Fair and Private Data Preprocessing through Microaggregation

VLADIMIRO GONZÁLEZ-ZELAYA, Universidad Panamericana, Mexico
JULIÁN SALAS and DAVID MEGÍAS, Universitat Oberta de Catalunya (UOC), Spain
PAOLO MISSIER, Newcastle University, UK

Privacy protection for personal data and fairness in automated decisions are fundamental requirements for responsible Machine Learning. Both may be enforced through data preprocessing and share a common target: data should remain useful for a task, while becoming uninformative of the sensitive information. The intrinsic connection between privacy and fairness implies that modifications performed to guarantee one of these goals, may have an effect on the other, e.g., hiding a sensitive attribute from a classification algorithm might prevent a biased decision rule having such attribute as a criterion. This work resides at the intersection of algorithmic fairness and privacy. We show how the two goals are compatible, and may be simultaneously achieved, with a small loss in predictive performance. Our results are competitive with both state-of-the-art fairness correcting algorithms and hybrid privacy-fairness methods. Experiments were performed on three widely used benchmark datasets: *Adult Income*, *COMPAS,* and *German Credit*.

CCS Concepts: • **Security and privacy** → **Social aspects of security and privacy**; **Privacy protections**; **Privacy-preserving protocols**; • **Information systems** → **Data cleaning**; *Clustering;*

Additional Key Words and Phrases: Responsible machine learning, fair classification, privacy preserving data mining, algorithmic fairness, ethical AI

ACM Reference format:
Vladimiro González-Zelaya, Julián Salas, David Megías, and Paolo Missier. 2023. Fair and Private Data Preprocessing through Microaggregation. *ACM Trans. Knowl. Discov. Data.* 18, 3, Article 49 (December 2023), 24 pages.
https://doi.org/10.1145/3617377

1 INTRODUCTION AND RELATED WORK

Privacy protection for personal data and fairness in automated decisions are two fundamental demands of 21st century society. Both ideals share a common ingredient: the need to hide or protect

V. González-Zelaya and J. Salas authors contributed equally to this research.
V. González-Zelaya also with Newcastle University.
J. Salas also with Universitat Rovira i Virgili (URV).

This research was partly supported by the by the Spanish Ministry of Science and Innovation under projects RTI2018-095094-B-C21, RTI2018-095094-B-C22 “CONSENT” and PID2021-125962OB-C31 “SECURING”.

Authors’ addresses: V. González-Zelaya, Universidad Panamericana, Facultad de Ciencias Económicas y Empresariales, Mexico City, Mexico; e-mail: cvgonzalez@up.edu.mx; J. Salas and D. Megías, Universitat Oberta de Catalunya (UOC), Barcelona, Spain; e-mails: cvgonzalez@up.edu.mx, jsalaspi@uoc.edu; P. Missier, Newcastle University, Newcastle upon Tyne, UK; e-mail: paolo.missier@ncl.ac.uk.

1556-4681/2023/12-ART49
https://doi.org/10.1145/3617377

certain attributes of the available data. Who the data are being hidden from and the reason for hiding them conforms to the essential difference between these two disciplines. While privacy seeks to protect the **sensitive attributes** (**SAs**) from being collected (or inferred) by a third-party, fairness strives to prevent the decision mechanism from learning potentially discriminatory biases with respect to these attributes. In the most elementary kind of fairness, a classification algorithm is denied access to the attribute for which discrimination is to be prevented. However, this doesn't guarantee that the learnt decision rule will be free from discrimination, since the bias may not only lie in the removed attributes, but on some other features correlated with them.

Fair ML seeks to correct these biases in order for classifiers to produce non-discriminatory decisions. Chouldechova and Roth [12] present a road-map of work to be done in ML fairness. They present a compendium of known facts about fairness, particularly that it is impossible to simultaneously equalise false positive rates, false negative rates and positive predictive values across **protected attributes** (**PAs**), i.e., most fairness definitions are incompatible with one another [11].

This work is related to different areas of responsible ML, such as fair clustering, fair classification and privacy-preserving data mining, as described in the following paragraphs.

1.1 Fairness Correction

A classifier's predictive fairness may be adjusted by the combination of one or more of the following approaches: *preprocessing* the training data [26, 31], *in-processing* the learning algorithm [1, 8, 51, 53, 54], or *post-processing* a classifier's predictions [28].

Fairness-correcting preprocessing is defined by Friedler et al. [24] as a set of preprocessing techniques that modify input data so that any classifier trained on such data will be fair. Preprocessing presents two distinct advantages over in-processing and post-processing correction:

(1) Preprocessing methods are classifier-agnostic, i.e., they will work regardless of the chosen classifier. In contrast, most in-processing approaches consist of a modification over an existing classifier, e.g., fairness-aware **Logistic Regression** (**LR**) [51] or Naíve Bayes [7], or by adding a fairness regularisation term, a strategy that would only apply to certain classifiers such as LR [33].
(2) The introduced corrections are transparent and auditable by the user: it is possible to quantify and report the changes introduced into data.

According to Kamiran and Calders [32], there are four ways in which to make adequate adjustments to data to enforce fairness: suppressing certain features, also known as *fairness through unawareness* [25], reweighing features [35], re-sampling data instances [26, 31, 43, 47], and *massaging* variable values [32], e.g., through data relabelling. According to Berk et al. [5], one of the main problems with fairness correction is the loss of prediction accuracy caused by such interventions. Our method corrects fairness through massaging keeping a low accuracy loss. Label massaging creates synthetic data and thus, when combined with microaggregation, it provides effective privacy protection for data anonymisation.

1.2 Fairness and Privacy

Privacy and certain fairness definitions, e.g., equalised odds, aim at concealing sensitive information while preserving the utility of data with respect to the task output [41]. While privacy aims at protecting information from disclosure, fairness seeks to prevent certain classification behaviours related to the PA [14]. Ekstrand et al. [20] argue in favour of integrating fairness research to socio-technical systems that provide privacy protection. A k-anonymisation method was used in Hajian et al. [27] to protect frequent patterns with fairness. However, most of the work connecting privacy and fairness, has to do with differential privacy.

Differential privacy may have disparate impact on both performance and fairness. Bagdasaryan et al. [3] show that the reduction in accuracy incurred by deep differential private models disproportionately impacts *underrepresented* PA groups, with differential privacy amplifying the model's bias towards the most popular groups. Regarding fairness, Cummings et al. [13] show that if demographic parity is satisfied by an ε-differentially-private algorithm, this is caused by trivial-accuracy classifiers. They also show that it is impossible to achieve exact fairness with non-trivial accuracy and give approximate fairness bounds. Pujol et al. [40] show that the use of small ε values also has a disparate impact on different PA groups. They propose a modified Laplace mechanism prioritising correct positive classifications. This corrects the disparity, at the cost of more misclassifications.

Foulds et al. [22] define a multi-attribute fairness metric, *differential fairness*, inspired by the use of ε in differential privacy, and extends the notion of the "80% rule" [4]: in order to satisfy ε-differential fairness, the quotient of any two PA subgroups' **positive ratio** (**PR**) should be greater than $e^{-\varepsilon}$ and smaller than e^{ε}. Jagielski et al. [30] introduce a post-processing method, *private equalised odds*, which is differentially private with respect to the PA and achieves equalised odds fairness, at the cost of a high error-rate. Dwork et al. [18] discuss conditions upon which fairness in classification implies privacy, and how differential privacy may be related to fairness. Models of differential private and fair LR are provided in Xu et al. [50].

1.3 Contributions

In this article, we present Fair-MDAV, a fairness correction method for classification tasks with t-closeness and k-anonymity guarantees. It is a modular system allowing for both privacy and fairness to be enhanced either separately or all at once without any compromises, that may be adjusted through protection and correction parameters.

Benchmark experiments were performed over three commonly used benchmarks for algorithmic fairness and privacy: Adult Income, COMPAS and German Credit, measuring several performance metrics as well as three fairness metrics. Fair-MDAV is competitive at enforcing demographic parity, and outperforms state-of-the-art equalised odds methods at fairness/accuracy tradeoff. This article is an extension of the position paper [45]: Fair-MDAV now allows to protect the privacy of an SA, correct the fairness with respect to it, or do both things at once. This version now permits to enforce fairness in two different ways, namely, *positive* and *negative* correction, each performing better than the other with respect to *recall* and *precision*, respectively.

Ntoutsi et al. [39] emphasise on the need to consider the unsupervised case, particularly for unlabelled data. While Fair-MDAV is not an unsupervised learning *per se*, it makes use of the MDAV clustering algorithm [16] and borrows the notion of *fairlet* from the area of fair clustering [10].

2 BACKGROUND AND DEFINITIONS

The scope of this article lies between the two communities of fairness and privacy research. We provide an extensive list of definitions to clarify the similarities and differences between both areas.

2.1 Fairness

The following definitions are useful in understanding *group fairness*, the family of fairness metrics on which we focus.

A binary task's labels may usually take *positive* or *negative* values, referring to how desirable that outcome may be, e.g., whether an application for college admission is successful or not. A dataset's *PA* refers to a feature prone to discrimination, due to many possible factors. In our case we will be dealing with a single binary PA, e.g., every data-point will belong to one of two possible groups. The ratio of the number of positive instances to the total number of instances in a group will be referred to as the PR of the group. The PA value having the highest PR will be referred to as the *favoured* (F) group, while the other one will be the *unfavoured* (U) group.

Although there are many different fairness definitions [34, 36], in this work we'll focus on two of them: *demographic (or statistical) parity*, which aims at achieving an *equality of outcomes* across groups, and *equalised odds*, a stronger condition which seeks to equalise the PRs across correctly-classified individuals.

Definition 2.1 (Demographic Parity). A classifier satisfies *demographic parity* (dPar) if the probability of being classified as positive is the same across PA subgroups:

$$\mathbb{P}(\,\hat{Y} = 1 \mid \mathrm{PA} = U\,) = \mathbb{P}(\,\hat{Y} = 1 \mid \mathrm{PA} = F\,).$$

Definition 2.2 (Equalised Odds). A classifier satisfies *equalised odds* if the probability of being correctly classified is the same across PA subgroups:

$$\mathbb{P}(\,\hat{Y} = i \mid Y = i,\ \mathrm{PA} = U\,) = \mathbb{P}(\,\hat{Y} = i \mid Y = i,\ \mathrm{PA} = F\,) \text{ for } i \in \{0, 1\}.$$

Since these equalities rarely hold, they may be thought of as differences between groups, with smaller differences being more desirable with respect to a particular fairness definition. Therefore, in this article we use the following scores:

Definition 2.3 (Demographic Parity Score). The *demographic parity score* (dPar) for a classifier $\hat{Y}$ is

$$\mathrm{dPar}(\hat{Y}) := \left|\mathbb{P}(\,\hat{Y} = 1 \mid \mathrm{PA} = U\,) - \mathbb{P}(\,\hat{Y} = 1 \mid \mathrm{PA} = F\,)\right|. \tag{1}$$

Definition 2.4 (Equalised Odds Score). The *equalised odds score* (eOdds) for a classifier $\hat{Y}$ is

$$\mathrm{eOdds}(\hat{Y}) := \sum_{i \in \{0,1\}} \left|\mathbb{P}(\,\hat{Y} = i \mid Y = i,\ \mathrm{PA} = U\,) - \mathbb{P}(\,\hat{Y} = i \mid Y = i,\ \mathrm{PA} = F\,)\right|. \tag{2}$$

The *fairlet* concept, borrowed from *fair clustering*, is defined as follows:

Definition 2.5 (Fairlet). Given a dataset D with binary PA taking values F and U, an (m, n)*-fairlet* of D is defined as a subset of D with m instances such that $\mathrm{PA} = U$ and n instances such that $\mathrm{PA} = F$.

In the clustering context [2, 10], fairlets are used for obtaining *fair clusters*, i.e., clusters in which the PA distribution is similar to the whole dataset. In our case, the purpose of fairlets is to locally correct fairness and to anonymise the data.

2.2 Privacy

The following definitions are basic definitions for privacy-preserving data publishing that are used throughout this article.

Removing all the *identifiers* from a database, such as social security number or name-surname, does not protect individuals from re-identification. They may still be linked to unique combinations of attribute values, and their *SAs* revealed in the data mining process. Two of the main models for privacy-preserving data mining are k-anonymity and its enhancements, e.g., t-closeness, and differential privacy. The differences and interactions between these two models are discussed in the context of big data in Salas and Domingo-Ferrer [44].

The confidential or SAs are attributes that contain sensitive information on the individual, e.g., salary, medical conditions, religious beliefs, and so on. The ***quasi-identifiers*** (**QIs**) are attributes, such that a unique combination of its values may be used to single out an individual for re-identification. In our setting, the data features may be separated in SAs and QIs, with the SAs consisting of the PA and the label. In [49], Latanya Sweeney showed that by using gender, birth dates, and postal codes as QIs it is possible to identify 87% of individuals in the United States.

To prevent re-identification, k-anonymity is defined in Samarati [48], Sweeney [49] as follows.

Definition 2.6 (k-Anonymity). A dataset is *k-anonymous* if each record is indistinguishable from at least other $k - 1$ records within the dataset, when considering the values of its QIs.

There are several techniques for obtaining k-anonymous datasets, e.g., microaggregation, generalisation, or suppression; Fair-MDAV is based on microaggregation.

Definition 2.7 (Microaggregation). Microaggregation is a family of masking methods for statistical disclosure control, that obtain microaggregates in a dataset with n records, by combining the records to form groups of size at least k. For each attribute, an aggregate value (usually the average) of each group is computed and is used to replace each of the original values.

Definition 2.8 (k-group). We define a *k-group* as a group of k elements used for microaggregation.

Definition 2.9 (Information Loss). We define the information loss of a microaggregation procedure over dataset D as the square root of the average within-group sum-of-squared-errors, i.e.,

$$\text{Information Loss} = \sqrt{\frac{1}{|D|}\sum_{j=1}^{n_g}\sum_{i=1}^{n_j} d(x_{ij}, \bar{x}_j)^2}, \tag{3}$$

where d is the Euclidean distance, x_{ij} denotes record i of group j, $\bar{x}_j$ is the average record of group j, n_g is the number of groups in the dataset and n_j is the number of elements in group j.

The optimal k-partition for a microaggregation procedure is the one that maximises within-group homogeneity, that is, the one that minimises the information loss. By design, k-anonymity guarantees that the probability of linkage of an individual's record is lower than $1/k$, i.e., an adversary knowing some of their characteristics cannot distinguish their record among a group of k similar records. However, if unmodified, the distribution of the SAs in a k-group may be used to infer the SA value of an individual, even if re-identification is not possible.

For example, if all the SA-values in a k-group are the same, an adversary may use their knowledge of a record to link it with a k-group and learn the SA of the record. To prevent attribute disclosure from to k-anonymity, t-closeness [38] is defined as follows:

Definition 2.10 (t-Closeness). A k-anonymous dataset D satisfies *t-closeness* if all its k-groups satisfy t-closeness. A group of records of a dataset satisfies t-closeness if the distance between the distribution of the SAs of the individuals in the group to the distribution of the SAs in the whole table is not greater than a threshold t.

The rationale behind t-closeness is that by making the distribution of the SAs in the k-groups similar to the distribution on the entire dataset, an adversary will not improve his knowledge of the SAs of a record, even when knowing the k-group to which the record belongs.

In our case, we consider that the proportions of favoured and unfavoured attributes (SAs) in the k-groups are similar to its proportions in the entire dataset. This is consistent with the concept of fairlet (Definition 2.5).

Differential privacy may not be used for fairness-correcting preprocessing. The most widely used approach to privacy protection is differential privacy [19]. Its promise is that including or excluding an individual's record in the database does not significantly affect the output of a learning algorithm, and thus the same inferences about an individual may be made regardless of their record belonging to the dataset. Such strong privacy guarantees come with a large cost on precision, as there are inevitable tradeoffs between privacy and utility.

ALGORITHM 1: Fair-MDAV algorithm.

```
input : D: dataset to process, with binary PA-and-label,
        m: number of unfavoured records per fairlet,
        n: number of favoured records per fairlet,
        ma: boolean whether entries are microaggregated or not,
        tau: float, the level of fairness correction,
        nc: boolean whether negative correction is performed.
G ← MakeFairlets(D, m, n);                               // Algorithm 2
G_micro ← G;                       // No microaggregation if ma is False
if ma is True then
    G_micro ← Microaggregate(G);                        // Algorithm 3
D_corrected ← CorrectFairness(D, G_micro, tau, nc);     // Algorithm 4
return D_corrected
```

More importantly, differential privacy is typically used in an interactive setting in which the result of a query is returned with noise added to provide it with differential privacy. Therefore, differential privacy may not be used for fairness-correcting preprocessing.

3 FAIR-MDAV DESCRIPTION

Fair-MDAV, presented in Algorithm 1, consists of three methods:

(1) Given a desired cluster size k, `MakeFairlets` (Algorithm 2) clusters the training set D into G, a collection of (m, n)-fairlets such that the fairlet's elements are close to each other, where m/n approximates the $|U| / |F|$ proportion of D as closely as possible subject to $m + n = k$. Clusters are formed in the following way: Let $\bar{r}$ be D's "average record", i.e.,

$$\bar{r} := \frac{1}{|D|} \sum_{r \in D} r.$$

`MakeFairlets` first locates e^*, the furthest element in D from $\bar{r}$, i.e.,

$$e^* = \arg\max_{e \in D} d(e, \bar{r}),$$

where d is the Euclidean distance.
Depending on $\mathrm{PA}(e^*)$, we define the auxiliary sets F_{e^*}, U_{e^*} as follows: if $e^* \in U$, then F_{e^*} are the m nearest *favoured* records to e^*, while $U_{e^*} \subset U$ are the $n - 1$ nearest *unfavoured* records to e^*. Otherwise, if $e^* \in F$, then take $m - 1$ records for F_{e^*} and n records for U_{e^*}. Then, $g := \{e^*\} \cup F_{e^*} \cup U_{e^*}$ will be an (m, n)-fairlet where F_{e^*} and U_{e^*} are the nearest *favoured* and *unfavoured* neighbours of e^*, respectively. Fairlet g is appended to G (D's fairlet partition), the elements of g are removed from D and `MakeFairlets` iterates over the remaining records in D until no more fairlets can be formed, discarding the remaining records.

(2) `Microaggregate` (Algorithm 3) replaces the original records' feature values with the corresponding aggregated feature values of the fairlet they belong to, e.g., with the mean values. The only exceptions to this are the PA and the label, for which the original values are kept.

(3) `CorrectFairness` (Algorithm 4) locally corrects the fairness of each fairlet by relabelling its records depending on their PA values so that $\mathrm{PR}(U) \geq \tau \cdot \mathrm{PR}(F)$, where τ modulates the amount of correction introduced to the data. Fairness correction may occur by relabelling negative *unfavoured* instances as positive (*positive correction*), or by relabelling positive *favoured* instances as negative (*negative correction*).

ALGORITHM 2: MakeFairlets method

```
input: D, m, n
G ← {};
while $|\{x \in D \mid pa(x) = 0\}| \geq m$ and $|\{x \in D \mid pa(x) = 1\}| \geq n$ do
    x_mean ← Mean({x ∈ D});
    x_r ← $\arg\max_{x \in D}$(distance(x, x_mean));
    // Get m, n nearest records to x_r with pa = 0, pa = 1. Depending on its PA, x_r will belong
       to either g_u or g_f.
    g_u ← GetNearestU(x_r, m);
    g_f ← GetNearestF(x_r, n);
    g ← Append(g_u, g_f);
    D ← Drop(D, g);
    G ← Append(G, g);
return G
```

ALGORITHM 3: Microaggregate method

```
input: G
// Replace all the records in a fairlet by their mean value
G_micro ← {};
for g in G do
    g_mean ← {};
    x_mean ← Mean({x ∈ g});
    foreach x in g do
        Append(g_mean, x_mean);
    Append(G_micro, g_mean);
return G_micro
```

ALGORITHM 4: CorrectFairness method

```
input: G
/* Calculate the positive ratio for the favoured and unfavoured groups, replacing as many
   unfavoured negatives with favoured positives as needed, dependant on tau          */
for g in G do
    fpr ← PositiveRatio(g, 1);
    upr ← PositiveRatio(g, 0);
    U_0 ← {x in g | pa(x) = 0 and y(x) = 0};
    F_1 ← {x in g | pa(x) = 1 and y(x) = 1};
    if nc is False then
        while upr < tau ∗ fpr and |U_0| > 0 do
            y(x) = 1 for x in U_0
    else
        while upr < tau ∗ fpr and |F_1| > 0 do
            y(x) = 0 for x in F_1
D_corrected ← Concatenate({g in G});
return D_corrected
```

In summary, Algorithm 1 consists of three components: Algorithm 2 calculates the training-set fairlets, Algorithm 3 provides privacy protection, and Algorithm 4 introduces fairness correction.

Example Data

id	X	PA	label
A	1	1	1
B	2	0	0
C	3	0	1
D	11	1	0
E	12	1	0
F	13	1	1
G	14	1	1

$(2, 1)$-Fairlets

id	X	X_{ma}	PA	label	PC	NC
A	1	4.67	1	1	1	0
B	2	4.67	0	0	1	0
C	3	9.33	0	1	1	1
D	11	4.67	1	0	0	0
E	12	9.33	1	0	0	0
F	13	9.33	1	1	1	1
G			*Dropped*			

Fig. 1. Microaggregates generated by Fair-MDAV, coloured in orange and purple, with microaggregated features (X_{ma}) and labels corrected positively (PC) and negatively (NC).

3.1 Simplified Example

Figure 1 exemplifies Fair-MDAV being applied to a 7-entry minimal dataset consisting of one continuous feature (X) and binary PA and label. In this particular example, Fair-MDAV microaggregates the data into $(2, 1)$-fairlets in the following way:

(1) The average value of X is 8, and the record with the farthest X-value from 8 is A.
(2) Since A has PA $= 1$, to complete a $(2, 1)$-fairlet two records must be added: one with PA $= 0$ and one with PA $= 1$. The valid records closest to A are B and D. Hence, the first fairlet will be $\{$A, B, D$\}$, coloured orange in Figure 1.
(3) After removing these three records, the process is repeated. The second fairlet becomes $\{$C, E, F$\}$, coloured purple.
(4) Since three records are required to build a $(2, 1)$-fairlet and only one element remains, it is dropped (G).
(5) Once the fairlets are grouped, the PA may be positively (PC) or negatively (NC) corrected. For the orange fairlet, the uncorrected PRs are 0 for its *unfavoured* elements and 0.5 for its *favoured* elements. When applying PC, record B's label changes from 0 to 1, increasing the unfavoured records' PR from 0 to 1. Conversely, when applying NC, records A and C are relabelled from 1 to 0, which changes the favoured records' PR from 1 to 0. On the other hand, the purple fairlet needs no correction, as the unfavoured PR is already greater than the favoured PR.

Assuming $\tau = 1$, the PC and NC columns display the positive and negative fairness correction relabellings, respectively, which are performed fairlet-wise. For PC, entries with PA $= 0$ and label $= 0$ get relabelled as 1, as long as the proportion of PA $= 1$ with label $= 1$ is higher than the proportion of PA $= 0$ with label $= 1$. For NC, it is PA $= 1$ with label $= 1$ entries that get relabelled to 0 as long as the same condition holds.

4 EXPERIMENTS

Our experiments were run over three datasets commonly used as benchmarks in both the privacy and fairness literature: *Adult Income (Income)* [17], *COMPAS* [37], and *German Credit (Credit)* [17], described next and in Table 1.

Income is a subset of the 1994 U.S. census, where the usual task assigned to it is to predict whether individuals will earn over \$50, 000 per year given their demographics; it is biased against women. *COMPAS* (Correctional Offender Management Profiling for Alternative Sanctions) is a subset of criminal records from Broward County, Florida, regarding the recidivism of former

Table 1. Datasets Used for Our Experiments

Dataset	SAs		QIs	Instances
	PA	Label		
Income	Sex	Earns Over $50 k	12	48,842
COMPAS	Race	Recidivated	8	6,907
Credit	Sex	Repaid Loan	20	1,000

Table 2. Experimental Parameter Values

Parameter	Values	Description
m	$\lfloor 10i \cdot \frac{\lvert U \rvert}{\lvert D \rvert} + 0.5 \rfloor$ for $i \in 1 \dots 10$	(m, n)-fairlet size $[n = 10i - m]$
k	$10i$ for $i \in 1 \dots 10$	k-group size
τ	$0.1i$ for $i \in 0 \dots 10$	Fairness correction level
ma	`True, False`	Microaggregation
nc	`True, False`	Negative correction

felons in the two years following their release; it is shown to be biased against black felons. *Credit* contains financial and demographic attributes along with a classification of each individual as a good or bad credit risk, i.e., whether they repaid their loans or not. It is also biased against women.

Given the difference in number of instances between these datasets, the patterns are clearer on *Income*, while a lot of noise is present on our *Credit* analyses. We evaluated a selection of fairness and performance metrics exhaustively across the parameter values presented in Table 2.

For each parameter combination, five-fold **cross validation** (**CV**) train-test splits were performed, applying Fair-MDAV to the training sets and followed by training several distinct classifiers—LR, **random forests** (**RF**), **stochastic gradient descent** (**SGD**) with hinge loss and **Gaussian naíve bayes** (**GNB**)—over modified training sets. Fairness and performance metrics were evaluated over the corresponding test sets, and the resulting metrics are represented with their average across CV splits, and the 95% confidence intervals. We use *scikit-learn*'s [15] implementations to learn the corresponding classifiers.

We measure the quality of the clustering by microaggregation through information loss (3), averaged by attribute, considering that each dataset has a different number of attributes for comparison. We compare our results on information loss to MDAV [16], a microaggregation-based algorithm (without fairness guarantees) that has had several modifications and improvements, such as adapting it for dynamic data [46] or improving its efficiency [42].

Figure 2 shows that MDAV incurs lower information loss than Fair-MDAV since the fairlets are more restrictive than the k-groups, and the group size ($m+n$ or k) is related to the information loss, however the difference between both methods almost remains constant with fairlet size increments. In the following section we show how our method is able to correct for group fairness while providing good utility and privacy.

4.1 Fairness, Privacy, and Accuracy Tradeoffs

The modularity of Fair-MDAV makes its privacy and fairness-correction components independent of each other. It is nonetheless interesting to analyse the impact of choosing certain parameter values on the effectiveness of the correction. Figure 3 shows that Fair-MDAV is effective regardless of the classifier that we tested on the data.

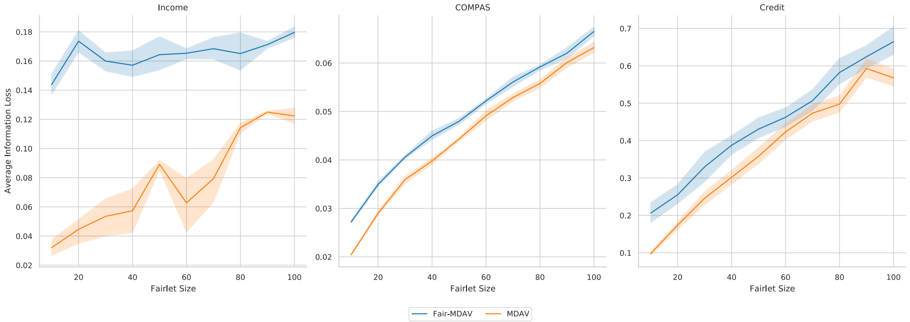

Fig. 2. Information loss for MDAV and FAIR-MDAV related to the group size on *Income*, *COMPAS,* and *Credit*.

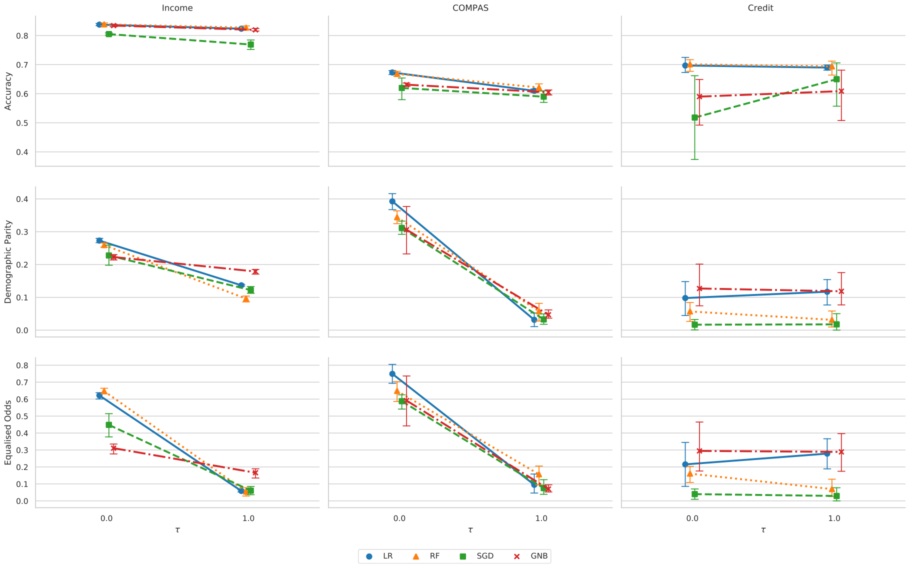

Fig. 3. Fairness/accuracy tradeoffs obtained considering $\tau = 0$ (no correction) and $\tau = 1$ (full correction) with a fixed fairlet size of 10 on *Income*, *COMPAS,* and *Credit*.

It also shows that the accuracy-loss caused by τ is negligible, while the disparity is steeply reduced towards zero (as measured with dPar and eOdds). FAIR-MDAV produced similar results across the aforementioned classifiers. Therefore, we focus the rest of our analysis on LR alone, which additionally allows us to compare FAIR-MDAV with Xu et al. [50]'s differentially-private-and-fair LR.

In Figure 4 we show the tradeoffs between fairness and accuracy, averaged across the CV folds on the test set. It shows all the possible tradeoffs on fairness and accuracy across all parameter value combinations of fairlet size $(n + m)$ and correction (τ) presented in Table 2 for the *Income* dataset. Analogous plots for *COMPAS* and *Credit*—displaying similar behaviours—are presented in Figures 11, 12, 13, 14 15, 16, 17, and 18 in Appendix B.

Figure 4 showcases the effect of τ correction and privacy guarantees measured by fairlet sizes. The sparsity in accuracy-loss is mostly explained by micro-aggregation, with cluster-size being the dominant factor that determines it: the larger the fairlet, the more accuracy will be lost with respect to the original dataset. Fairness correction, on the other hand, has a slight impact on accuracy loss,

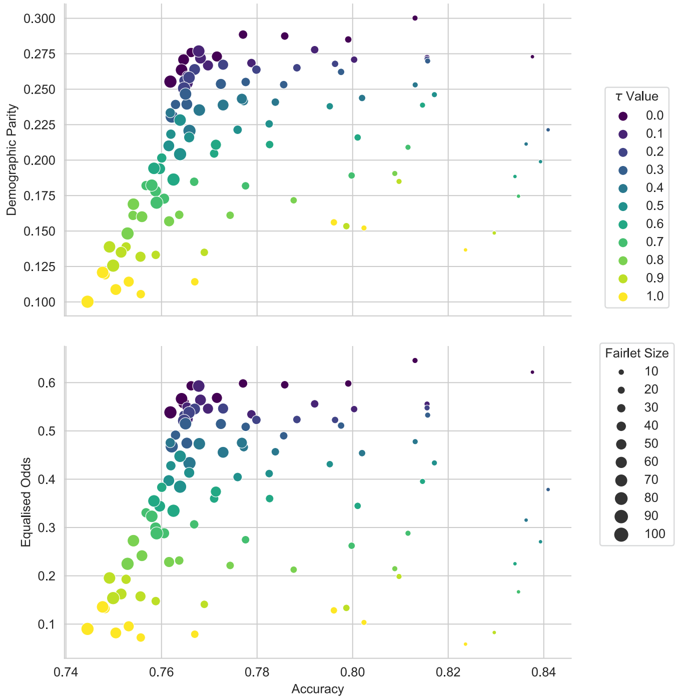

Fig. 4. Fairness/accuracy tradeoff on microaggregated *Income* across Fair-MDAV parameters.

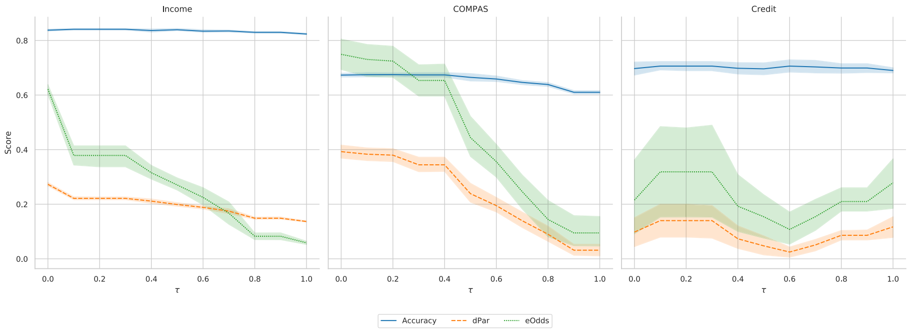

Fig. 5. Fairness/accuracy tradeoffs obtained for correction parameter τ and fixed fairlet size of 10 on *Income*, *COMPAS,* and *Credit*. The lines represent the average for the 5-folds and 95% confidence interval.

while having a large, descending impact on equalised odds and demographic parity scores. In Figure 5, we show such tradeoffs (with their confidence intervals) for a fixed fairlet size of 10 as a function of τ on *Income*, *COMPAS,* and *Credit*. It can be observed that the best scores for both dPar and eOdds are achieved for τ values closer to 1.

In Figure 6, we can see that the negative correction parameter of Fair-MDAV always improves fairness, however, depending on the correction it may improve classifiers' precision or recall. We consider that negative correction is harder to ethically justify, as it implies taking away from the favoured group instead of compensating the unfavoured group, i.e., fairness is attained with a lessened amount of overall well-being.

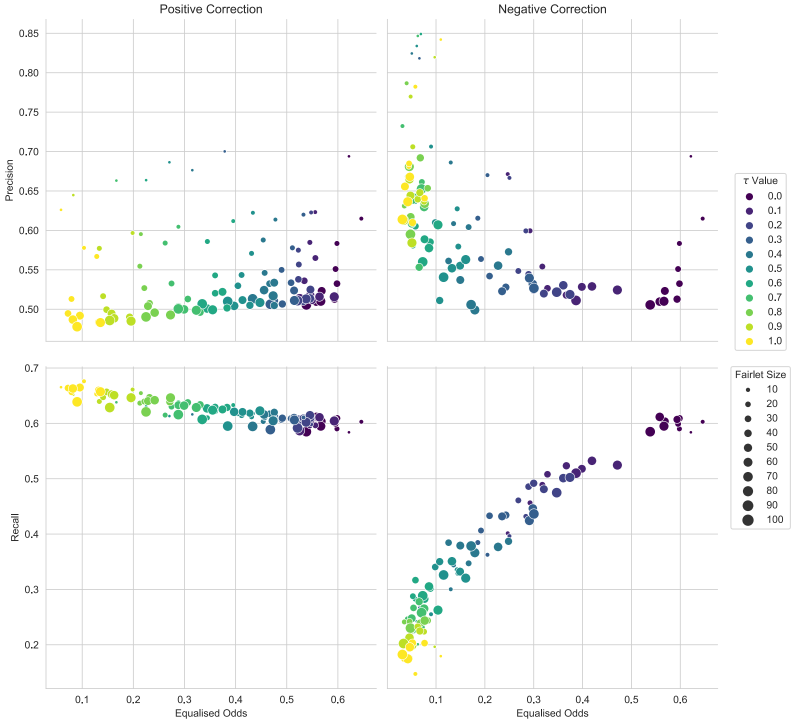

Fig. 6. Precision and recall vs. equalised odds for positive and negative correction experiments over microaggregated *Income*.

4.2 Using Fair-MDAV for Fairness Correction Alone

The privacy and fairness components of Fair-MDAV may be independently applied. In this subsection, we consider the properties of Fair-MDAV as a fairness correcting algorithm, without the additional privacy guarantees provided by microaggregation.

As may be observed in Figure 7, when data is not microaggregated fairness correction and accuracy loss are strongly correlated, i.e., higher τ values not only lead to improved outcomes for the unfavoured group, but also cause the performance of the learnt classifiers to drop. The "fairest" non-microaggregated classifiers produced an accuracy higher than 84%, with much less correction being required for it than in the *microaggregated* case. Depending on the correction and cluster types, τ values between 0.6 and 0.8 produced optimal eOdds, while $\tau = 1$ is required to attain optimal demographic parity. Larger τ values lead to a fairness over-correction, i.e., the unfavoured group's PRs become much better than the favoured group's, and hence unfairness happens, only in a reversed way, i.e., the favoured group becomes unfavoured and viceversa. This behaviour contrasts with the microaggregated case, in which setting $\tau = 1$ achieves the best possible fairness for every cluster size and type.

The need for larger τ values to enforce demographic parity may be explained by the severity of the introduced constraint: while for eOdds true positives and true negatives usually sit nearby a decision boundary, forcing positive rates across PA groups may require a much greater change in the aforementioned decision boundary.

In Figure 8, we show the tradeoffs between accuracy and fairness obtained for a fixed fairlet size of 10 (without applying microaggregation) and positive correction parameter τ on *Income*, *COMPAS* and *Credit*. It can be seen that the minimum score for demographic parity is achieved for

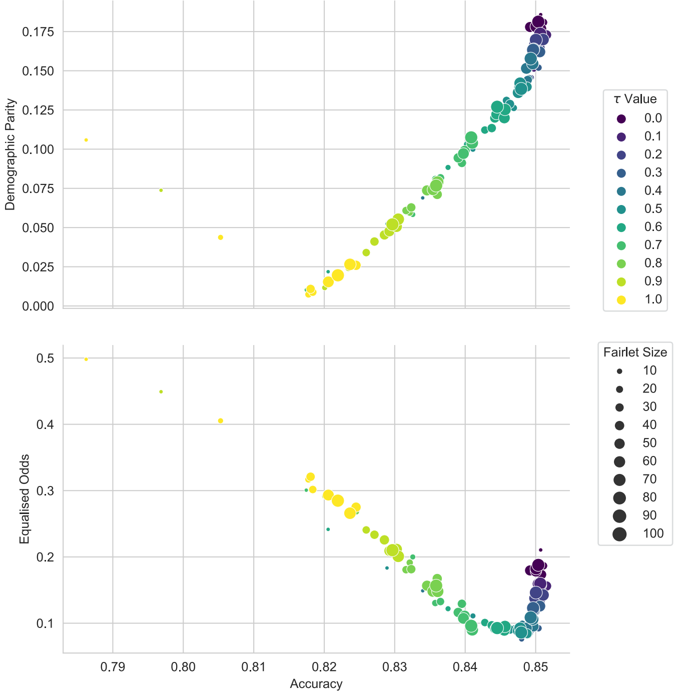

Fig. 7. Fairness/accuracy tradeoff on non-microaggregated *Income* across Fair-MDAV parameters.

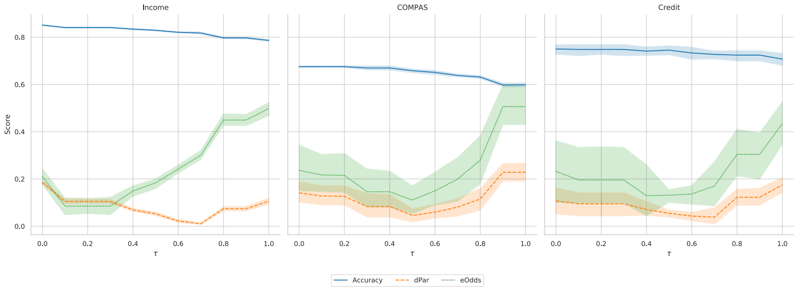

Fig. 8. Average and 95% confidence interval for accuracy and fairness obtained without microaggregation for each positive correction parameter τ and fixed fairlet size of 10 on *Income*, *COMPAS*, and *Credit*.

the values of τ between 0.5 and 0.7 for all three datasets. This value may be automatically obtained for any specific dataset through an optimisation procedure, e.g., Bayesian optimisation [23]. Note that increasing the τ after achieving the minimum fairness score actually decreases the dataset's fairness by over-compensating the unfavoured group.

Figure 9 shows the resulting precision and recall across the non-microaggregated experiments on *Income*. As may be seen, fairness correction impacts precision and recall in opposite directions, and it does so differently for positive and negative correction: while for positive correction higher τ-values cause worse precision and better recall scores, for negative correction higher τ-values imply better precision but worse recall.

The precision/recall tradeoff is a well-known phenomenon [6], but in our case it follows from the way Fair-MDAV corrects fairness: positive correction relabels negative records as positive,

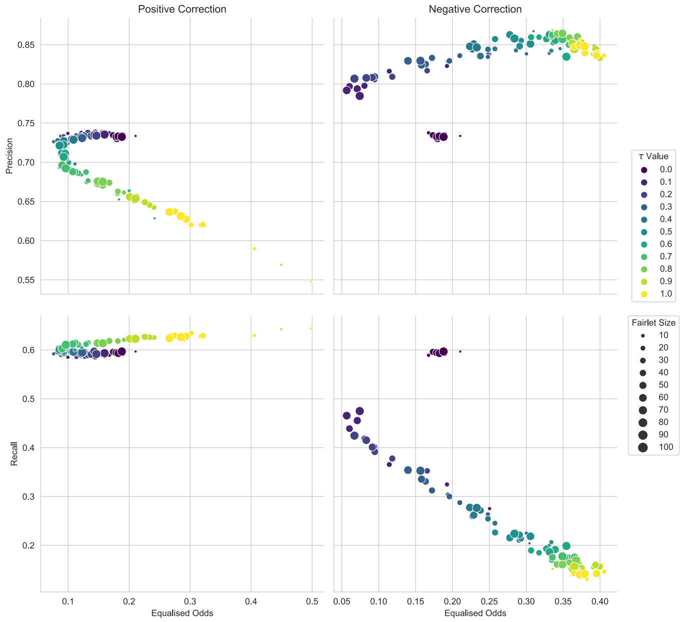

Fig. 9. Precision and recall vs. equalised odds for positive and negative correction experiments over non-microaggregated *Income*.

relaxing the constraints upon which a data instance would be classified as positive. This, in turn, causes the learnt classifier to predict less false negatives (improving recall) at the cost of labelling more false positives (worsening precision). With negative correction the effect is precisely the opposite: the constraints for being classified as positive are stronger, hence there will be less false positives (better precision) but more false negatives (worse recall). Since classification tasks may find one performance metric to be more relevant than the other (e.g., for fraud detection the false negative rate, i.e., recall, should be as small as possible), the choice of positive vs. negative correction can aid in improving the relevant performance metric, as a bonus to correcting fairness.

4.3 Comparison vs. Existing Fairness Methods

In this subsection, we compare Fair-MDAV with other existing fairness methods that are optimised for improving either Demographic Parity or Equalised Odds. We perform a comparison of the Pareto fronts of Fair-MDAV, *ParDS* [26], *Fair Classification* [52], and *FairLearn Threshold Optimiser* [28] on all three datasets in our study: *Income*, *COMPAS,* and *Credit*. Additionally, we compare Fair-MDAV with the reported values on *Income* for the following algorithms: PFLR* [50], ***preferential sampling (PREF)*** [32], ***learning fair representations (LFR)*** [54], ***adaptive sensitive reweighting (ASR)*** [35], *ADAFAIR* [29], *SMOTEBOOST* [9], and ***parametrised correction (ParDS)*** [26].

4.3.1 Demographic Parity Comparison. We perform a comparison of the Pareto fronts of preprocessing (*ParDS*), in-processing (*Fair Classification*), and post-processing (*FairLearn*) state-of-the-art methods.

Since *FairLearn Threshold Optimiser* (like many other in-processing methods) returns a classifier with no adjustable parameters, it can not be tuned for different fairness/accuracy tradeoffs.

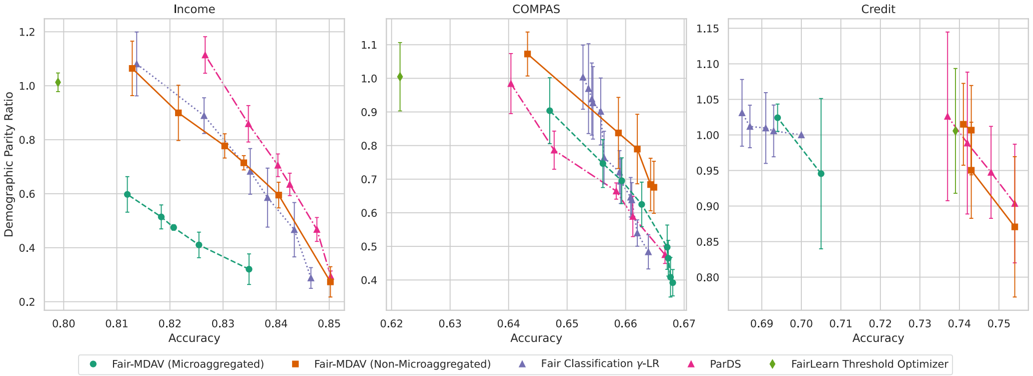

Fig. 10. Fairness/accuracy tradeoffs for Fair-MDAV compared to pre-, in-, and post-processing methods over the benchmark datasets.

Table 3. Fairness and Accuracy Values for Different Privacy Guarantees

Algorithm	Privacy	dPar	Accu
Fair-MDAV	$t = 0.05$ $(k = 10)$	0.02	**0.79**
Fair-MDAV	$t = 0.19$ $(k = 20)$	0.04	**0.79**
Fair-MDAV	$k = 100$	0.05	0.78
PFLR*	$\varepsilon = 0.1$	**0.00**	0.75
PFLR*	$\varepsilon = 1$	0.01	0.76
PFLR*	$\varepsilon = 10$	0.02	0.76

Fair-MDAV, *Fair Classification*, and *ParDS*, on the other hand, provide different tradeoffs and are classifier-agnostic.

The results presented in Figure 10 were produced by measuring Demographic Parity Ratio, as this is the metric for which *Fair Classification* optimises. 5-Fold CV was performed over the same folds for every method; this accounts for the confidence intervals displayed in Figure 10. For every instance, Fair-MDAV was set to produce fairlets of size 10, i.e., $u + f = 10$. On all five tested methods, LR was chosen as the classification algorithm, since *Fair Classification* is based on it.

We observe that Fair-MDAV outperforms *FairLearn*, i.e., a better fairness ratio is obtained at similar accuracy, and is competitive with *Fair Classification* when not microaggregated, i.e., without adding privacy, and it is outperformed by *ParDS* only over *Income*. We also note that the microaggregated Fair-MDAV is dominated by the non-microaggregated Fair-MDAV on every tested dataset. These results are expected, as there is always a privacy/accuracy tradeoff. It is also worth mentioning that the performance ordering for the methods varies across different datasets, showing that there is no universal-best method.

Table 3 presents demographic parity and accuracy metrics for different privacy definitions (namely, k-anonymity, t-closeness, and ε-differential privacy). Although Fair-MDAV and PFLR* aim at different privacy guarantees, we include this comparison as both methods lie within the Fairness/Privacy intersection. LR classifiers learnt from Fair-MDAV-protected data are competitive in both fairness and accuracy with PFLR*. However, PFLR* provides differential privacy for LR only, and it is optimised for improving Demographic Parity exclusively, while Fair-MDAV can be used with any classifier and for any fairness metric. When guaranteeing t-closeness ($t = 0.05$ $k = 10$),

Table 4. Demographic Parity and Accuracy Comparison of Fair-MDAV with Related Fairness-correcting Methods over *Income*

Method	Clf	dPar	Accu
Fair-MDAV	LR	**0.01**	**0.80**
LFR [54]	LR	0.20	0.68
Fair-MDAV	DT	0.09	0.76
PREF [32]	DT	0.03	**0.84**

The compared metrics were obtained by training different classification algorithms (Clf), namely, LR and **decision trees** (**DT**). The best result for each metric is highlighted.

Table 5. Equalised Odds and Accuracy Comparison of Fair-MDAV with Related Fairness-correcting Methods over *Income*

Method	Clf	eOdds	Accu
Fair-MDAV	LR	0.05	**0.85**
ASR [35]	LR	0.05	0.82
Fair-MDAV	AB	0.22	**0.85**
ADAFAIR [29]	AB	**0.08**	0.83
SMOTEBOOST [9]	AB	0.47	0.81

The compared metrics were obtained by training different classification algorithms (Clf), namely, LR and **AdaBoost** (**AB**). The best result for each metric is highlighted.

Fair-MDAV has better accuracy for the same dPar as PFLR* [50] with $\varepsilon = 10$. PFLR* with $\varepsilon = 0.1$ obtains the overall best dPar at the cost of higher accuracy loss.

Table 4 compares dPar and accuracy (Accu) scores on *Income* for Fair-MDAV when used purely for fairness correction against existing fairness correction methods. Fair-MDAV achieves slightly worse dPar scores than *ParDS*, but they are better than PREF. However, it does so at a higher cost in accuracy loss.

4.3.2 Equalised Odds Comparison. In Table 5, we observe that Fair-MDAV and ASR achieve the best eOdds score; however, Fair-MDAV provides a much better accuracy than ASR.

5 LIMITATIONS

There are two kinds of limitations in this study: technical specifications and legal regulations. From the technical standpoint, our study focused on group-fairness definitions, considered the binary PA-and-label case and using the euclidean distance exclusively during Fair-MDAV's clustering phase. These limitations could be addressed in future research by extending Fair-MDAV to work on multiple and multi-class PA and label, experimenting with different distance definitions and analysing the impact of Fair-MDAV on individual-fairness definitions.

On the legal front, there is a potential issue as well: label massaging alone may be in conflict with the accuracy principle of legislations such as the European Union's ***General Data Protection Regulation*** (**GDPR**) [21]. Hence, the usage of Fair-MDAV for pure fairness-correction (not making use of the privacy component) may not be suitable for some tasks. Nonetheless, anonymised data does not fall within the scope of GDPR—as it can not be associated to specific individuals—and, therefore, is no longer considered to be personal data. Therefore, making use of Fair-MDAV's privacy component overcomes such limitations.

6 CONCLUSION

Fair-MDAV is a modular and parametrisable fairness-correcting preprocessing method with privacy guarantees. Fair-MDAV is also definition agnostic for group-fairness definitions and classification algorithms. Its modularity allows for publishing data with k-anonymity and t-closeness guarantees and then optimising for fair-classification. Precision and recall may be further prioritised through the selection of negative or positive correction, respectively.

We tested Fair-MDAV's effectiveness over three datasets (*Income, COMPAS,* and *Credit*) using four classifiers (*LR, Random Forest, Support Vector Machine,* and *Gaussian Naíve Bayes*), as measured by three performance (accuracy, precision, and recall) and two fairness (demographic parity and equalised odds) scores.

Fair-MDAV outperforms existing fairness-correcting methods, producing a better equalised odds/accuracy tradeoff, and is competitive with respect to the resulting demographic parity/accuracy tradeoff as well. It also provides better accuracy for the same demographic parity scores achieved by PFLR*.

APPENDICES

A ONLINE RESOURCES

A.1 Code Availability

Fair-MDAV is available in the form of a *Jupyter* notebook at https://github.com/vladoxNCL/fairMDAVplus.

A.2 Availability of Data and Material

The datasets on which our experiments were run are available at:

Adult Income: https://archive.ics.uci.edu/ml/datasets/adult
COMPAS: https://github.com/propublica/compas-analysis
German Credit: https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

They were all prepared using the `data_cleanup.ipynb` notebook available at the project's repository.

B ADDITIONAL PLOTS

This section contains analogous plots to Figures 4, 7, 6, and 9 corresponding to the *COMPAS* and *Credit* datasets.

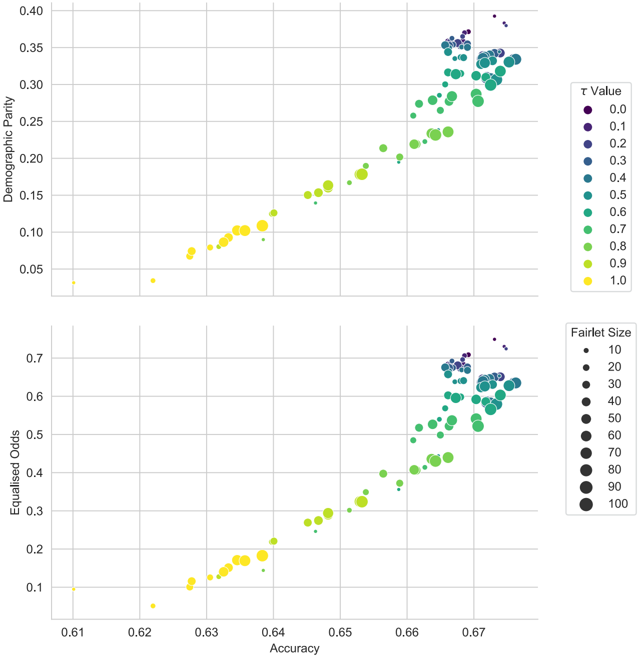

Fig. 11. Fairness/accuracy tradeoff on microaggregated *COMPAS* across Fair-MDAV parameters.

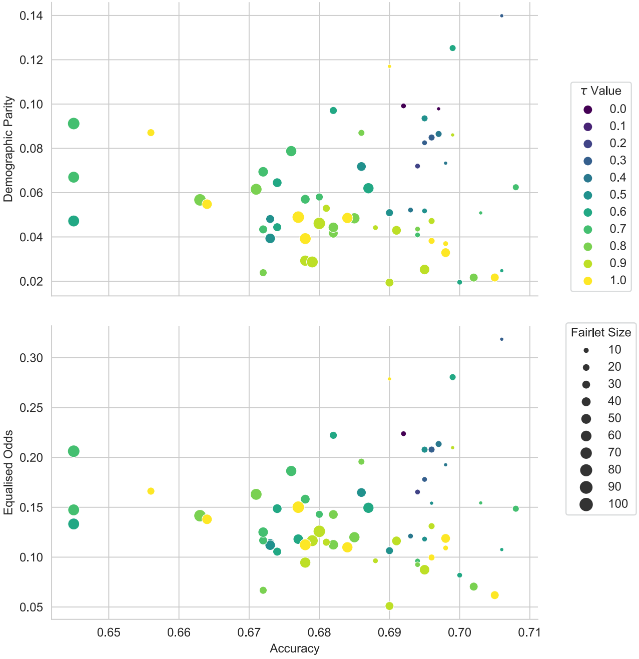

Fig. 12. Fairness/accuracy tradeoff on microaggregated *Credit* across Fair-MDAV parameters.

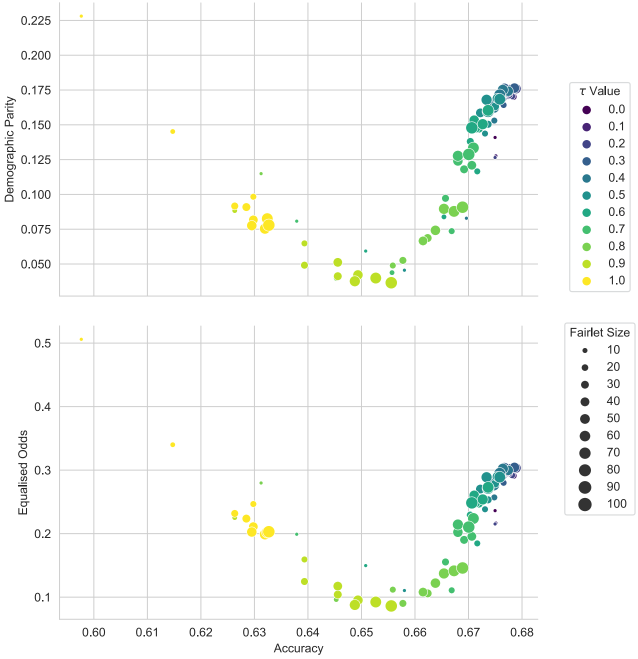

Fig. 13. Fairness/accuracy tradeoff on non-microaggregated *COMPAS* across Fair-MDAV parameters.

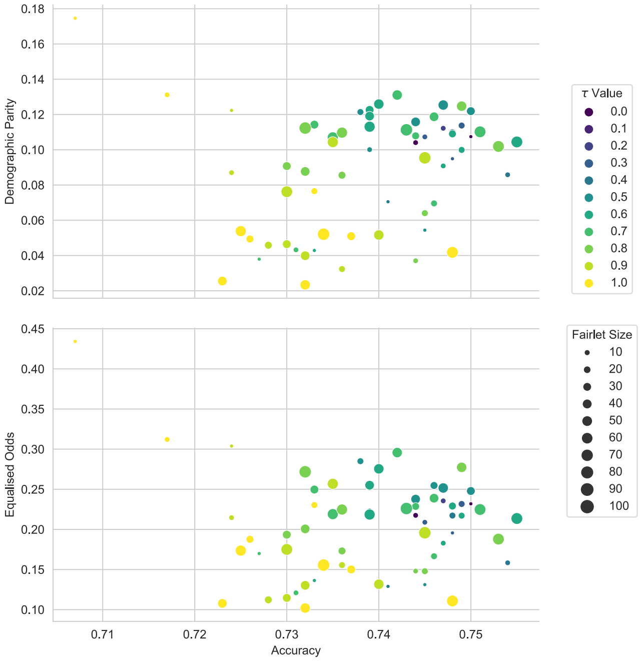

Fig. 14. Fairness/accuracy tradeoff on non-microaggregated *Credit* across Fair-MDAV parameters.

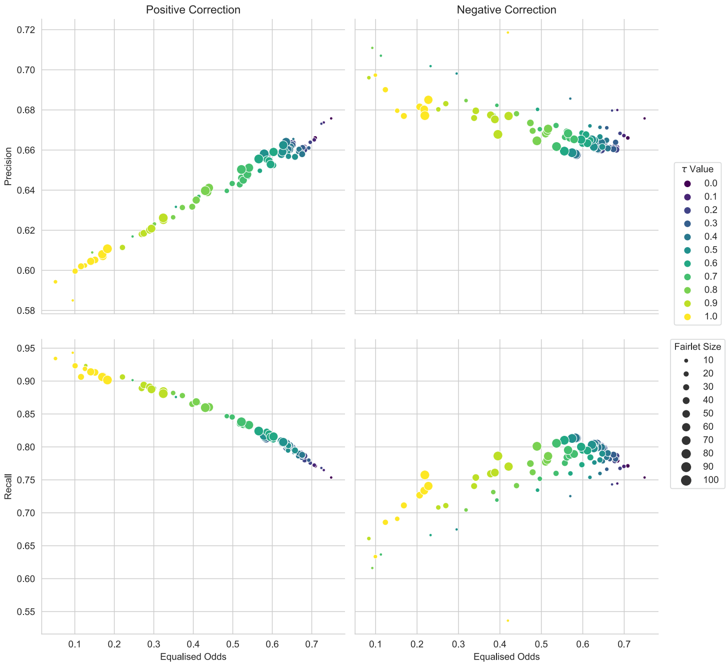

Fig. 15. Precision and recall vs. equalised odds for positive and negative correction experiments on microaggregated *COMPAS*.

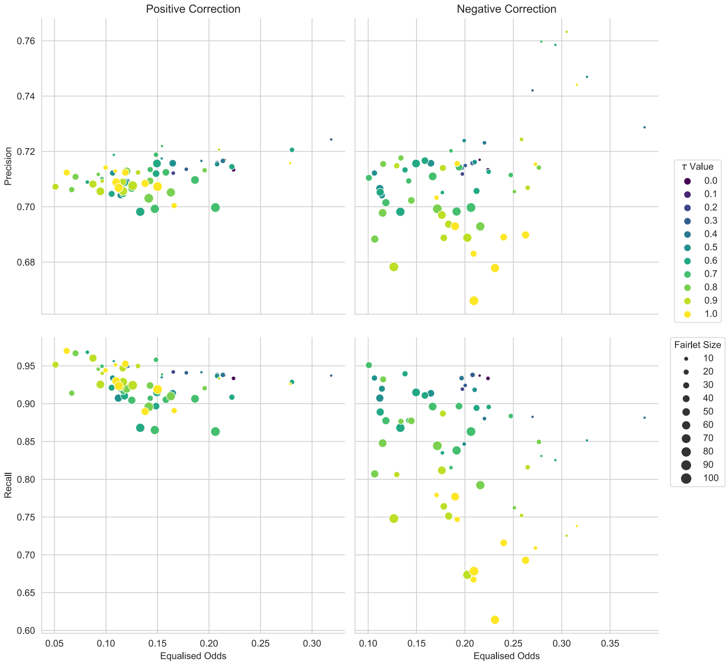

Fig. 16. Precision and recall vs. equalised odds for positive and negative correction experiments on microaggregated *Credit*.

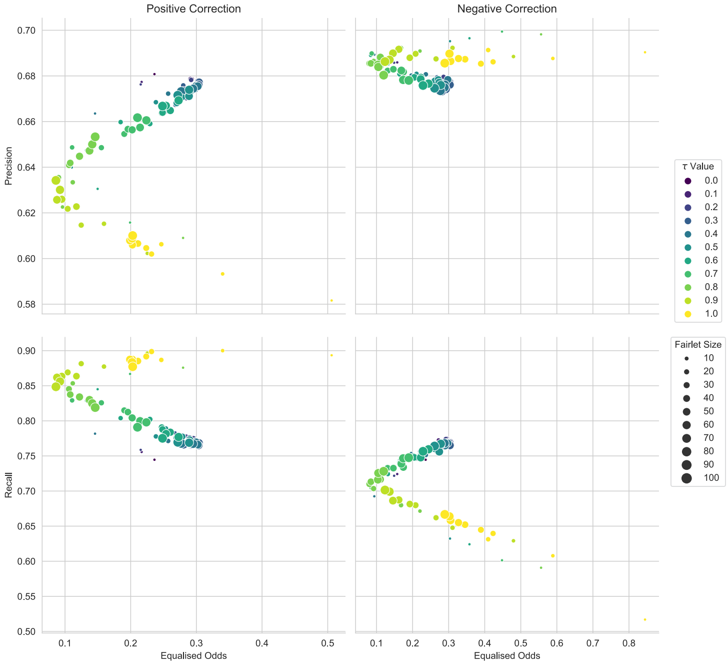

Fig. 17. Precision and recall vs. equalised odds for positive and negative correction experiments on non-microaggregated *COMPAS*.

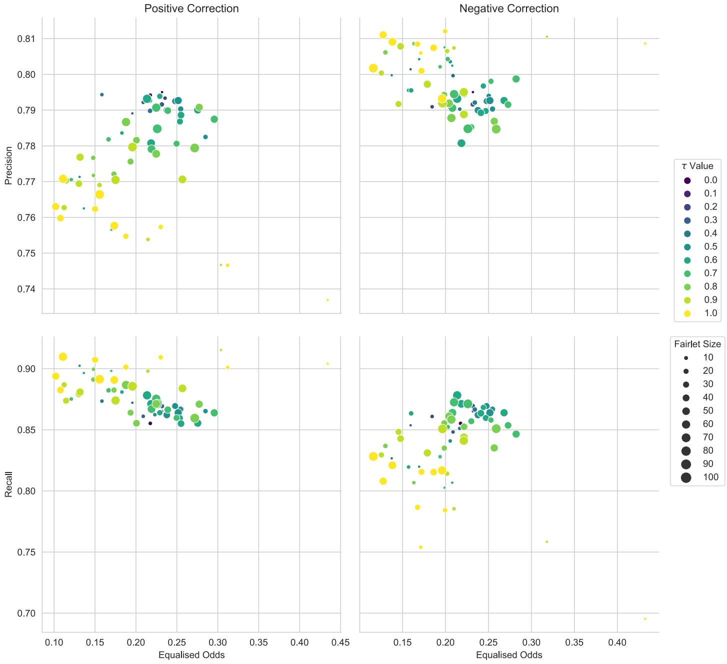

Fig. 18. Precision and recall vs. equalised odds for positive and negative correction experiments on microaggregated *Credit*.

REFERENCES

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. arXiv:1803.02453. Retrieved from https://arxiv.org/abs/1803.02453

[2] Arturs Backurs, Piotr Indyk, Krzysztof Onak, Baruch Schieber, Ali Vakilian, and Tal Wagner. 2019. Scalable fair clustering. In *Proceedings of the International Conference on Machine Learning*. 405–413.

[3] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. 2019. Differential privacy has disparate impact on model accuracy. In *Proceedings of the Advances in Neural Information Processing Systems*. 15479–15488.

[4] Solon Barocas and Andrew D. Selbst. 2016. Big data's disparate impact. *California Law Review* 104 (2016), 671.

[5] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. Fairness in criminal justice risk assessments: The state-of-the-art. *Sociological Methods and Research* (2018).

[6] Michael Buckland and Fredric Gey. 1994. The relationship between recall and precision. *Journal of the American Society for Information Science* 45, 1 (1994), 12–19.

[7] Toon Calders and Sicco Verwer. 2010. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* 21, 2 (2010), 277–292.

[8] L. Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K. Vishnoi. 2019. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 319–328.

[9] Nitesh V. Chawla, Aleksandar Lazarevic, Lawrence O. Hall, and Kevin W. Bowyer. 2003. SMOTEBoost: Improving prediction of the minority class in boosting. In *Proceedings of the European Conference on Principles of Data Mining and Knowledge Discovery*. Springer, 107–119.

[10] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. 2017. Fair clustering through fairlets. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Curran Associates Inc., 5036–5044.

[11] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* 5, 2 (2017), 153–163.

[12] Alexandra Chouldechova and Aaron Roth. 2018. The frontiers of fairness in machine learning. arXiv:1810.08810. Retrieved from https://arxiv.org/abs/1810.08810

[13] Rachel Cummings, Varun Gupta, Dhamma Kimpara, and Jamie Morgenstern. 2019. On the compatibility of privacy and fairness. In *Proceedings of the Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*. 309–315.

[14] Anupam Datta, Shayak Sen, and Michael Carl Tschantz. 2018. Correspondences between privacy and nondiscrimination: Why they should be studied together. arXiv:1808.01735. Retrieved from https://arxiv.org/abs/1808.01735

[15] Sci-kit Learn Developers. 2019. scikit-learn: machine learning in Python. (2019).

[16] Josep Domingo-Ferrer and Vicenç Torra. 2005. Ordinal, continuous and heterogeneous k-anonymity through microaggregation. *Data Mining and Knowledge Discovery* 11, 2 (2005), 195–212.

[17] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. (2017). Retrieved from http://archive.ics.uci.edu/ml

[18] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS'12)*. Association for Computing Machinery, 214–226.

[19] Cynthia Dwork and Aaron Roth. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science* 9, 3–4 (2014), 211–407. DOI: https://doi.org/10.1561/0400000042

[20] Michael D. Ekstrand, Rezvan Joshaghani, and Hoda Mehrpouyan. 2018. Privacy for All: Ensuring fair and equitable privacy protections. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency.* Sorelle A. Friedler and Christo Wilson (Eds.), Proceedings of Machine Learning Research, Vol. 81. PMLR, 35–47.

[21] European Parliament and Council of the European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council. (2016). Retrieved from https://data.europa.eu/eli/reg/2016/679/oj

[22] James R. Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. 2020. An intersectional definition of fairness. In *Proceedings of the 2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE, 1918–1921.

[23] Peter I. Frazier. 2018. A tutorial on Bayesian optimization. arXiv:1807.02811. Retrieved from https://arxiv.org/abs/1807.02811

[24] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 329–338.

[25] Pratik Gajane and Mykola Pechenizkiy. 2017. On formalizing fairness in prediction with machine learning. arXiv:1710.03184. Retrieved from https://arxiv.org/abs/1710.03184

[26] Vladimiro González-Zelaya, Julián Salas, Dennis Prangle, and Paolo Missier. 2021. Optimising fairness through parametrised data sampling. In *Proceedings of the EDBT*. 445–450.

[27] Sara Hajian, Josep Domingo-Ferrer, Anna Monreale, Dino Pedreschi, and Fosca Giannotti. 2015. Discrimination- and privacy-aware patterns. *Data Mining and Knowledge Discovery* 29, 6 (2015), 1733–1782.

[28] Moritz Hardt, Eric Price, Nati Srebro, et al. 2016. Equality of opportunity in supervised learning. In *Proceedings of the Advances in Neural Information Processing Systems*. 3315–3323.

[29] Vasileios Iosifidis and Eirini Ntoutsi. 2019. AdaFair: Cumulative fairness adaptive boosting. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 781–790.

[30] Matthew Jagielski, Michael Aaron Kearns, Saeed Sharifi-Malvajerdi, Jieming Mao, Alina Oprea, Aaron Roth, and Jonathan Ullman. 2019. Differentially private fair learning. In *Proceedings of the International Conference on Machine Learning*. PMLR, 3000–3008.

[31] Faisal Kamiran and Toon Calders. 2010. Classification with no discrimination by preferential sampling. In *Proceedings of the 19th Machine Learning Conf. Belgium and The Netherlands*. 1–6.

[32] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1 (2012), 1–33.

[33] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 35–50.

[34] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. In *Proceedings of the Advances in Neural Information Processing Systems*. 656–666.

[35] Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2018. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 853–862.

[36] Matt J. Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Proceedings of the Advances in Neural Information Processing Systems*. 4066–4076.

[37] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How we analyzed the COMPAS recidivism algorithm. *ProPublica* 9 (2016).

[38] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. 2007. t-Closeness: Privacy beyond k-anonymity and l-diversity. In *Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering*. 106–115.

[39] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, Ioannis Kompatsiaris, Katharina Kinder-Kurlanda, Claudia Wagner, Fariba Karimi, Miriam Fernandez, Harith Alani, Bettina Berendt, Tina Kruegel, Christian Heinze, Klaus Broelemann, Gjergji Kasneci, Thanassis Tiropanis, and Steffen Staab. 2020. Bias in data-driven artificial intelligence systems - an introductory survey. *WIREs Data Mining and Knowledge Discovery* (2020).

[40] David Pujol, Ryan McKenna, Satya Kuppam, Michael Hay, Ashwin Machanavajjhala, and Gerome Miklau. 2020. Fair decision making using privacy-protected data. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*'20)*. Association for Computing Machinery, 189–199.

[41] Borja Rodríguez-Gálvez, Ragnar Thobaben, and Mikael Skoglund. 2020. A variational approach to privacy and fairness. *arXiv* (2020), arXiv–2006.

[42] Ana Rodríguez-Hoyos, José Estrada-Jiménez, David Rebollo-Monedero, Ahmad Mohamad Mezher, Javier Parra-Arnau, and Jordi Forné. 2020. The fast maximum distance to average vector (F-MDAV): An algorithm for k-anonymous microaggregation in big data. *Engineering Applications of Artificial Intelligence* 90 (2020).

[43] Donald B. Rubin. 1973. The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics* (1973), 185–203.

[44] Julián Salas and Josep Domingo-Ferrer. 2018. Some basics on privacy techniques, anonymization and their big data challenges. *Mathematics in Computer Science* 12, 3 (2018), 263–274.

[45] Julián Salas and Vladimiro González-Zelaya. 2020. Fair-MDAV: An algorithm for fair privacy by microaggregation. In *Proceedings of the Modeling Decisions for Artificial Intelligence*. Vicenç Torra, Yasuo Narukawa, Jordi Nin, and Núria Agell (Eds.), Springer International Publishing, Cham, 286–297.

[46] Julián Salas and Vicenç Torra. 2018. A general algorithm for k-anonymity on dynamic databases. In *Proceedings of the Data Privacy Management, Cryptocurrencies and Blockchain Technology*. Springer International Publishing, Cham, 407–414.

[47] Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. 2019. Capuchin: Causal database repair for algorithmic fairness. arXiv:1902.08283. Retrieved from https://arxiv.org/abs/1902.08283

[48] Pierangela Samarati. 2001. Protecting respondents identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering* 13, 6 (2001), 1010–1027.

[49] Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 05 (2002), 557–570.

[50] Depeng Xu, Shuhan Yuan, and Xintao Wu. 2019. Achieving differential privacy and fairness in logistic regression. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW'19)*. Association for Computing Machinery, 594–599.

[51] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness beyond disparate treatment and disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*. 1171–1180.

[52] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. 2019. Fairness constraints: A flexible approach for fair classification. *The Journal of Machine Learning Research* 20, 1 (2019), 2737–2778.

[53] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2015. Fairness constraints: Mechanisms for fair classification. arXiv:1507.05259. Retrieved from https://arxiv.org/abs/1507.05259

[54] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *Proceedings of the International Conference on Machine Learning*. 325–333.

Received 20 July 2022; revised 8 August 2023; accepted 18 August 2023