Utility-Embraced Microaggregation for Machine Learning Applications

SOOBIN LEE 1 AND WON-YONG SHIN 2 (Senior Member, IEEE)

1School of Business and Machine Intelligence & Data Science Laboratory, Yonsei University, Seoul 03722, Republic of Korea
2School of Mathematics and Computing (Computational Science and Engineering), Yonsei University, Seoul 03722, Republic of Korea

Corresponding author: Won-Yong Shin (wy.shin@yonsei.ac.kr).

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A2C3004345), by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2021-0-00347, 6G Post-MAC (POsitioning- & Spectrum-aware intelligent MAC for Computing & Communication Convergence)), and by the Yonsei University Research Fund of 2021 (2021-22-0083).

I. INTRODUCTION
A. BACKGROUND

In the past decade, the advent of technologies to collect and process vast amounts of data has created new opportunities for various business sectors such as healthcare, transportation, finance, and marketing [1], [2]. Such big data technologies enable business entities to make better decisions based on data analytics and to discover unprecedented business products [3]. While the amount of accessible information has become vast, the usefulness of the data in decision-making through collecting and analyzing personal data of individuals has also been emphasized [4]. On the other hand, as massive volumes of personal data are available and machine learning (ML) models become sophisticated, privacy disclosure has become a critical issue [5]. Although the data containing personal records of individuals help improve decision-making via ML models in various business sectors, it should be taken into serious consideration how to prevent re-identification when such data are to be released [6]. Among anonymization techniques, k-anonymous microaggregation has received considerable attention in the literature since it provides a good and flexible balance between privacy protection and data usefulness [7]–[9]. Microaggregation is a family of the most widely used perturbation methods in which microdata sets (i.e., personal records) are aggregated in the sense of preserving privacy through k-anonymity. In other words, perturbed microdata sets are created by aggregating the attributes’ values of
groups of \( k \) records in order to reduce the re-identification risk.

### B. MOTIVATION AND MAIN CONTRIBUTIONS

Despite various studies on designing \( k \)-anonymous microaggregation, their results are not often effective when the microaggregated datasets are used for conducting various ML tasks [3], [10]. In other words, it is largely underexplored how to take into account the data utility (e.g., the test accuracy) for ML applications when microaggregation models are designed. To cope with this problem, it is as important as data privacy protection to use appropriate measures for data utility in ML applications. To date, the degree of data utility has generally been measured at the level of distortion resulting from data perturbation rather than the performance of the intended ML task. Although the performance of ML applications is usually evaluated with respect to the test accuracy, there is a lack of concern on verifying the effectiveness of microaggregated data in various aspects other than information loss [4]. Prior studies on microaggregation models dealing with both data utility and data privacy have not yet utilized the information loss as an indicator of data utility [11], [12].

In this paper, we introduce Util-MA, a novel utility-embraced microaggregation framework for effective ML applications. Unlike previous studies that apply microaggregation directly to raw data [3], [5], [12], [13], we aim to design an end-to-end integrated microaggregation framework that can potentially improve data utility while preserving the \( k \)-anonymity through dimensionality reduction and clustering as preprocessing stages.

More specifically, our Util-MA framework consists of three stages: dimensionality reduction, clustering, and \( k \)-anonymous microaggregation. In the first stage, quasi-identifiers (QIDs) of an original dataset are used as input and are converted into a low-dimensional dataset through one of two different types of dimensionality reduction techniques, including the principal component analysis (PCA)\(^1\) and autoencoder (AE), for not only denoising but also computational efficiency. In the second stage, the dataset transformed by dimensionality reduction is divided into multiple clusters through one of the three clustering methods such as the \( K \)-means++ [14], the Gaussian mixed model (GMM) [15]–[18], and the density-based spatial clustering of noisy applications (DBSCAN) [19], [20]. In this step, the compressed QID dataset is partitioned into several groups, each of which is fed into microaggregation models in a separate manner. This pre-clustering before microaggregation is inspired by the fact that almost dissimilar records would not be grouped into the same microaggregation cell [21]. After the two preprocessing stages, we perform \( k \)-anonymous microaggregation using the MDAV-generic [22].

\(^1\)PCA is the process of computing only a few principal components and using them to perform a change of basis on the data.

To validate the superiority of our Util-MA framework, we perform empirical evaluations using various real-world datasets whose number of attributes ranges from 9 to 34. Depending on the types of datasets, we solve either binary or multi-class classification problems as intended ML tasks. We empirically prove the effectiveness of data utility by adopting the \( F_1 \)-score and the macro/micro-averaged \( F_1 \) score for binary and multi-class classifications, respectively, as performance metrics. We comprehensively perform empirical comparisons with the benchmark microaggregation method without preprocessing stages including dimensionality reduction and clustering.

Experimental results demonstrate that the proposed Util-MA framework almost outperforms the benchmark method in terms of several test accuracy metrics. Our Util-MA framework is shown to achieve substantial gains up to 85.42\% compared with the benchmark method. In addition, we offer interpretations of our experimental results. Although the Util-MA framework is superior to the benchmark method for most cases, it is unclear how such a gain is achieved. To explain this superiority, we investigate the importance of preprocessing by measuring the following key performance indicators (KPIs) of clustering given ground truth labels [23]–[26]: homogeneity, completeness, and \( V \)-measure. The empirical findings indicate that the clustering stage of Util-MA leads to high performance on the ML task (i.e., classification) when the clustering results substantially coincide with the ground truth labels. Moreover, if a specific combination of dimensionality reduction and clustering in Util-MA shows high clustering performance in terms of all three KPIs, then the classification along with the microaggregated dataset tends to exhibit relatively high classification accuracies compared with the benchmark method. In other words, a close relationship between the KPIs of clustering and the classification accuracies is more evident when the improvement rate of Util-MA over the benchmark method is established. Therefore, our pre-clustering stage turns out to be indeed effective to further enhance the data utility when the released dataset is used for diverse ML applications.

### C. ORGANIZATION AND NOTATIONS

The remainder of this paper is organized as follows. In Section II, we summarize significant studies that are related to our work. In Section III, we explain the methodology of our study, including the basic settings and an overview of our Util-MA framework. Implementation details and experimental results are discussed in Section IV. Finally, we provide a summary and concluding remarks in Section V.

Table 1 summarizes the notation that is used in this paper. This notation will be formally defined in the following sections when we introduce our methodology and the technical details.

### II. RELATED WORK

The framework that we propose in this paper is related to three broader areas of research, namely standard \( k \)-
TABLE 1. Summary of notations.

| Notation | Description |
|----------|-------------|
| X        | Original normalized dataset |
| Ź        | Microaggregated QID dataset |
| V        | Low-dimensional dataset |
| n        | Number of records |
| m        | Number of QIDs |
| ̃n        | Dimension of the QID after dimensionality reduction |
| G_i      | Each partitioned dataset after pre-clustering |
| C        | Set of microcells from each G_i |

anonymous microaggregation models, MDAV models, and applications of microaggregation.

**Standard k-anonymous microaggregation models.** It has been of paramount concern to maintain the re-identification risk below a certain level since the top priority of the microdata statistical disclosure control (SDC)

2 is “privacy-first”. The best-known approach for data privacy protection is k-anonymous microaggregation, which is a perturbing technique that focuses on masking QIDs. The microaggregation approach consists of two steps, namely partitioning and aggregation. First, the approach creates clusters, each of which contains at least k records or records of similar size. Second, it replaces the records within each cluster with an aggregate value, e.g., the arithmetic mean value [8], [27]. Despite its popularity, some refinements of the k-anonymity were presented in order to alleviate an inherent risk of attacks based on the potential lack of disclosure: p-sensitive k-anonymity, l-diversity, and t-closeness microaggregation models [6], [28]–[33]. Designing the optimal multivariate microaggregation model is an NP-hard problem. Due to this complexity issue, there have been many heuristic methods in the literature. As a recent heuristic approach for reducing the perturbation error caused by microaggregation, transforming the multivariate microaggregation problem into its univariate counterpart by ordering microdata records with a proper Hamiltonian path and applying an optimal univariate solution were presented in [34].

**MDAV models.** The goal of k-anonymous microaggregation models is to minimize the information loss while preserving k-anonymity. Many heuristic microaggregation methods are categorized into fixed-size and variable-size microaggregation models. For the fixed-size microaggregation, the most widely-used model is the maximum distance to average vector (MDAV), which is a less computationally demanding variation of the maximum distance (MD) model [35]. Alternatively, there are a number of variable-size microaggregation models as variants of the original MDAV model because fixed-size models often do not fully reflect the characteristics of the data. Unlike the original MDAV model, a variant of MDAV in [36] assigns the remaining records to their nearest cluster at the last step. For a more general application, MDAV-generic was proposed to be used along with any type of attributes, aggregation operators, and distance measures [22]. V-MDAV [13] allows records to be grouped into the same cluster if their similarity is greater than a certain level controlled by a user-defined parameter. In addition, there have also been efforts to improve the large computational complexity of MDAV for practical applications. For example, F-MDAV [5] reduces the computation time using precomputation and partial selection. Significant gains in terms of complexity were achieved by performing PCA before partitioning [21].

**Applications of microaggregation.** When de-identified datasets are created and distributed to the public, the data can be used for a variety of ML tasks related to healthcare, finance, and social media [4]. While these data have quite private individual information that contains the risk of privacy leakage, they can also be useful for effectively solving many ML problems. For example, various anonymization techniques have been developed in either recommender systems such as collaborative filtering based on users’ preferences and profiles [37]–[39] or location-based services (LBSs) that also deal with private user information such as the geographic location of users [40].

**Discussion.** Besides MDAV, there are more microaggregation models employing different clustering strategies: the density-based algorithm (DBA) [36] first partitions records in descending order of their densities; and a microaggregation model via sorting before partitioning was shown to prevent very dissimilar records to be grouped into the same cluster [41]. Similar to MDAV, other microaggregation models such as the minimum spanning tree (MST) [42] and the two fixed-reference points (TFRP) [43] repeatedly build a cluster at a time while preserving the k-anonymity and low information loss (e.g., the \( \mu \)-Approx [44]). In addition, recent microaggregation studies focus on more effective partitioning by splitting data beforehand. Models in [10], [45] use two-step clustering strategies to better configure microcells. The transformation based method (TBM) [3] uses a pre-processing step to handle the categorical attributes appropriately. Despite these contributions, no prior model has suggested the microaggregation-model-agnostic framework applicable to general microaggregation or clustering methods.

Furthermore, previous studies have focused mostly on reducing the information loss and/or computational complexity when designing microaggregation models, whereas they have not yet explored data utility such as the test accuracy in ML tasks as the most popular measure. More precisely, the performance of microaggregation models was evaluated in terms of the information loss in [3], [13], [22], [36], [41]–[45] and the runtime complexity in [3], [5], [41], [43]. The performance was also evaluated from the perspective of the distance-based linkage disclosure (e.g., the model in [3]) and the diversity of sensitive attributes within the fixed size groups (e.g., the model in [45]).

On the other hand, in the research field of de-identification, it is necessary to find an appropriate balance between the
privacy protection and the usefulness of data. Since the privacy is a sensitive issue, MDAV and other follow-up microaggregation studies have primarily focused on preventing re-identification. However, the higher the privacy level, the greater the distortion of the data, which inevitably reduces the data utility. Note that, although several models were proposed to increase the data utility while maintaining privacy levels above a certain level, they still used data utility metrics as an error from the original dataset [11], [12].

III. METHODOLOGY
In this section, we introduce the overview of the proposed Util-MA framework. We first describe our microaggregation model with basic settings. Then, we delineate each step of the framework precisely.

A. BASIC SETTINGS
We start with a microdata set consisting of confidential attributes and QIDs [8], [22], [46]. Before describing our Util-MA framework, we briefly introduce the basic structure of the microdata set. First, identifiers (IDs) are attributes that unambiguously identify a user. These include passport numbers, social security numbers, and full names. They are usually removed or encrypted during the pre-processing phase. Unlike IDs, QIDs cannot be removed from the dataset. They are a set of attributes that can be associated with external information, making themselves a potential source for re-identification. Confidential attributes are attributes that contain sensitive information about users (e.g., diagnoses). Since there exist many opportunities to utilize the confidential attributes, they are not removed or encrypted.

Unperturbed confidential attributes are often necessary to effectively build an ML model (e.g., an ML classifier) in practical circumstances. However, QIDs typically contain demographic information including age, gender, address, or physical features, which can be used to re-identify the original dataset with other available information [7], [47]. Therefore, for privacy protection, the QIDs must be anonymized before publication or release. We denote an original multivariate QID dataset that should be microaggregated as $X \in \mathbb{R}^{n \times m}$, which consists of $n$ records and $m$ QIDs. For computational convenience, we initially normalize the dataset by following the common practice in SDC [4], [48].

The datasets used in this paper are binary or multi-class datasets. Since microaggregation models employed in this paper operate only with numerical attributes, the datasets including the categorical attributes need to be preprocessed beforehand [4]. Although there are some studies on how to vectorize categorical attributes, we convert them to one-hot encoded vectors to simplify experimental evaluations since, otherwise, we need other preprocessing steps along with the categorical attributes in order to build an ML model.
B. OVERVIEW OF OUR Util-MA FRAMEWORK

In this subsection, we briefly describe our Util-MA framework composed of the following three stages: 1) dimensionality reduction, 2) clustering, and 3) microaggregation. More specifically, we first reduce the dimensionality of the QIDs in the original dataset and then construct multiple clusters in such a way that similar records are grouped into the same cluster beforehand. Within each cluster, we run a microaggregation model corresponding. The microaggregated data are finally combined with the original dataset and are distributed to the public.

Fig. 1 illustrates the schematic overview of Util-MA. First, let us mention the dimensionality reduction stage. The original dataset is often difficult to be microaggregated appropriately due to its large size and/or the intrinsic noise in the data. Thus, the normalized QID dataset \(X\) is transformed to a low-dimensional dataset, denoted by a matrix \(V \in \mathbb{R}^{n \times m}\), using dimensionality reduction techniques such as PCA and AE belonging to linear projection and nonlinear representation approaches, respectively, where \(n\) is the dimension of the QID after reduction. Here, the dataset \(V\) in a low-dimensional space retains some meaningful properties of the original dataset while keeping closer to its intrinsic dimension by denoising the original data so that it is more computationally efficient. In the second stage, transformed records are partitioned into multiple clusters. This pre-clustering before microaggregation is basically inspired by the insight that the almost dissimilar records would not be grouped into the same microaggregation cell. After the pre-clustering step, we run \(k\)-anonymous microaggregation models repeatedly only for the records within the belonging cluster. For example, when there are three clusters, we conduct the microaggregation three times along with the intra-cluster records. Then, the \(k\)-anonymous microaggregated QID dataset, denoted by a matrix \(\tilde{X} \in \mathbb{R}^{n \times m}\), is attained accordingly. Finally, the de-identified dataset, consisting of the microaggregated QIDs \(\tilde{X}\) and confidential attributes, is ready for being released to the public for potential ML applications (e.g., binary and multi-class classification problems).

By using the de-identified dataset, we are able to solve a variety of ML application problems without concerns about data privacy. In our study, we aim to quantitatively evaluate the performance of microaggregation in terms of data utility. Due to the fact that the information loss itself may be less relevant to the performance of ML tasks, we adopt the \(F_1\)-score as a popular measure of a test’s accuracy when we pay our attention to classification problems.

IV. EMPIRICAL EVALUATION

In this section, we first present Util-MA, an end-to-end solution to the problem of microaggregation embracing data utility. Then, we describe real-world datasets used in the evaluation. After describing our performance metrics and experimental settings, we comprehensively evaluate the performance of our Util-MA framework and benchmark approaches.

A. IMPLEMENTATION DETAILS

The overall procedure of our framework is described in Algorithm 1, where the normalized QID dataset \(X\) is used as input. We elaborate on each stage in the proposed Util-MA framework along with technical details.

For the dimensionality reduction stage, we use both PCA and AE. The PCA projects the QID dataset \(X\) onto its first few principal components to obtain a low-dimensional approximation \(V\) of \(X\) along with a pre-defined target dimension size as input \([21],[49],[50]\). The PCA problem can be represented as the following optimization:

\[
\hat{V} = \arg \min_{V} \| (I_n - VV^T) X \|_F^2,
\]

\[
\text{s.t. } V^TV = I_n,
\]

where \(I_n\) is the identify matrix of size \(n \times n\); the superscript \(T\) indicates the transpose of a matrix, and \(\| \cdot \|_F\) is the Frobenius norm of a matrix. On the other hand, the AE is generally built upon a multi-layer neural network including an encoder and a decoder according to multiple hyperparameters (e.g., a learning rate), nonlinear activation function(s), and a given optimizer (e.g., stochastic gradient descent (SGD)) \([51]\). We obtain a compressed representation \(V\) of the input data \(X\) at the hidden layer of the AE \([52]\). The AE problem can be represented as the following optimization:

\[
\{\hat{W}_e,\hat{W}_d,\hat{b}_e,\hat{b}_d\} = \arg \min_{W_e,W_d,b_e,b_d} \mathcal{J}(W_e,W_d,b_e,b_d),
\]

where the loss function is defined as

\[
\mathcal{J}(W_e,W_d,b_e,b_d) = \mathcal{L}(X,Z) + \lambda g(W_e,W_d).
\]

Here, \(W_e\) and \(b_e\) are the weight matrix and bias in the encoder, respectively; \(W_d\) and \(b_d\) are the weight matrix and bias in the decoder, respectively; \(Z\) is the output of the decoder with its activation function; \(\mathcal{L}(\cdot,\cdot)\) is the reconstruction loss term; and \(g(\cdot,\cdot)\) is a regularization with a coefficient \(\lambda\). The dimensionality reduction function is represented as \(f_{DR}\) whose input is given by the QID dataset \(X\) and dimension size \(n\) (refer to line 1 of Algorithm 1).

Next, we turn to the clustering stage after dimensionality reduction, which partitions the low-dimensional dataset \(V\) into several groups via one of clustering methods. More specifically, as the most popular method, we first employ \(k\)-means clustering, which assigns records into \(K\) clusters in such a way that each record belongs to the cluster with the nearest mean, minimizing intra-cluster variances \([53],[54]\). For the \(K\)-means clustering problem, we aim to choose \(K\) centers so as to minimize the objective function \(\phi\):

\[
\phi = \sum_{v_i \in C} \min_{c \in C} ||v_i - c||^2,
\]

where \(v_i \in \mathbb{R}^n\) is \(i\)-th row vector of \(V\) and \(C\) is the set of \(K\) centers. That is, we choose \(K\) centers in the sense of minimizing the sum of the squared distances between each point and its closest center in an iterative manner. In our study, we select initial points along the \(K\)-means++ algorithm \([14]\).
for better optimization. For comparison, we also adopt other clustering methods such as the GMM [15]–[18] and the DBSCAN [19], [20]. For the GMM clustering problem, the marginal probability $P(V = v_i)$ can be expressed as

$$P(V = v_i) = \sum_{l=1}^{K} \pi_l N(v_i | \mu_l, \Sigma_l),$$  \hspace{1cm} (1)$$

where $\mu_l$ and $\Sigma_l$ are the mean and the variance, respectively, of the $l$-th Gaussian distribution of $N(\cdot)$, and $\pi_l$ is a mixing coefficient. Then, using (1), the log-likelihood is given by

$$\ln P(V | \pi_l, \mu_l, \Sigma_l) = \sum_{i=1}^{n} \ln \left( \sum_{l=1}^{K} \pi_l N(v_i | \mu_l, \Sigma_l) \right).$$  \hspace{1cm} (2)$$

We perform an expectation maximization (EM) algorithm in the context of the GMM in order to find the set of $K$ centers. Using (1) and (2), we alternately perform an E-step (expectation) and an M-step (maximization) until convergence. For the DBSCAN clustering problem, we use two input parameters $\epsilon$ and $\text{minPts}$ to estimate the density of a particular point’s neighborhood. Here, $\epsilon$ is the radius of a point’s neighborhood, and $\text{minPts}$ is the minimum number of records in an $\epsilon$-neighborhood of a point in order to be a cluster. Thus, for a given dataset $V$, the DBSCAN algorithm calculates the local density of a point $v_i \in V$ as the total number of points in its $\epsilon$-neighborhood (i.e., the cardinality of $N_{\epsilon}(v_i)$), where

$$N_{\epsilon}(v_i) = \{ v_j : \forall j, d(v_i, v_j) < \epsilon \}.$$

Here, $d(v_i, v_j)$ is the Euclidean distance between two points $v_i$ and $v_j$. For more detailed description of DBSCAN, we refer to [55], [56]. The clustering function is represented as $f_{\text{DBSCAN}}$ whose input is given by the low-dimensional dataset $V$ and the number of clusters, $K$ (refer to line 2 of Algorithm 1). Note that, unlike the $K$-means++ and GMM, we do not feed $K$ into the DBSCAN model as input. Instead, for the DBSCAN algorithm, two input parameters $\epsilon$ and $\text{minPts}$ are appropriately determined by parameter tuning via grid search. In our study, we slightly modify the original DBSCAN in such a way that the values of $\epsilon$ and $\text{minPts}$ are chosen in the sense of achieving the highest V-measure in each dataset after running the DBSCAN, where the V-measure is a KPI of clustering and will be specified in Section IV-E2.

After the above preprocessing stages, we perform $k$-anonymous microaggregation (refer to lines 3–10 of Algorithm 1). First, we construct a set $C_i$ of microcells from each clustered dataset $G_i$ through the quantization function $f_{\text{micro}}$, where $i$ indicates the cluster index (refer to line 4 of Algorithm 1). Then, we pass each record $v_p \in V$ in a microcell $C_{i,j}$ through the reconstruction function $f_{\text{agg}}$ to obtain an aggregated value $\tilde{v}_p$, where $j$ denotes the index of microcells (refer to line 5–9 of Algorithm 1). The aggregated value $\tilde{v}_p$ is determined by the given aggregation operator. Therefore, every record in each microcell is changed to the same aggregated value. In our study, we apply the MDAV-generic [22] for the $k$-anonymous microaggregation, which is summarized in Algorithm 2 for each partitioned dataset $G_i$.

### B. DATASETS

Seven real-world attributed datasets described in Table 2 are used for experimental evaluations. Note that, since each dataset has a different range of attributes, it may distort the clustering results. Thus, we normalize the above datasets by the min-max scaling technique [45]. We replace missing values with the most frequent value in each associated attribute.
| Name                              | # of attributes | # of records | # of classes |
|----------------------------------|----------------|-------------|--------------|
| Credit Approval                  | 14             | 690         | 2            |
| Breast Cancer Wisconsin          | 9              | 699         | 2            |
| Breast Cancer                    | 9              | 286         | 2            |
| Diabetic Retinopathy Debrecen    | 19             | 1,151       | 2            |
| Cardiotocography                 | 21             | 2,126       | 3            |
| Heart Disease                    | 13             | 303         | 5            |
| Dermatology                      | 34             | 366         | 6            |

**C. PERFORMANCE METRICS**

We validate the performance of the proposed Util-MA framework and the benchmark method by adopting several metrics as follows. We verify whether the disclosed de-identification dataset achieves high utility in practice.

As a classification method, we chose logistic regression for binary classification and one-vs-rest logistic regression for multi-class classification. The classification is executed using a de-identified dataset, and then the predicted output is compared with the true class label masked from the original dataset. More precisely, given a series of real class labels, the performance of ML applications is evaluated using the test accuracy. As a popular performance metric for binary classification, we adopt the F1-score, which is defined as the harmonic mean of recall and precision [4], [57]. The precision is the number of true positives divided by the number of all positive predictions, and the recall is the number of true positives divided by the number of all records that originally have the positive label. For multi-class classification, we adopt the macro- and micro-averaged F1-scores. Since we use both binary and multi-class datasets, we use the accuracy and the F1-score for binary classification, and the macro/micro-averaged F1-score for multi-class classification. Moreover, we adopt the area under the ROC curve (AUC) score, where an ROC curve is a graph showing the performance of a classification model and plots the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.

**D. EXPERIMENTAL SETUP**

We first describe the experimental settings of our Util-MA framework. For the first phase in Algorithm 1, the dimension size $m$ in the function $f_{DR}$ is set to $m/3$, where $m$ is the number of QID attributes. For the parameter $K$, indicating the number of clusters in the function $f_{CL}$, finding the optimal value of $K$ in terms of maximizing our performance metrics would not be possible unless ground truth labels are available. Since a large value of $K$ may interrupt effective microaggregation, we set $K$ to the number of classes in each dataset, which is assumed to be available beforehand. Note that DBSCAN does not use the parameter $K$ as input. For the DBSCAN algorithm, since two input parameters $\epsilon$ and $\text{minPts}$ are appropriately determined, the values of $\epsilon$ and $\text{minPts}$ vary over different datasets. In our experiments, we choose the values of $\epsilon$ and $\text{minPts}$ in such a way that the highest V-measure (to be specified in Section IV-E2) in each dataset is achieved after running the DBSCAN. The parameter $k$ in the function $f_{micro}$ is set as both 11 and 19. For comparison, we employ MDAV-generic [22] as a benchmark method in our study.

Unless otherwise stated, for each dataset, each experiment runs over 30 different random splits of training and test sets to

---

3https://archive.ics.uci.edu/ml/datasets/credit+approval.
4https://archive.ics.uci.edu/ml/datasets/breast-cancer+wisconsin+(original).
5https://archive.ics.uci.edu/ml/datasets/Breast+Cancer.
6https://archive.ics.uci.edu/ml/datasets/Diabetic+Retinopathy+Debrecen+Data+Set.
7https://archive.ics.uci.edu/ml/datasets/Cardiotocography.
8https://archive.ics.uci.edu/ml/datasets/heart+disease.
9https://archive.ics.uci.edu/ml/datasets/dermatology.
calculate the average performance. We use 40% of the dataset as the test set.

E. EXPERIMENTAL RESULTS

Our empirical study in this subsection is designed to answer the following two key research questions:

- Q1. Which combination of \( f_{DR} \) and \( f_{CL} \) dominate others in comparison with the benchmark method?
- Q2. Does our clustering stage positively influence the performance on the ML problem?
- Q3. How does the performance behave when cross-validation is used?

To answer these questions, we carry out comprehensive experiments in the following.

1) Comparative Study Among Various Methods

The performance comparison between the benchmark method (i.e., MDAV-generic [22]) and our Util-MA framework with possible combinations of dimensionality reduction and clustering is comprehensively presented in Table 3 for seven real-world datasets. In the table, the performance is shown with respect to the \( F_1^{(1)} \) and \( F_1^{(2)} \), where \( F_1^{(1)} \) indicates the accuracy and the macro-averaged \( F_1 \)-score when the task is binary and multi-class classifications, respectively; and \( F_1^{(2)} \) refers to the general \( F_1 \)-score and the micro-averaged \( F_1 \)-score for the binary and multi-class classification problems, respectively.

From Table 3, our findings are as follows:

- The proposed Util-MA framework outperforms the benchmark method without any dimensionality reduction and clustering except only for one case where the Credit Approval is used for \( k = 19 \). In other words, at least one pair of dimensionality reduction and clustering in Util-MA performs better than the benchmark method.
- In particular, the improvement rate of Util-MA (A) over the benchmark method (B) is the largest when the Breast Cancer dataset is used; the maximum improvement rate of 85.42% is achieved, where the improvement rate (%) is given by \( \frac{A-B}{B} \times 100 \). Thus, it is confirmed that significant gains can be achieved using our framework.
- In our Util-MA framework, the GMM-based clustering exhibits higher accuracies than those using other two clustering approaches for 50% of all cases.
- However, in Util-MA, a certain pair of dimensionality reduction and clustering does not dominate other pairs. That is, the best combination of dimensionality reduction and clustering in Util-MA varies depending on datasets.

Additionally, the performance comparison between the benchmark method and our Util-MA framework with six combinations of dimensionality reduction and clustering is comprehensively presented in Table 4 with respect to the AUC for seven real-world datasets. From the table, experimental results essentially show a tendency similar to those in Table 3 while consistently exhibiting the superiority of Util-MA over the benchmark method.

2) The Effect of Clustering

Although the Util-MA framework is superior to the benchmark method for most cases, it is unclear where the gain is originated. To provide interpretations for our results, we analyze the effect of clustering when the number of clusters, \( K \), is set to the number of classes in each dataset. To this end, we adopt the following three KPIs [23]–[26]:

- Homogeneity: This KPI measures how many samples in a cluster are similar to each other using Shannon’s entropy based on the ground truth labels. Homogeneity is met when each cluster contains only records belonging to a single class.
- Completeness: Completeness is also characterized given the ground truth labels. If all the records belonging to a single class are assigned to the same cluster, then the completeness is fulfilled with the value of 1.
- V-measure: V-measure is an entropy-based measure representing how successfully the criteria of both homogeneity and completeness are fulfilled. In general, this KPI is calculated as the harmonic mean of homogeneity and completeness similarly as in the \( F_1 \)-score [58].

Note that the above KPIs lie between 0 and 1; higher values represent better performance. In Fig. 2, clustering performance is illustrated in terms of three KPIs for each dataset, where all combinations of dimensionality reduction and clustering in the Util-MA framework are taken into account. From the figure, we would like to make the following insightful observations:

- It is likely that using the clustering stage leads to high performance on the ML task (i.e., classification) when the created clusters are similar to the ground truth labels. For example, it is seen that, in the Breast Cancer and Cardiotocography datasets, the cases of AE+DBSCAN and PCA+DBSCAN achieve high clustering performance in three KPIs while exhibiting high accuracies in \( F_1^{(1)} \) and \( F_1^{(2)} \) (refer to Table 3). In addition, in the Dermatology dataset, the cases of AE+GMM and PCA+GMM reveal high performance on both the three KPIs of clustering and the classification accuracies.
- Such a close relationship between the KPIs of clustering and the classification accuracies is more evident when the improvement rate of Util-MA over the benchmark method is investigated. For example, in the Breast Cancer dataset, the first and second best performers with respect to the clustering performance are the cases of AE+DBSCAN and PCA+DBSCAN, which show 62.04% and 85.42% of improvement rates, respectively, in the \( F_1 \)-score. The difference in clustering performance between two cases, namely AE+GMM and PCA+GMM, and others is the most significant in the Dermatology dataset; the cases of AE+GMM and PCA+GMM show 26.58% and 24.55% of improvement.
rates, respectively, over the benchmark method in terms of the macro-averaged $F_1$ score.

- This implies that clustering plays a crucial role as a preprocessing step in ensuring the data utility.

### 3) Results Using 5-Fold Cross-Validation

Instead of multiple random splits of each dataset, we evaluate the performance by conducting 5-fold cross-validation similarly as in [59]. We split each dataset into two sets: 80% as the training set and 20% as the test set. The performance comparison between the benchmark method [22] and our Util-MA framework with 6 combinations is comprehensively presented in Table 5 for seven real-world datasets. From the table, it is seen that the performance of our Util-MA framework is superior to that of the benchmark method without any dimensionality reduction and clustering by up to the improvement rate of 103.27% when the Breast Cancer dataset is used. Overall, the results using 5-fold cross validation exhibit a similar tendency to those in Table 3.

### V. CONCLUDING REMARKS

In this paper, we explored a simple yet important problem of how to take into account data utility when developing microaggregation models. To this end, we proposed Util-MA, a new utility-embraced microaggregation framework for effectively solving ML problems. More specifically, we designed a unified and model-agnostic framework that can potentially enhance the data utility while preserving the $k$-anonymity through preprocessing stages including dimensionality reduction and clustering. By using various real-world datasets, we empirically demonstrated that the Util-MA framework outperforms the benchmark method without any preprocessing stages for almost all cases in terms of test accuracy metrics while exhibiting substantial gains up to 103.27%. Furthermore, we investigated the effect of clus-
Clustering performance among all combinations of dimensionality reduction and clustering in Util-MA in terms of the homogeneity, completeness, and V-measure.

Potential avenues of future research include the design of deep learning-based microaggregation methods incorporating preprocessing stages. Future research directions also include the design of explanation models for the target ML task when a microaggregated dataset is used.

REFERENCES

[1] D. Li, X. Shen, and L. Wang, “Connected geomatics in the big data era,” Int. J. Digit. Earth, vol. 11, no. 2, pp. 139–153, Apr. 2017.
[2] W. Raghupathi and V. Raghupathi, “Big data analytics in healthcare: Promise and potential,” Health Inf. Sci. Syst., vol. 2, no. 1, p. 3, Feb. 2014.
[3] M. Salari, S. Jalili, and R. Mortazavi, “TBM, a transformation based
method for microaggregation of large volume mixed data,” Data Min. Knowl. Discov., vol. 31, no. 1, pp. 65–91, Jan. 2017.

[4] A. Rodríguez-Hoyos, J. Estrada-Jiménez, D. Rebollo-Monedero, J. Parra-Arnau, and J. Forné, “Does \( k \)-anonymous microaggregation affect machine-learned macrotrends?” IEEE Access, vol. 6, no. 1, pp. 28 238–28 277, May 2018.

[5] A. Rodríguez-Hoyos, J. Estrada-Jiménez, D. Rebollo-Monedero, A. M. Mezher, J. Parra-Arnau, and J. Forné, “The fast maximum distance to average vector (F-MDAV): An algorithm for \( k \)-anonymous microaggregation in big data,” Eng. Appl. Artif. Intell., vol. 90, no. 1, pp. 1–12, Apr. 2020.

[6] R. W. Wong, J. Li, A. W. C. Fu, and K. Wang, “(\( \alpha \), \( \beta \))-anonymity: an enhanced \( k \)-anonymity model for privacy preserving data publishing,” in Proc. 12th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD ’06), New York, NY, USA, Aug. 2006, pp. 754–759.

[7] M. Tempel, B. Meindl, A. Kowarik, and S. Chen, “Introduction to statistical disclosure control (SDC),” IHSN Working Paper No. 007, Aug. 2014.

[8] P. Samarati and L. Sweeney, “Protecting privacy when disclosing information: \( k \)-anonymity and its enforcement through generalization and suppression,” SRI Computer Science Laboratory, Tech. Rep. SRI-CSL-98-199, Mar. 1998.

[9] A. Zigomilotos, F. Casino, A. Solanas, and C. Pataskis, “A survey on privacy properties for data publishing of relational data,” IEEE Access, vol. 8, pp. 51 071–51 099, 2020.

[10] J. Han, J. Yu, Y. Mo, J. Lu, and H. Liu, “MAGE: A semantics retaining \( k \)-anonymization method for mixed data,” Knowl.-Based Syst., vol. 55, no. 1, pp. 75–86, Jan. 2014.

[11] A. C. Fadel, L. S. Ochi, J. A. de Moura Brito, and G. S. Seman, “Microaggregation heuristic applied to statistical disclosure control,” Inf. Sci., vol. 548, no. 1, pp. 37–55, Feb. 2021.

[12] R. Mortazavi and S. Jalili, “Fast-data oriented microaggregation algorithm for large numerical datasets,” Knowl.-Based Syst., vol. 67, no. 1, pp. 195–205, Sep. 2014.

[13] A. Solanas, A. Martinez-Balleste, and J. Domingo-Ferrer, “\( V \)-MDA: A multivariate microaggregation with variable group size,” in Proc. 17th COMPSTAT Symp. IASC, Rome, Italy, Aug. 2006, pp. 106–115.

[14] D. Arthur and S. Vassilvitskii, “\( k \)-means++: The advantages of careful seeding,” in Proc. 18th Annu. ACM-SIAM Symp. Discrete Algorithms (SODA ’07), New Orleans, LA, USA, Jan. 2007, pp. 1027–1035.

[15] L. Xu and M. I. Jordan, “On convergence properties of the EM algorithm for Gaussian mixtures,” Neural Comput., vol. 8, no. 1, pp. 129–151, Jan. 1996.

[16] C. Maugis, G. Celeux, and M.-L. Martin-Magniette, “Variable selection for clustering with Gaussian mixture models,” Biometrics, vol. 65, no. 3, pp. 701–709, Sep. 2009.

[17] C. Amendola, J.-C. Fauergere, and B. Sturmels, “Moment variances of Gaussian mixtures,” J. Algebra. Stat., vol. 7, no. 1, pp. 14–28, Jul. 2016.

[18] E. Miquelis, N. Koshev, and E. S. Helou, “A backprojection slice theorem for tomographic reconstruction,” IEEE Trans. Image Process., vol. 27, no. 2, pp. 894–906, Feb. 2018.

[19] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al., “A density-based algorithm for discovering clusters in large spatial databases with noise,” in Proc. 2nd Int. Conf. Knowledge Discovery and Data Mining (KDD ’96), Portland, OR, USA, 1996, pp. 226–231.

[20] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, “DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN,” ACM Trans. Database Syst., vol. 42, no. 3, pp. 1–21, Sep. 2017.

[21] D. Rebollo-Monedero, A. M. Mezher, X. C. Colomé, J. Forné, and M. So- riano, “Efficient \( k \)-anonymous microaggregation of multivariate numerical data via principal component analysis,” Inf. Sci., vol. 503, no. 1, pp. 417–443, Nov. 2019.

[22] J. Domingo-Ferrer and V. Torra, “Ordinal, continuous and heterogeneous \( k \)-anonymity through microaggregation,” Data Mining & Knowledge Discovery, vol. 11, no. 2, pp. 195–212, Aug. 2005.

[23] M. Meilă, “Comparing clusterings—an information based distance,” J. Multivar. Anal., vol. 98, no. 5, pp. 873–895, May 2007.

[24] R. Reichart and A. Rappoport, “The NVI clustering evaluation measure,” in Proc. 13th Conf. Comput. Nat. Lang. Learn. (CoNLL 2009), Boulder, CO, USA, Jun. 2009, pp. 165–173.

[25] A. Rosenberg and J. Hirschberg, “\( V\)-measure: A conditional entropy-based external cluster evaluation measure,” in Proc. 2007 Int. Conf. Empir. Methods Nat. Lang. Process. Comput. Nat. Lang. Learn. (EMNLP-CoNLL), Prague, Czech Republic, Jun. 2007, pp. 410–420.

[26] E. Aldana-Bobadilla and A. Kuri-Morales, “A clustering method based on the maximum entropy principle,” Entropy, vol. 17, no. 1, pp. 151–180, Jan. 2015.

[27] D. Defays and M. N. Anwar, “Masking microdata using microaggregation,” Journal of Official Statistics, vol. 14, no. 4, pp. 449–461, Dec. 1998.

[28] T. M. Truta and B. Vinay, “Privacy protection: \( p \)-sensitive \( k \)-anonymity property,” in Proc. 22nd IEEE Int. Conf. Data Eng. Workshops (ICDEW ’06), Atlanta, GA, USA, Apr. 2006, pp. 1–10.

[29] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, “\( l \)-diversity: Privacy beyond \( k \)-anonymity,” ACM Trans. Knowl. Discov. Data, vol. 1, no. 1, pp. 1–52, Mar. 2007.

[30] N. Li, T. Li, and S. Venkitasubramaniam, “\( \ell \)-closeness: Privacy beyond \( k \)-anonymity and \( l \)-diversity,” in Proc. 23rd IEEE Int. Conf. Data Eng., Istanbul, Turkey, Apr. 2007, pp. 106–115.

[31] J. Soria-Comas, J. Domingo-Ferrer, D. Sanchez, and S. Martinez, “\( \ell \)-closeness through microaggregation: Strict privacy with enhanced utility preservation,” IEEE Trans. Knowl. Data Eng., vol. 27, no. 11, pp. 3096–3110, Nov. 2015.

[32] Y. Yan, A. H. Eyeloko, A. Mahmood, J. Li, Z. Dong, and F. Xu, “Privacy preserving dynamic data release against synonymous linkage based on microaggregation,” Sci. Reports, vol. 12, no. 2352, Feb. 2022.
J. L. Lin, T. H. Wen, J. C. Hsieh, and P. C. Chang, “Density-based N. Senavirathne and V. Torra, “On the role of data anonymization in J. Domingo-Ferrer and J. M. Mateo-Sanz, “Practical data-oriented mi- croaggregation through Hamiltonian paths and optimal univariate mi- croaggregation,” Symmetry, vol. 13, no. 6, pp. 1–27, 2021. J. L. Lin, T. H. Wen, J. C. Hsieh, and P. C. Chang, “Density-based microaggregation for statistical disclosure control,” IEEE Trans. Knowl. Data Eng., vol. 14, no. 1, pp. 189–201, Jan./Feb. 2002. J. M. Kabir, A. N. Mahmood, H. Wang, and A. K. Mustafa, “Microaggre- gation sorting framework for k-anonymity statistical disclosure control in cloud computing,” IEEE Trans. Cloud Comput., vol. 8, no. 2, pp. 408–417, Apr.-Jun. 2020. F. Casino, J. Domingo-Ferrer, C. Patsakis, D. Puig, and A. Solanaz, “A k-anonymous approach to privacy preserving collaborative filtering,” J. Comput. Syst. Sci., vol. 81, no. 6, pp. 1000–1011, Sep. 2015. —, “Privacy preserving collaborative filtering with k-anonymity through microaggregation,” in Proc. 2013 IEEE 10th Int. Conf. e-Bus. Eng. (ICEBE 2013), Coventry, UK, Sep. 2013, pp. 490–497. F. Casino, C. Patsakis, and A. Solanas, “Privacy-preserving collaborative filtering: A new approach based on variable-group-size microaggregation,” Electron. Commerce Res. Appl., vol. 38, no. 1, pp. 1–12, Nov./Dec 2019. D. Rebollo-Monedero, J. Forné, and M. Soriano, “An algorithm for k-anonymous microaggregation and clustering inspired by the design of distortion-optimized quantizers,” Data Knowl. Knowl. Eng., vol. 70, no. 10, pp. 892–921, Oct. 2011. M. E. Kabir, A. N. Mahmoud, H. Wang, and A. K. Mustafa, “Microaggre- gation sorting framework for k-anonymity statistical disclosure control in cloud computing,” IEEE Trans. Cloud Comput., vol. 8, no. 2, pp. 408–417, Apr.-Jun. 2020. M. Laszlo and S. Mukherjee, “Minimum spanning tree partitioning algo- rithm for microaggregation,” IEEE Trans. Knowl. Data Eng., vol.17, no. 7, pp. 902–911, Jul. 2005. C.-C. Chang, Y.-C. Li, and W.-H. Huang, “TFRP: An efficient microaggregation algorithm for statistical disclosure control,” J. Syst. Softw., vol. 80, no. 11, pp. 1866–1878, Nov. 2007. J. Domingo-Ferrer, F. Sebé, and A. Solanas, “A polynomial-time approx- imation to optimal multivariate microaggregation,” Comput. Math. with Appl., vol. 55, no. 4, pp. 714–732, Feb. 2008. B. Abidi, S. Ben Yahia, and C. Perera, “Hybrid microaggregation for privacy preserving data mining,” J. Ambient Intell. Humaniz. Comput., vol. 11, no. 1, pp. 23–38, Jan. 2020. P. Samarati, “Protecting respondents’ identities in microdata release,” IEEE Trans. Knowl. Data Eng., vol. 13, no. 6, pp. 1010–1027, Nov./Dec 2001. J. Domingo-Ferrer, A. Martínez-Ballesté, J. M. Mateo-Sanz, and F. Sebé, “Efficient multivariate data-oriented microaggregation,” VLDJ J., vol. 15, no. 4, pp. 355–369, Aug. 2006. D. Rebollo-Monedero, J. Forné, M. Soriano, and J. Puiggali Allepuz, “k-anonymous microaggregation with preservation of statistical dependence,” Inf. Sci., vol. 342, no. 1, pp. 1–23, May 2016. T. Jolliffe, Principal Component Analysis, 1st ed. New York, NY, USA: Springer, 1986. N. Halko, P. G. Martinsson, and J. A. Tropp, “Finding structure with randomness: Stochastic algorithms for constructing approximate matrix decompositions,” California Institute of Technology, Pasadena, CA, USA, Tech. Rep. 2009-05, May 2009. W. Wang, Y. Huang, Y. Wang, and L. Wang, “Generalized autoencoder: A neural network framework for dimensionality reduction,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Columbus, OH, USA, Jun. 2014, pp. 490–497. Y. Wang, H. Yao, and S. Zhao, “Auto-encoder based dimensionality reduction,” Neurocomputing, vol. 184, no. 1, pp. 232–242, Apr. 2016. S. Lloyd, “Least squares quantization in PCM,” IEEE Trans. Inf. Theory, vol. 28, no. 2, pp. 129–137, Mar. 1982. H. H. Bock, “Clustering methods: A history of k-means algorithms,” in Selected Contributions in Data Analysis and Classification, 1st ed., P. Brito, G. Cucumel, P. Bertrand, and F. de Carvalho, Eds. Springer, 2007, pp. 161–172. A. A.rusha and G. Yi, “Comparative analysis review of pioneering DBSCAN and successive density-based clustering algorithms,” IEEE Access, vol. 9, pp. 87918–87 935, Jun. 2021. M. D. Nguyen and W.-Y. Shin, “An improved density-based approach to spatio-textual clustering on social media,” IEEE Access, vol. 7, pp. 27 217– 27 230, Mar. 2019. A. Fujino, H. Isozaki, and J. Suzuki, “Multi-label text classification with model combination based on f1-score maximization,” in Proc. 3rd Int. Jt. Conf. Nat. Lang. Process.: Volume-II (IJCNLP 2008), Hyderabad, India, Jan. 2008, pp. 823–828. C. J. V. Rijsbergen, Information Retrieval, 2nd ed. London, UK: Butterworth, 1979. K. T. Chui, D. C. L. Fung, M. D. Lytras, and L. T. Miu, “Predicting at- risk university students in a virtual learning environment via a machine learning algorithm,” Comput. Human Behavior, vol. 107, pp. 1–7, Jun. 2020.

SOOBIN LEE received the B.S. degree in german language and literature and business administration from Yonsei University, Seoul, Republic of Korea, in 2019. She received the M.S. degree in management science from the same university in 2021. Since November 2021, she has been with Intelligence Labs of Nexon Korea, Seongnam-si Gyeonggi-do, Republic of Korea, as a data scientist. Her research interests include social network analysis, financial engineering, data mining, and machine learning.

WON-YONG SHIN (S’02–M’08–SM’16) received the B.S. degree in electrical engineering from Yonsei University, Seoul, Republic of Korea, in 2002. He received the M.S. and the Ph.D. degrees in electrical engineering and computer science from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea, in 2004 and 2008, respectively. In May 2009, Dr. Shin joined the School of Engineering and Applied Sciences, Harvard University, Cambridge, MA USA, as a Postdoctoral Fellow and was promoted to a Research Associate in October 2011. From March 2012 to February 2019, he was a Faculty Member (with tenure) of the Department of Computer Science and Engineering, Dankook University, Yongin, Republic of Korea. Since March 2019, he has been with the School of Mathematics and Computing (Computational Science and Engineering), Yonsei University, Seoul, Republic of Korea, where he is currently an Associate Professor. His research interests are in the areas of information theory, communications, mobile computing, data mining, and machine learning.

From 2014 to 2018, Dr. Shin served as an Associate Editor of the IIEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences. He has now served as an Organizing Committee member for the 2023 IEEE Consumer Communications & Networking Conference. He served as an Organizing Committee member for the 2015 IEEE Information Theory Workshop. He also served as a Technical Program Committee member or a Chair for various conferences. He received the Bronze Prize of the Samsung Humantech Paper Contest (2008), the KICS Haedong Young Scholar Award (2016), and the ICT Express Best Guest Editor Award (2021).