbalance, FOS shows greater resilience



of model weights. Because a LG model is shallow, there is a direct correspondence between features and weights. More important features have higher weight magnitudes. Since the model uses a single classification layer with summation, both negatively and positively signed weights are equally important; thus, we take the absolute value of the weights. The magnitudes are averaged across 5 cross-validation runs. In all cases, FOS changes the magnitude, and sometimes the sequence, of feature importance (model weights).

Importantly, FOS increases the magnitude of *almost every* feature, except one—gender. Thus, FOS reduces the magnitude, or importance, of the protected feature (sex). By reducing the importance of the protected feature, it reduces model bias toward that attribute. **(RQ3 answered.)**

7 Conclusion

A key facet in reducing algorithmic discrimination is the simultaneous reduction of class and protected feature bias in training data. We show that reducing data imbalance facilitates improvements in model accuracy and that debiasing protected features improves group fairness. We discussed the importance of bridging imbalanced learning and group fairness, by showing how key concepts in these fields overlap, and proposed a novel oversampling algorithm, *Fair*

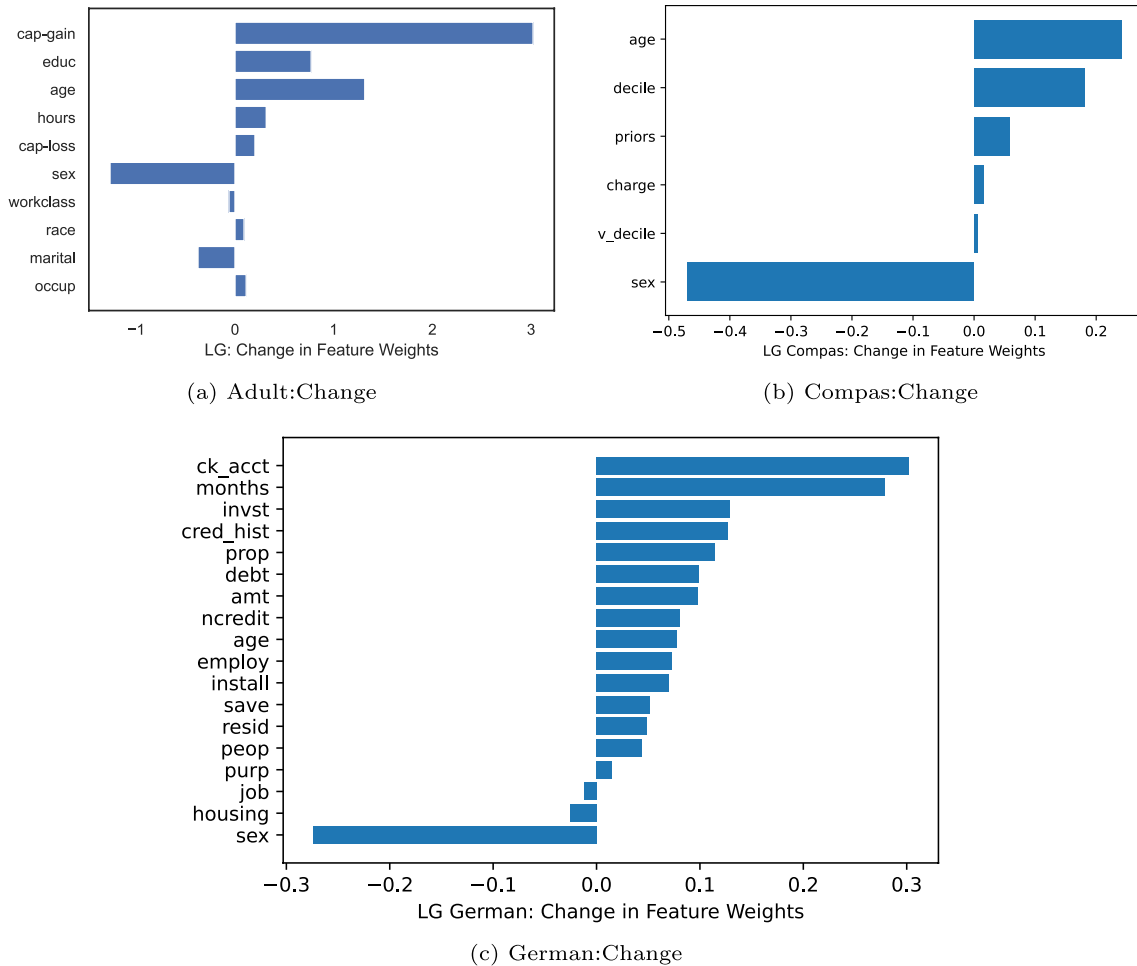


Fig. 3 These figures show that FOS generally *increases* the magnitude of all logistic regression model coefficients (except one), which is the *opposite* of weight regularization. Thus, it is able to improve sensitive feature and minority class generalization *without* weight regularization; notably, it markedly reduces the feature importance of gender (sex), which is the tested protected feature here

Oversampling, that addresses both class and protected feature bias. We take a step toward bridging the gap between fairness and imbalanced learning with a new metric, *Fair Utility*, that combines balanced accuracy with group fairness measures.

Author contributions All authors read and wrote the paper.

Funding No funding.

Data availability Data and code are publicly available.

Declarations

Ethics approval and consent to participate Not applicable. All authors consent to participate.

Consent to publication All authors consent to publication.

Competing interests The authors declare that they have no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article

are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. Smote: synthetic minority over-sampling technique. *J Artif Intell Res.* 2002;16:321–57.
2. Khandani AE, Kim AJ, Lo AW. Consumer credit-risk models via machine-learning algorithms. *J Banking Finance.* 2010;34(11):2767–87.
3. Berk R, Heidari H, Jabbari S, Kearns M, Roth A. Fairness in criminal justice risk assessments: the state of the art. *Sociol Methods Res.* 2021;50(1):3–44.
4. Chouldechova A, Benavides-Prado D, Fialko O, Vaithianathan R. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In: *Conference on Fairness, Accountability and Transparency*, 2018; pp. 134–148. PMLR.
5. Schumann C, Foster J, Mattei N, Dickerson J. We need fairness and explainability in algorithmic hiring. In: *International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS) 2020*.
6. Shroff R. Predictive analytics for city agencies: lessons from children's services. *Big Data.* 2017;5(3):189–96.
7. Dastin J. Amazon scraps secret AI recruiting tool that showed bias against women (2018). <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G>.
8. Raji ID, Gebru T, Mitchell M, Buolamwini J, Lee J, Denton E. Saving face: investigating the ethical concerns of facial recognition auditing. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020; pp. 145–151.
9. Caliskan A, Bryson JJ, Narayanan A. Semantics derived automatically from language corpora contain human-like biases. *Science.* 2017;356(6334):183–6.
10. Barocas S, Selbst AD. Big data's disparate impact. *Calif L Rev.* 2016;104:671.
11. Commission E. Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) 2021: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206>.
12. Kleinberg J. Inherent trade-offs in algorithmic fairness. In: *Abstracts of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems*, pp. 40–40 2018.
13. Flores AW, Bechtel K, Lowenkamp CT. False positives, false negatives, and false analyses: a rejoinder to machine bias: there's software used across the country to predict future criminals and it's biased against blacks. *Fed Probation.* 2016;80:38.
14. Hashimoto T, Srivastava M, Namkoong H, Liang P. Fairness without demographics in repeated loss minimization. In: *International Conference on Machine Learning*, 2018; pp. 1929–1938 PMLR.
15. Chouldechova A. Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. *Big Data.* 2017;5(2):153–63.
16. Corbett-Davies S, Pierson E, Feller A, Goel S, Huq A. Algorithmic decision making and the cost of fairness. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017;797–806.
17. Romei A, Ruggieri S. A multidisciplinary survey on discrimination analysis. *Knowl Eng Rev.* 2014;29(5):582–638.
18. Žliobaite I. Measuring discrimination in algorithmic decision making. *Data Min Knowl Disc.* 2017;31(4):1060–89.
19. Calmon FP, Wei D, Vinzamuri B, Ramamurthy KN, Varshney KR. Optimized pre-processing for discrimination prevention. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017;3995–4004.
20. Speicher T, Heidari H, Grgic-Hlaca N, Gummadi KP, Singla A, Weller A, Zafar MB. A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018;2239–2248.
21. Narayanan A. Translation tutorial: 21 fairness definitions and their politics. In: *Proc. Conf. Fairness Accountability Transp.*, New York, USA, vol. 2018;1170.
22. Corbett-Davies S, Goel S. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* 2018.
23. Bonchi F, Hajian S, Mishra B, Ramazzotti D. Exposing the probabilistic causal structure of discrimination. *Int J Data Sci Anal.* 2017;3(1):1–21.
24. Grgic-Hlaca N, Zafar MB, Gummadi KP, Weller A. The case for process fairness in learning: Feature selection for fair decision making. In: *NIPS Symposium on Machine Learning and the Law*, 2016;1:2.
25. Zemel R, Wu Y, Swersky K, Pitassi T, Dwork C. Learning fair representations. In: *International Conference on Machine Learning*, 2013:325–333. PMLR.
26. Agarwal A, Beygelzimer A, Dudík M, Langford J, Wallach H. A reductions approach to fair classification. In: *International Conference on Machine Learning*, 2018:60–69. PMLR.
27. Calders T, Verwer S. Three naive bayes approaches for discrimination-free classification. *Data Min Knowl Disc.* 2010;21(2):277–92.
28. Edwards H, Storkey A. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897* 2015.
29. Feldman M, Friedler SA, Moeller J, Scheidegger C, Venkatasubramanian S. Certifying and removing disparate impact. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015;259–268.
30. Kamiran F, Žliobaite I, Calders T. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowl Inf Syst.* 2013;35(3):613–44.
31. Li P, Liu H. Achieving fairness at no utility cost via data reweighing with influence. In: *International Conference on Machine Learning*, 2022:12917–12930. PMLR.
32. Han H, Wang W-Y, Mao B-H. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In: *International Conference on Intelligent Computing*, 2005:878–887. Springer.
33. Bunkhumpornpat C, Sinapiromsaran K, Lursinsap C. Safe-level-smote: safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2009:475–482. Springer.

34. Dablain D, Krawczyk B, Chawla NV. DeepSMOTE: fusing deep learning and smote for imbalanced data. *IEEE Trans Neural Netw Learn Syst.* 2022. <https://doi.org/10.1109/TNNLS.2021.3136503>.
35. Sleeman WC, Krawczyk B. Multi-class imbalanced big data classification on spark. *Knowl Based Syst.* 2021;212: 106598.
36. Krawczyk B, Koziarski M, Wozniak M. Radial-based oversampling for multiclass imbalanced data classification. *IEEE Trans Neural Netw Learn Syst.* 2020;31(8):2818–31.
37. Sharma S, Bellinger C, Krawczyk B, Zaiane OR, Japkowicz N. Synthetic oversampling with the majority class: a new perspective on handling extreme imbalance. In: *IEEE International Conference on Data Mining, ICDM 2018, Singapore, November 17-20, 2018, IEEE Computer Society, 2018:447–456.*
38. Bej S, Davtyan N, Wolfien M, Nassar M, Wolkenhauer O. Loras: an oversampling approach for imbalanced datasets. *Mach Learn.* 2021;110(2):279–301.
39. Kahn H, Marshall AW. Methods of reducing sample size in monte Carlo computations. *J Oper Res Soc Am.* 1953;1(5):263–78.
40. Lin T-Y, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: *Proceedings of the IEEE International Conference on Computer Vision, 2017:2980–2988.*
41. Cui Y, Jia M, Lin T-Y, Song Y, Belongie S. Class-balanced loss based on effective number of samples. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019:9268–9277.*
42. Cao K, Wei C, Gaidon A, Arechiga N, Ma T. Learning imbalanced datasets with label-distribution-aware margin loss. *arXiv preprint arXiv:1906.07413.* 2019.
43. Ridnik T, Ben-Baruch E, Zamir N, Noy A, Friedman I, Protter M, Zelnik-Manor L. Asymmetric loss for multi-label classification. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021:82–91.*
44. Woźniak M, Grana M, Corchado E. A survey of multiple classifier systems as hybrid systems. *Inf Fusion.* 2014;16:3–17.
45. Lango M, Stefanowski J. Multi-class and feature selection extensions of roughly balanced bagging for imbalanced data. *J Intell Inf Syst.* 2018;50(1):97–127.
46. Zhang W, Ramezani R, Naeim A. Wotboost: Weighted oversampling technique in boosting for imbalanced learning. In: *2019 IEEE International Conference on Big Data (IEEE BigData), Los Angeles, CA, USA, December 9-12, 2019, pp. 2523–2531. IEEE, 2019.*
47. Tao X, Li Q, Guo W, Ren C, Li C, Liu R, Zou J. Self-adaptive cost weights-based support vector machine cost-sensitive ensemble for imbalanced data classification. *Inf Sci.* 2019;487:31–56.
48. Zybiewski P, Sabourin R, Wozniak M. Preprocessed dynamic classifier ensemble selection for highly imbalanced drifted data streams. *Inf Fusion.* 2021;66:138–54.
49. Shui C, Xu G, Chen Q, Li J, Ling CX, Arbel T, Wang B, Gagné C. On learning fairness and accuracy on multiple subgroups. *Adv Neural Inf Process Syst.* 2022;35:34121–35.
50. Subramanian S, Rahimi A, Baldwin T, Cohn T, Frermann L. Fairness-aware class imbalanced learning. *arXiv preprint arXiv:2109.10444* 2021.
51. Yan S, Kao H-t, Ferrara E. Fair class balancing: enhancing model fairness without observing sensitive attributes. In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020:1715–1724.*
52. Douzas G, Bacao F, Last F. Improving imbalanced learning through a heuristic oversampling method based on k-means and smote. *Inf Sci.* 2018;465:1–20.
53. Ferrari E, Bacciu D. Addressing fairness, bias and class imbalance in machine learning: the fbi-loss. *arXiv preprint arXiv:2105.06345* 2021.
54. Iosifidis V, Ntoutsi E. Dealing with bias via data augmentation in supervised learning scenarios. *Jo Bates Paul D. Clough Robert Jäschke* 2018:24.
55. Chakraborty J, Majumder S, Menzies T. Bias in machine learning software: why? how? what to do? *arXiv preprint arXiv:2105.12195* 2021.
56. Wang Z, Saxena N, Yu T, Karki S, Zetty T, Haque I, Zhou S, Kc D, Stockwell I, Wang X, et al. Preventing discriminatory decision-making in evolving data streams. In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 2023:149–159.*
57. Tarzanagh DA, Hou B, Tong B, Long Q, Shen L. Fairness-aware class imbalanced learning on multiple subgroups. In: *Uncertainty in Artificial Intelligence, 2023:2123–2133. PMLR.*
58. Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. *Adv Neural Inf Process Syst.* 2016;29:3315–23.
59. Biswas S, Rajan H. Do the machine learning models on a crowd sourced platform exhibit bias? An empirical study on model fairness. In: *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2020:642–653.*
60. Bellamy RK, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, Lohia P, Martino J, Mehta S, Mojsilovic A, et al. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943* 2018.
61. Halevy A, Norvig P, Pereira F. The unreasonable effectiveness of data. *IEEE Intell Syst.* 2009;24(2):8–12.
62. Quy TL, Roy A, Iosifidis V, Ntoutsi E. A survey on datasets for fairness-aware machine learning. *CoRR abs/2110.00530* 2021. [arXiv:2110.00530](https://arxiv.org/abs/2110.00530).
63. Hofmann DH. Statlog (German Credit Data) Data Set (1994). <https://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29>.
64. Kohavi R, et al. Scaling up the accuracy of naive-bayes classifiers. A decision-tree hybrid In: *Kdd.* 1996;96:202–7.
65. Angwin J. Machine Bias: There’s software used across the country to predict future criminals and its biased against blacks 2016: <https://github.com/Trusted-AI/AIF360/blob/master/aif360/data/raw/compas/README.md>.

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.