FairBound: Boundary-Preserving Fair Downsampling for Accurate and Equitable Imbalanced Classification

Chrysostomos Kalousios*, Kenji Kobayashi*, Virginia Ghiara*, and Hiroya Inakoshi†
*Fujitsu Research of Europe, Slough, UK
†Fujitsu Ltd., Kawasaki, Japan
Email: {chrysostomos.kalousios, kobayashi_kenji, virginia.ghiara, inakoshi.hiroya}@fujitsu.com

***Abstract*—We present a fair downsampling classification method applicable to a wide range of imbalanced datasets. The imbalance appears both at the class (majority vs. minority) as well as at the group level (e.g. gender, age, race). In traditional downsampling methods, such as Near Miss, there is usually a trade-off between accuracy and fairness, and in many cases downsampling deteriorates fairness. We propose FairBound, a novel downsampling method designed to prevent underfitting while balancing fairness and accuracy by retaining samples near the decision boundaries. We demonstrate the utility of our method in several examples.**

***Index Terms*—Fairness, Imbalanced Classification, Downsampling, Machine Learning**

I. Introduction

Machine learning classifiers have been widely used in various decision-making systems. Due to their impact on our lives, there are concerns regarding fairness, as the resulting models can unintentionally discriminate against groups. There are many application domains that range from the medical sector (e.g. [1], [2]) to education (e.g. [3], [4]) and the banking sector (e.g. [5]), where fairness is of crucial importance (diabetes prediction, performance evaluation of employees, credit lending, fraud detection, medical diagnosis, and intrusion detection).

A key source of unfairness is dataset imbalance, where certain classes (majority classes) appear with more instances than other classes (minority classes). Many machine learning algorithms assume that the distribution of samples is balanced [6] (see also [7], [8]). In an imbalanced dataset the algorithm usually spends less training time with the minority classes, thus making it difficult to build rules and prediction boundaries for those classes. As a result, samples from the minority classes are most often misclassified [6]. This can be a problem because, when working with an imbalanced system, practitioners are usually interested in predictions related to minority classes.

An example of the negative impact of a system trained on imbalanced datasets can be seen in algorithms trained on diagnostic patterns from X-ray imaging, which sometimes inherit biases due to the low representation of female patients. As a result, diagnostic algorithms tend to be less accurate for females compared to males. Reducing the risk of bias by making the datasets more gender-balanced not only improves fairness, but also allows the algorithms to perform better for both males and females [9].

It is important to recognize that identifying privileged and unprivileged groups in a dataset requires a deep understanding of the specific domain in question. While legally defined protected characteristics (such as age, gender, race, religion, disability, and national origin, as outlined in the UK *Equality Act 2010* [10] and the *European Convention on Human Rights* [11]) provide a foundational starting point, further guidance is available to support this process. For example, the Alan Turing Institute's *AI Ethics and Governance in Practice: AI Fairness in Practice* [12] workbook offers practical tools for reflecting on marginalised, vulnerable, historically discriminated against, or disadvantaged social groups.

Given the serious implications of biased algorithms, regulatory bodies have begun to intervene. In recent years, national and international regulators have introduced obligations for organizations to assess and mitigate biases in AI systems. For instance, Article 10 (Data and Data Governance) of the EU AI Act [13] requires that training, validation and testing datasets are subject to data governance practices to ensure they are representative and free of biases; while in Hong Kong, the Office of the Privacy Commissioner for Personal Data (PCPD) released the *Model Personal Data Protection Framework* [14] in June 2024, highlighting dataset imbalance as a source of unjust biases and recommending sampling techniques such as downsampling (p. 36) to address representation gaps. In the UK, the Information Commissioner's Office (ICO) included recommendations to address dataset imbalance in its *Guidance on AI and Data Protection* [15], such as adjusting the data by adding or removing samples from underrepresented or overrepresented population subsets to ensure fairer outcomes. Similar considerations have been included in sectorial guidelines, such as the *Artificial Intelligence Governance Principles* published by the European Insurance and Occupational Pensions Authority (EIOPA) [16]. Moreover, the study *Auditing the Quality of Datasets Used in Algorithmic Decision-making Systems* [17] published by the Panel for the Future of Science and Technology (STOA) investigated the possibility of introducing, in the European Union, certificates for datasets with the objective of avoiding biases.

In order to effectively work with an imbalanced dataset,

several approaches have been developed, such as modification of the dataset itself (e.g. downsampling, also known as undersampling), cost optimization methods and ensemble methods that combine weak learners to predict the minority class.

For downsampling, which is the main focus of this work, many different techniques have been developed, including a) randomly deleting samples from the majority class, b) methods that select samples to keep in the majority class (*Near Miss* and *Condensed Nearest Neighbour Rule*), c) methods that select samples to delete from the majority class (*Tomek Links* and *Edited Nearest Neighbours Rule*), and d) methods that combine the two previous approaches (*One-Sided Selection* and *Neighbourhood Cleaning Rule*). Random Downsampling and Near Miss allow the user to specify the desired amount of downsampling. Other downsampling methods clean the feature space by removing either noisy observations or observations that are too easy to classify. The open source library *imblearn* [18] contains a selection of common algorithms that deal with imbalanced datasets and can be also consulted for references on the aforementioned algorithms.

Traditional downsampling algorithms, such as those mentioned earlier, typically demonstrate a trade-off between accuracy and fairness (accuracy would improve at the cost of fairness). The reason for this is that downsampling algorithms only consider class imbalances, thus ignoring group imbalances.

In this work we focus on improving fairness in downsampling by proposing an effective boundary-preserving algorithm. Literature on fair downsampling appears to be more limited than the broader literature on imbalanced learning and fair pre-processing. One possible reason is that downsampling removes information and is therefore often treated more cautiously than other mitigation strategies.

Existing work, including [19], has primarily focused on random removal of samples from the majority class to create a more balanced dataset. In such approaches, fairness improvements typically arise from equalizing class and group distributions. Our contribution differs in that the proposed method performs boundary-aware selection within subgroup-specific subsets, with the aim of preserving informative samples while improving fairness.

We argue that further research into fair downsampling is necessary, particularly when guided by socio-technical considerations. Datasets used to train machine learning models are not neutral; they are socially constructed artifacts that can reflect prevailing power structures and systemic biases. As such, they should be critically examined and managed in ways that recognize their social and ethical implications. Likewise, data transformation, model development, and the practices of data scientists are all embedded within a broader socio-technical context [20]. Developing a downsampling technique that accounts for fairness without compromising accuracy is therefore not only a technical challenge, but also an ethical one, because such techniques can shape outcomes in areas such as healthcare, criminal justice, and employment. Ensuring fairness in these contexts is a matter not only of model performance, but also of social responsibility.

Apart from balancing the size of groups and classes, benefits can also be obtained from preserving the boundaries, while deleting samples away from them. It is known that downsampling (and by extension fair downsampling) can lead to underfitting (see for example [21]).

By considering both group/class imbalance and decision boundaries at the same time, we propose a general-purpose Boundary-Preserving Fair Downsampling (FairBound) method. It is based on splitting the given dataset into smaller subsets and performing downsampling between those smaller subsets. The key idea is to equalize the number of records between classes (such as majority and minority) as well as between the two groups defined for a given protected attribute, while considering boundaries between groups and classes and deleting samples away from them. Our results show that our method improves fairness while maintaining a competitive level of accuracy.

The rest of the paper is organized as follows. In the next section we explain our notation and define the utility and fairness metrics used in this work. Our method is explained in the following section, entitled Method. In the section entitled Experiments we describe the various datasets used in this work, the parameters used for training our models, the experiments performed, and the results obtained. We finish our work with Conclusions.

II. Preliminaries

We define $\hat{Y}$ to be the predicted outcome of a classifier and Y to be the actual labels of the dataset. We further define the following abbreviations TP: True Positive, FN: False Negative, FP: False Positive, TN: True Negative.

Using the aforementioned quantities, we can construct other useful metrics such as TPR, the True Positive Rate (also known as sensitivity or recall), and FPR, the False Positive Rate, defined as

$$\mathrm{TPR} \equiv \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \quad \mathrm{FPR} \equiv \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}.$$

In order to measure the accuracy of our models we use the **Balanced Accuracy** (BA) metric, appropriate for imbalanced datasets. It is defined as

$$\mathrm{BA} \equiv \frac{1}{2}\left(\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} + \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}\right).$$

We now introduce three popular fairness metrics used for the evaluation of our models.

Statistical Parity Difference (SPD) [22]: privileged and unprivileged groups should receive an equal proportion of positive labels. This can be done by computing their difference

$$\mathrm{P}(\hat{Y}=1|A=0) - \mathrm{P}(\hat{Y}=1|A=1),$$

where the positive rate P is defined

$$\mathrm{P}(\hat{Y}=1|A) \equiv \frac{\mathrm{TP}+\mathrm{FP}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}},$$

Algorithm 1 Boundary-Preserving Fair Downsampling (Fair-Bound)

Input: Training dataset D
Hyperparameters: ϵ (amount of downsampling), n (number of closest or farthest points)
Output: Downsampled dataset

1: Split D into majority, M, and minority, m, classes.
2: Split D into privileged, A_1, and unprivileged, A_0, groups.
3: Split $M \cap A_0$ randomly without replacement into three (all boundaries) subsets, S_i, $i = 1, 2, 3$, of equal size.
4: Split $M \cap A_1$ randomly without replacement into three (all boundaries) subsets, T_i, $i = 1, 2, 3$, of equal size.
5: • Version 1: $\forall j \in S_1$ find the n *closest* points from $m \cap A_0$ and calculate their average distance $\bar{d}_j$, then keep as many points of S_1 with the smallest $\bar{d}_j$ as dictated by ϵ.
• Version 2: $\forall j \in S_1$ find the n *farthest* points from $m \cap A_0$ and calculate their average distance $\bar{d}_j$, then keep as many points of S_1 with the smallest $\bar{d}_j$ as dictated by ϵ.
6: Repeat the previous step between S_2 and $m \cap A_1$, between T_1 and $m \cap A_0$, and between T_2 and $m \cap A_1$.
7: Downsample S_3 into $\tilde{S}_3$ using T_3. Downsample T_3 into $\tilde{T}_3$ using the original S_3. Obtain $\tilde{S}_3 \cup \tilde{T}_3$.
8: Return union of all downsampled datasets.

and the parameter A refers to the demographic groups and takes the value 1 for the privileged group and 0 for the unprivileged group.

Equal Opportunity Difference (EOD) [23]: it measures equality of TPR across population groups and is defined as

$$\mathrm{P}(\hat{Y} = 1 | A = 0, Y = 1) - \mathrm{P}(\hat{Y} = 1 | A = 1, Y = 1).$$

Average Odds Difference (AOD) [23]: It measures equality of TPR and FPR across population groups and is defined as

$$\frac{1}{2}\Big(\mathrm{P}(\hat{Y} = 1 | A = 0, Y = 1) - \mathrm{P}(\hat{Y} = 1 | A = 1, Y = 1) + \mathrm{P}(\hat{Y} = 1 | A = 0, Y = 0) - \mathrm{P}(\hat{Y} = 1 | A = 1, Y = 0)\Big).$$

The focus of our proposed method on equalizing the number of records between classes and groups directly relates to the idea of demographic parity, which Statistical Parity Difference measures. Similarly, by preserving boundaries, our method aims to reduce disparities in True Positive Rate and False Positive Rate across groups, as reflected by Equal Opportunity Difference and Average Odds Difference. Since fairness is multi-dimensional, improvements in one metric do not necessarily imply improvements in the others; for this reason we report all three.

III. Method

In this section we describe our fair downsampling algorithm. The main idea and design philosophy of our proposed methodology is to keep data points close to the decision boundaries

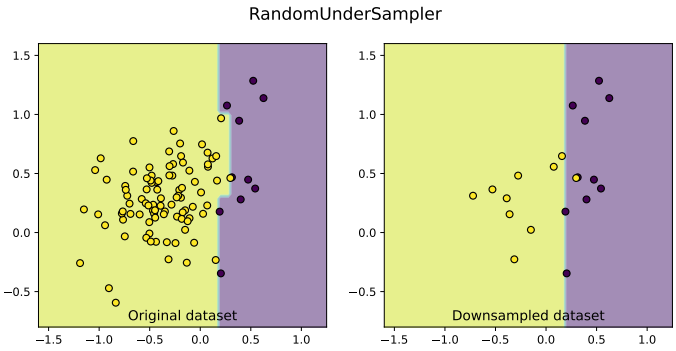


Fig. 1: Demonstration of underfitting in the case of Random Downsampling using the Decision Tree classifier and a simple synthetic dataset. The decision boundary can be thought of as separating either different classes or different groups.

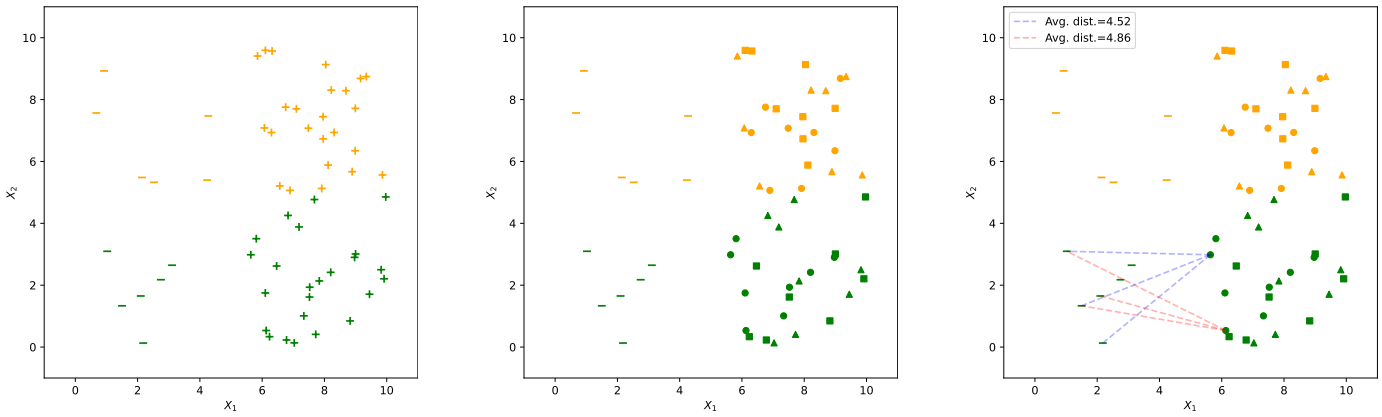
while deleting samples away from them. Intuitively, points near class or group boundaries are more likely to influence classification errors and disparity gaps, so removing mainly points far from those boundaries can preserve decision structure while reducing imbalance. By decision boundaries we refer to boundaries that either separate classes or groups. This offers an improvement compared to traditional downsampling methods, where only decision boundaries between classes are considered, thus negatively impacting the fairness of the algorithm. Keeping data points near the boundaries is also encouraged by the fact that downsampling can lead to underfitting. As an example, in Fig. 1 we demonstrate the phenomenon in the case of a simple synthetic dataset. The classifier used is Decision Tree and the downsampling is done by randomly deleting samples. After downsampling, the decision boundary has been reduced to a straight line, potentially leading to accuracy and fairness loss.

Our proposed method is described in Algorithm 1, and below we explain its main steps and our design choices.

Hyperparameters: Our algorithm introduces two hyperparameters. The first one expresses the amount of downsampling (also known as *sampling strategy*), ϵ, and it is defined as the ratio of the number of samples of the minority class in the original dataset to the number of samples of the majority class after downsampling. The second hyperparameter, n, expresses the number of nearest or farthest points chosen in Line 5 of our algorithm. Although the optimal choice of the hyperparameters depends on the dataset at hand and is subject to fine-tuning, our experiments suggest that values around $\epsilon \approx 1$ and $n \approx 3$ can serve as reasonable starting points.

Step 1 (lines 1-2): Given an imbalanced dataset D we first split it into the majority (+) and the minority (-) class, that we respectively denote as M and m. We also split samples that belong to the protected attribute into two different sets (privileged and unprivileged groups), that we denote as A_1 and A_0 respectively (see Fig. 2a). After splitting our initial dataset D we end up with the following four subsets, $M \cap A_0$, $M \cap A_1$, $m \cap A_0$, and $m \cap A_1$ whose union is D.

Step 2 (lines 3-4): Our downsampling proposal reduces the



(a) Step 1 (b) Step 2 (c) Step 3

Fig. 2: Step 1: Split a given imbalanced dataset into classes and groups. The two classes (majority and minority) are shown with (+) and (-), whereas the two groups are distinguished by different colours. X_1 and X_2 represent a two-dimensional feature space. Step 2: Further split the majority classes into three subsets. Step 3: We calculate and compare the average distance of two points with their $n = 3$ farthest points.

size of the subsets that contain points belonging to the majority classes. In our notation, these correspond to points in the two subsets, $M \cap A_0$ and $M \cap A_1$. We now further split each of $M \cap A_0$ and $M \cap A_1$, into three non-overlapping sets of equal size, that we respectively denote as S_i and T_i with $i = 1, 2, 3$. In Fig. 2b the splitting is demonstrated using three different shapes (triangles, squares, and circles).

A few comments regarding our design choices. We split $M \cap A_0$ and $M \cap A_1$ into three subsets to address three distinct types of decision boundaries: intra-group (S_1, T_2) for group-specific accuracy, cross-group (S_2, T_1) for intersectional fairness, and demographic (S_3, T_3) to balance the majority class itself.

We have chosen the subsets S_i (and similarly T_i) to be of equal size because this is one of the simplest choices and it treats the three boundary types symmetrically. Alternatively, data-dependent splitting strategies are possible, but we do not further discuss them in this work.

Finally, the splitting is done at random without replacement. We have chosen random splitting so that the derived subsets follow the distribution of the original subsets, $M \cap A_0$ and $M \cap A_1$, as much as possible. We use sampling without replacement to give all data points a chance of being removed.

Step 3 (line 5): This step describes how downsampling is performed. We introduce two different versions of our algorithm that differ only in whether we consider the closest points (Version 1) or the farthest points (Version 2).

For each data point in S_1 we find its n closest (farthest) points from the subset $m \cap A_0$ and calculate its average distance, $\bar{d}$. We keep only those points of S_1 with the smallest average distance. The exact number of points we keep is determined by the chosen amount of downsampling ϵ. The number n is conceptually akin to the number of neighbours in the `KNeighborsClassifier` classification algorithm (see [24]).

As an illustration, in Fig. 2c we calculated the average distances of two points with their $n = 3$ farthest points from $m \cap A_0$. We keep the point with the smallest average distance (Avg. dist. = 4.52) and delete the one with Avg. dist. = 4.86.

We now compare the two versions. Version 1 prioritizes points that remain close to the opposite class or group and is therefore more directly aligned with our objective of boundary preservation. Version 2 provides an alternative ranking criterion that may be useful when stronger class separation is preferred. In this work we report Version 1 because it matches our design objective more closely.

Step 4 (line 6): Using the same methodology presented in Step 3, we downsample the following three subsets, S_2 with $m \cap A_1$, T_1 with $m \cap A_0$, and T_2 with $m \cap A_1$.

Step 5 (line 7): So far we have downsampled the subsets S_1, S_2, T_1, and T_2. The final subsets to downsample are S_3 and T_3. We note that the subsets S_3 and T_3 have no common data points. In this final case, care must be taken because both S_3 and T_3 will be downsampled. We proceed as follows. We first downsample S_3 using T_3 and we denote the downsampled subset S_3 as $\tilde{S}_3$ (as in Step 3 this is done by considering the average distance of each data point in S_3 from its n closest (farthest) points from the subset T_3 and keeping only the points with the smallest average distance). Finally, we downsample T_3 using the original S_3 and we denote the downsampled subset as $\tilde{T}_3$. We obtain $\tilde{S}_3 \cup \tilde{T}_3$.

Step 6 (line 8): In this final step we collect the union of the sets obtained from the output of Steps 3-5. Explicitly, these are $m \cap A_0$, $m \cap A_1$, the three downsampled S_i and the three downsampled T_i.

Intersectional fairness: Algorithm 1 is presented for binary protected groups. A simple extension to an intersectional setting is to define composite groups, for example by treating one dominant intersection as privileged and aggregating the remaining intersections as unprivileged. A fuller treatment of multiple protected attributes and finer-grained intersectional

evaluation is not considered in this work.

IV. Experiments

A. Datasets

We demonstrate the applicability of our algorithm using three datasets with varying numbers of samples, imbalance levels, and features. These are

- **ACS Income** [25] which is often used as a modern alternative to the *Adult* dataset [26] and predicts whether income exceeds \$70K/yr based on 2018 census data. It contains 185489 samples and 11 features. We consider *race* as the protected attribute and *white* as the privileged group.
- **Bank Marketing** [27] contains data from a direct marketing campaign of a Portuguese banking institution. The goal is to predict if the client will subscribe to a term deposit. The protected attribute is *age* and the privileged group is $25 \leq \text{age} < 60$.
- **MEPS** (Medical Expenditure Panel Survey, [28]). It consists of large-scale surveys of families and individuals, medical providers, and employers, and collects data on health services used, costs and frequency of services, demographics, etc., of the respondents. The particular dataset we use contains survey data from 2015 obtained in rounds 3, 4, and 5 of Panel 19. The label is *utilization*, the protected attribute *race*, and the privileged group is *white*.

The datasets used and some of their features are summarized in the following Table I.

TABLE I: Dataset Characteristics

dataset	N_s	N_f	PA	CI	GI
ACS Income	185489	11	race	4.07	12.99
Bank	45211	16	age	7.55	16.43
MEPS	15830	138	race	4.82	1.80

N_s: number of samples, N_f: number of features, PA: protected attribute, CI: class imbalance (it expresses the ratio of the majority class to the minority class), GI: group imbalance (it expresses the ratio of the privileged group to the unprivileged group).

B. Experimental environment

We have downloaded the ACS Income dataset from [25], the Bank Marketing dataset from the *UCI Machine Learning Repository* [27], and the MEPS dataset from [28]. We have applied minimal preprocessing that involves transforming categorical to numeric variables, removing samples with missing values, deleting duplicates, and scaling data.

In order to train the models we have used the following six classifiers: Logistic Regression (LR), Random Forest (RF), XGBoost (XGB), k-nearest Neighbours (KN), Decision Tree (DT), and Multilayer Perceptron (MLP). We have used the implementations of the libraries *scikit-learn* [24] and *xgboost* [29]. We compare results using five different methods, No Downsampling (ND), Random Downsampling (RD), Near Miss (NM), Fair-SMOTE, and our Boundary-Preserving Fair Downsampling (FairBound) algorithm.

Although our focus is on downsampling methods we also included Fair-SMOTE [30] (a popular oversampling algorithm) in our analysis as a baseline. This allows us to compare our method with a state-of-the-art augmentation technique and highlight the efficiency benefits of downsampling.

Furthermore, we used the default settings for training our models. We have used 5-fold cross-validation. When necessary we have used the random seed 42 for reproducibility. We have used Version 1 of our algorithm. For the MLP we used a three-layer neural network with layer sizes 64, 32, and 1.

For the implementations of the downsampling algorithms (`RandomUnderSampler` and `NearMiss`) we have used the library *imblearn* [18].

For evaluating fairness we have used three common metrics, Statistical Parity Difference, Equal Opportunity Difference, and Average Odds Difference. For the implementation of these metrics we have used the library *aif360* [31].

C. Experimental results

In this subsection we report the results of our experiments. In Fig. 3 we show the results of accuracy and fairness metrics for our three datasets and six classification algorithms.

With appropriate tuning of the sampling-strategy hyperparameter for the different fairness metrics, we improve fairness while maintaining competitive accuracy compared to the results without downsampling. We observe that Random Downsampling tends to increase Balanced Accuracy at the cost of fairness, whereas Near Miss tends to deteriorate fairness. Our method performed well for the smaller dataset, MEPS, indicating that the range of applicability can be broad. Although Fair-SMOTE yields higher accuracy, FairBound achieves comparable and in many cases better fairness using a significantly smaller dataset.

D. Ablation studies

Sampling strategy We report how the sampling strategy affects the various metrics for a representative selection of experiments in Fig. 4. For varying sampling strategy we observe that Random Downsampling produces relatively consistent results when it comes to Balanced Accuracy, whereas both Near Miss and FairBound deteriorate performance. FairBound seems to outperform Near Miss. Regarding fairness we observe that Random Downsampling deteriorates fairness in most cases, whereas both Near Miss and FairBound improve fairness, with FairBound mostly outperforming Near Miss.

Number of neighbours We perform experiments for various numbers of neighbours in FairBound and we report the results in Fig. 5 in the case of the Bank Marketing dataset. Similar trends are also observed for the other two datasets. We do not observe a pronounced effect (perhaps because ϵ is already the dominant shaping factor).

Computational efficiency We report computational time and compare FairBound with NearMiss and Fair-SMOTE in Fig. 6. Our algorithm performs 2 to 3 times faster compared to Near

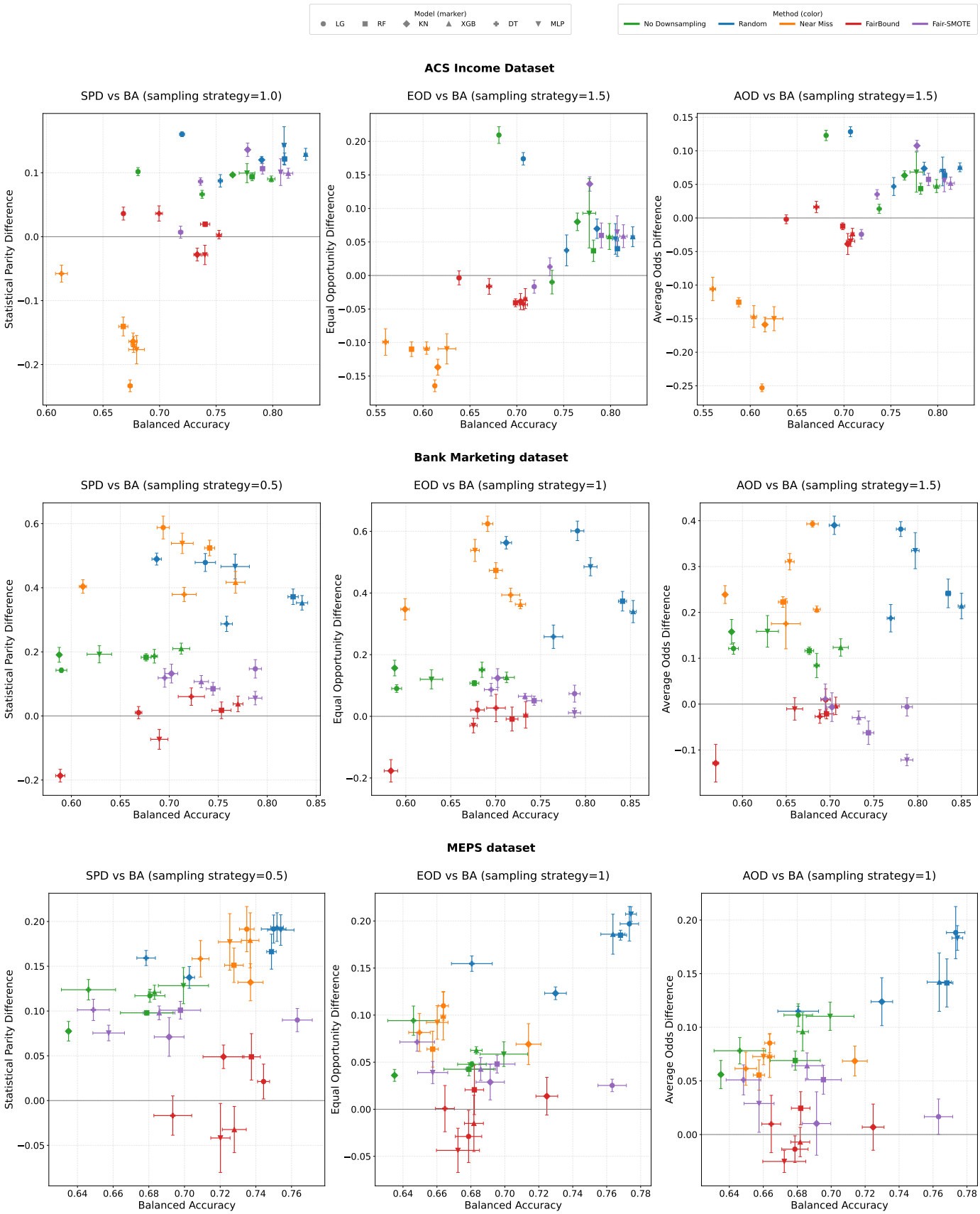


Fig. 3: Balanced Accuracy vs three common fairness metrics for three datasets and six classification algorithms.

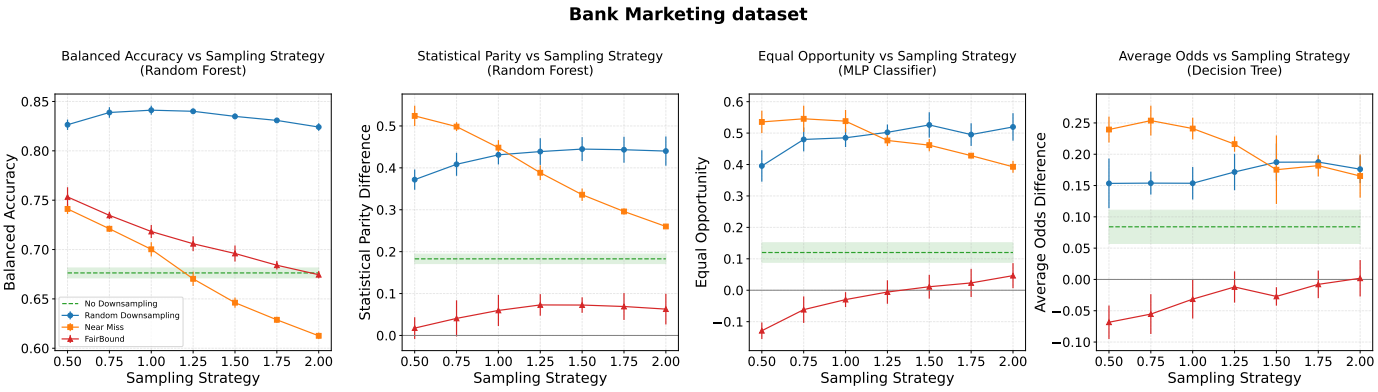


Fig. 4: Various metrics as functions of the sampling strategy.

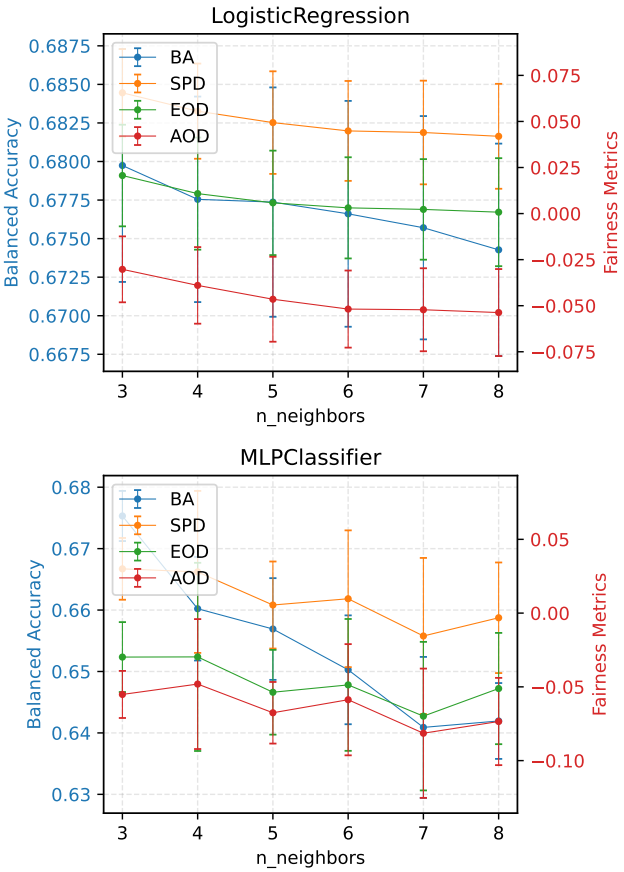


Fig. 5: Metrics vs number of neighbours for the Bank Marketing dataset.

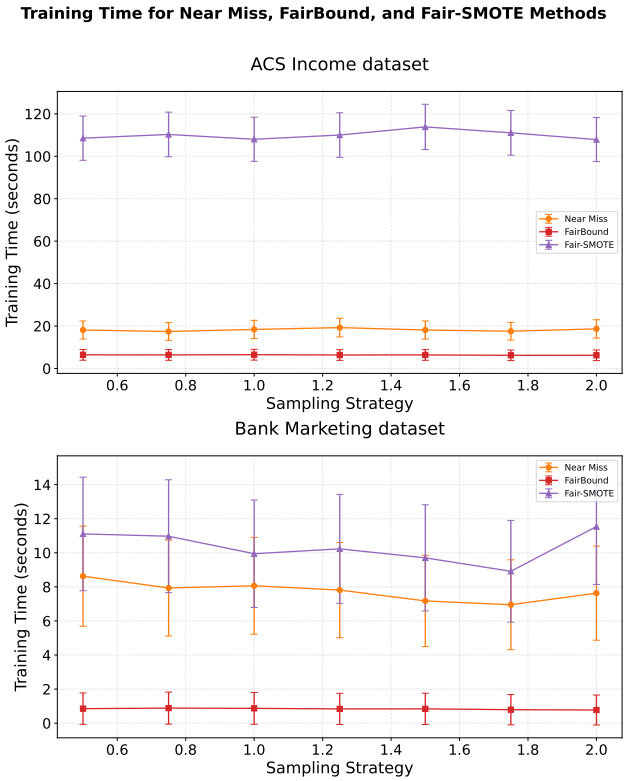


Fig. 6: Computational time for two datasets.

Miss and is up to one order of magnitude faster (depending on the size of the dataset) than the oversampling baseline, Fair-SMOTE, highlighting the scalability of our downsampling approach.

V. Conclusions

In this work we proposed a novel downsampling technique that considers fairness while maintaining competitive accuracy. In order to achieve this, we split the given datasets both into groups and classes. Then we considered all possible boundaries, deleting samples away from them in order to avoid as much as possible the problem of underfitting. The resulting dataset is smaller in size controlled by a sampling strategy hyperparameter. While oversampling methods such as Fair-SMOTE increase data volume and therefore training costs, FairBound reduces them. As a result, our approach is suitable for large-scale, resource-constrained environments with practical applications ranging from the health sector to the banking and educational sectors, where imbalanced datasets are prevalent and fairness is paramount.

We then tested our algorithm on three datasets (including one with a small number of samples). We measured the

performance using six classification algorithms, one utility metric, and three fairness metrics.

Traditional downsampling algorithms often result in a deterioration of fairness. Our approach significantly improves on that, outperforming the traditional approaches in most cases tested. By choosing different hyperparameters we were able to improve on all three fairness metrics tested. Our method performed well even with the smaller dataset (MEPS) that contains 15K samples.

FairBound offers computational advantages both at the training and inference levels. Compared to other downsampling methods, we have seen speed gains of up to three times and, compared to Fair-SMOTE, an oversampling method, of up to 10 times.

Our aim was to develop a downsampling method that preserves fairness while remaining technically effective, thereby supporting the development of more responsible AI systems. Decisions based on model predictions can have far-reaching consequences, especially in high-impact settings affected by class and group imbalance.

Datasets with a very small number of samples are generally a limitation of downsampling methods. Other limitations of our method include cases with very few samples for a particular group, the current focus on binary protected-group settings, and excessive downsampling, which could lead to significant data loss and reduced model performance.

The practical implications of our work extend to both algorithm developers and policymakers. For practitioners, our method offers actionable guidance for incorporating fairness into data preprocessing and sampling strategy design. It bridges theoretical concepts with real-world application, enabling model designers to integrate fairness without sacrificing accuracy. For policymakers and regulators, the findings inform the creation of ethical AI guidelines by shedding light on how fairness constraints affect model outcomes. In doing so, this work contributes to the ongoing effort to develop transparent, accountable, and equitable AI systems.

REFERENCES

[1] F. Dehghani, N. Malik, J. Lin, S. Bayat, and M. Bento, "Fairness in healthcare: Assessing data bias and algorithmic fairness," in *2024 20th International Symposium on Medical Information Processing and Analysis (SIPAIM)*, 2024, pp. 1–6.

[2] J. Huang, G. Galal, M. Etemadi, and M. Vaidyanathan, "Evaluation and mitigation of racial bias in clinical machine learning models: Scoping review," *JMIR Med Inform*, vol. 10, no. 5, p. e36388, May 2022. [Online]. Available: https://doi.org/10.2196/36388

[3] R. S. Baker and A. Hawn, "Algorithmic bias in education," *International Journal of Artificial Intelligence in Education*, vol. 32, no. 4, pp. 1052–1092, December 2022.

[4] H. M. Allam, B. Gyamfi, and B. AlOmar, "Sustainable innovation: Harnessing ai and living intelligence to transform higher education," *Education Sciences*, vol. 15, no. 4, 2025. [Online]. Available: https://www.mdpi.com/2227-7102/15/4/398

[5] A. Castelnovo, R. Crupi, G. D. Gamba, G. Greco, A. Naseer, D. Regoli, and B. S. Miguel Gonzalez, "Befair: Addressing fairness in the banking sector," in *2020 IEEE International Conference on Big Data (Big Data)*, 2020, pp. 3652–3661.

[6] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent Data Analysis*, vol. 6, no. 5, pp. 429–449, 2002. [Online]. Available: https://journals.sagepub.com/doi/abs/10.3233/IDA-2002-6504

[7] Fernández et al., *Learning from Imbalanced Data Sets*. Springer Cham, 2018. [Online]. Available: https://doi.org/10.1007/978-3-319-98074-4

[8] Y. M. Haibo He, Ed., *Imbalanced Learning: Foundations, Algorithms, and Applications*. Wiley, 2013.

[9] A. J. Larrazabal *et al.*, "Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis," *Proceedings of the National Academy of Sciences*, vol. 117, no. 23, pp. 12 592–12 594, 2020. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/32457147/

[10] H. Government, "Equality act 2010," https://www.legislation.gov.uk/ukpga/2010/15/contents, 2010.

[11] ECHR, "European convention on human rights," https://www.echr.coe.int/european-convention-on-human-rights.

[12] Leslie et al., "AI Fairness in Practice, The Alan Turing Institute," https://www.turing.ac.uk/sites/default/files/2023-12/aieg-ati-fairness_1.pdf, 2023.

[13] European Parliament, "Article 10: Data and data governance," https://artificialintelligenceact.eu/article/10/, 2024.

[14] Office of the Privacy Commissioner for Personal Data, "Model personal data protection framework," https://www.pcpd.org.hk/english/resources_centre/publications/files/ai_protection_framework.pdf, 2024.

[15] Information Commissioner's Office, "Guidance on ai and data protection," https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/, 2023.

[16] EIOPA, "Artificial intelligence governance principles," https://www.eiopa.europa.eu/publications/artificial-intelligence-governance-principles-towards-ethical-and-trustworthy-artificial_en, 2021.

[17] European Parliamentary Research Service, "Auditing the quality of datasets used in algorithmic decision-making systems," https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2022)729541, 2022.

[18] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn," *Journal of Machine Learning Research*, vol. 18, no. 17, pp. 1–5, 2017. [Online]. Available: http://jmlr.org/papers/v18/16-365.html

[19] T. S. Pias, Y. Su, X. Tang, H. Wang, and D. D. Yao, "Undersampling for fairness: Achieving more equitable predictions in diabetes and prediabetes," *medRxiv*, 2023. [Online]. Available: https://www.medrxiv.org/content/early/2023/05/05/2023.05.02.23289405

[20] J.-M. John-Mathews, D. Cardon, and C. Balagué, "From reality to world. a critical perspective on ai fairness," *Journal of Business Ethics*, vol. 178, no. 4, pp. 945–959, July 2022.

[21] J. Brownlee, *Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning*, 2021.

[22] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *ITCS '12*. ACM, 2012, pp. 214–226. [Online]. Available: http://doi.acm.org/10.1145/2090236.2090255

[23] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems*, vol. 29. Curran Associates, Inc., 2016. [Online]. Available: https://papers.nips.cc/paper_files/paper/2016/hash/6a9659feb1216f14f7384ba499518b38-Abstract.html

[24] Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[25] F. Ding, M. Hardt, J. Miller, and L. Schmidt, "Retiring adult: New datasets for fair machine learning," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 6478–6490.

[26] B. Becker and R. Kohavi, "Adult," https://archive.ics.uci.edu/dataset/2/adult, 1996.

[27] S. Moro, P. Rita, and P. Cortez, "Bank Marketing," https://archive.ics.uci.edu/dataset/222/bank+marketing, 2012.

[28] MEPS, "MEPS," https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-181, 2015.

[29] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. ACM, Aug. 2016. [Online]. Available: http://dx.doi.org/10.1145/2939672.2939785

[30] J. Chakraborty, S. Majumder, and T. Menzies, "Bias in machine learning software: why? how? what to do?" ser. ESEC/FSE 2021. New York, NY, USA: Association for Computing Machinery, 2021, p. 429–440.

[31] Bellamy et al., "AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," Oct. 2018. [Online]. Available: https://arxiv.org/abs/1810.01943