A Data-Centric ℓ-Diversity Model for Securely Publishing Personal Data with Enhanced Utility

Abdul Majeed, *Member, IEEE,* and Seong Oun Hwang, *Senior Member, IEEE*

Abstract—In this paper, we propose and implement a novel anonymization model, called data-centric ℓ-diversity, to effectively safeguard the privacy of individuals with considerably enhanced utility in data publishing scenarios. Through experimental analysis of real-life datasets, we found that when the data quality is poor (e.g., distributions are uneven), most of the existing methods only anonymize some parts of the data (where distributions are balanced) and leave other parts unprocessed, which can lead to explicit privacy disclosures. Furthermore, they do not identify and repair problematic parts of the data before anonymization, and therefore, they are not secure from the threat of privacy breaches. To address these technical problems, in this paper, we implement an automated method that identifies vulnerabilities in the underlying data to be anonymized w.r.t. distribution, and that repairs them by injecting virtual samples of good quality. Later, we implement a data partitioning strategy that creates compact and diverse classes of size k, where k is the privacy parameter. Finally, only shallow generalization (or no generalization) is applied to each class to minimally generalize the data, whereas existing methods overly distort data by not improving the quality beforehand, which can lead to poor utility in data-driven services. We conducted detailed experiments on four datasets to justify the performance of our model in realistic scenarios, and achieved promising results from the perspectives of boosted accuracy, privacy preservation, data utility enrichment, and reduced computing overheads. Compared with baseline methods, our model enhanced privacy preservation by 36.56% on three different metrics, and data utility was augmented with 18.65% less information loss and 14.37% greater accuracy. Lastly, our model, on average, has shown a 26.13% reduction in time overheads on four datasets compared to the SOTA baseline methods.

Index Terms—personal data, anonymization, ℓ-diversity, privacy disclosures, data publishing, data quality, generalization, utility

1 Introduction

Outsourced personal data from hospitals, banks, etc., enclose valuable knowledge that can augment the performance of data-driven services to improve the quality of people's lives. For example, in healthcare, outsourced data can improve the quality of medical services and can foster the discovery of new treatment methods [1]. However, personal data enclose sensitive information, and privacy issues can arise after release [2]. To prevent privacy breaches, the data are usually anonymized before release. However, when the quality is poor (e.g., a big gap exists in values; distributions are uneven), most privacy models cannot be applied directly, and if applied, only some parts of the data can be anonymized (where distributions are balanced) and other parts are left as they are, possibly leading to explicit privacy disclosures. Furthermore, poor-quality data can lead to unfair decisions or wrong conclusions when analyzed after release [3]. To address these issues, many innovative methods (data augmentation, information gap filling, etc.) have been introduced to increase quality. Data enhancement has become a new profession, and it is vital for resolving low accuracy, avoiding unfair decisions, and mitigating drift issues in data-driven services [4].

1.1 Background and Context

The syntactic methods such as k-anonymity, ℓ-diversity, etc. have specific parameters to be enforced on the dataset to be anonymized. For example, k-anonymity preserves at least k-user with identical values of non-sensitive attributes in each group/class while anonymizing datasets. The ℓ-diversity ensures that each class should have at least ℓ-diverse values of the SA. We found that in some cases, constraints like ℓ diverse values in each group cannot be enforced, particularly when data to anonymize is low quality (or imbalanced). We report a similar scenario in Table 1, where the real data of ten users need to be anonymized with ℓ-diversity model with $\ell = 2$. The disease column is SA and the main target of the ℓ-diversity model. Let's say $k = 2$ (we need to keep two users in each class), only two classes can fulfill the criteria of ℓ-diversity at $\ell = 2$ (as shown in Table 2), where three classes will have no diversity in disease/SA column, indicating a failure of ℓ-diversity model in practice. In some cases, the imbalance can be very high as indicated in Sec. 1.2 in two real-world datasets and a considerable number of classes cannot have any diversity, which can risk the users' SA or leave some parts of data unprocessed which is not acceptable in real-world cases. One of the

TABLE 1
Overview of tabular/relational data to be anonymized with ℓ-diversity.

ID	Name	Age	Sex	Citizenship	Race	Height	Disease
u_1	Chan	45	M	USA	Black	6	Cancer
u_2	Levin	32	F	England	White	5	Flu
u_3	Jae	35	F	England	White	6	Cancer
u_4	Kuran	43	M	USA	White	5	Flu
u_5	Jaeni	55	F	Italy	White	6	Flu
u_6	Dao	46	F	France	White	6	Flu
u_7	Hall	69	M	Spain	White	7	Flu
u_8	Zhu	39	M	Australia	Black	6	Flu
u_9	Ming	89	F	China	White	7	Flu
u_{10}	Zheng	80	F	China	White	5	Flu

• *Abdul Majeed and Seong Oun Hwang are with the Department of Computer Engineering, Gachon University, Korea.*
E-mails: {ab09,shwang}@gachon.ac.kr

Manuscript received April 19, 2005; revised August 26, 2015.

TABLE 2
Overview of tabular data which meets ℓ-diversity criteria at $\ell = 2$.

ID	Name	Age	Sex	Citizenship	Race	Height	Disease
u_1	Chan	45	M	USA	Black	6	Cancer
u_4	Kuran	43	M	USA	White	5	Flu
u_2	Levin	32	F	England	White	5	Flu
u_3	Jae	35	F	England	White	6	Cancer

main consequences of a higher imbalance in SA is 100% disclosure in some cases. For example, if the 2-anonymity (i.e., $k = 2$) is applied to the last two rows of Table 1, we can get the anonymized data (2 records) like <80-90, F, East Asia, White, 5-7, Flu>. Now if the attacker has some background information about either Ming/Zhang, he/she can uncover the disease information easily as the SA value is the same for both Ming and Zhang in the class/group. One solution is to overly generalize (e.g., put '*') the attributes other than the disease in all classes that have no diversity, but it will destroy the data utility, seriously impacting the objectives of data sharing. In this context, we need some practical approaches to analyze the characteristics of data to repair problematic parts of data before applying anonymity models to preserve privacy while guaranteeing utility from the published data, which is the main idea of our research. Recently, data-centric approaches have gained popularity, and we intend to harness the potential of such approaches in rectifying the established privacy models like ℓ-diversity in order to accomplish the objectives of data sharing.

Thanks to rapid developments in AI methods, we can deal with such a distribution-related vulnerability by creating virtual records that are very close to real records. We introduced this concept in the anonymization process while keeping the privacy model in a loop, and improved various critical problems in the ℓ-diversity model. It is worth noting that t-closeness (an advanced form of ℓ-diversity) can overcome the deficiencies of ℓ-diversity, but if one value's representation is extremely low, such as 4.5% in the stroke dataset from Kaggle, t-closeness cannot provide strong privacy guarantees. In addition, both models and their ramifications mostly ignore the homophily between records w.r.t. quasi-identifiers (QIDs) while establishing classes, leading to lower truthfulness in the anonymized data. Data with lower truthfulness has some discrepancies between statistical results and real results. For example, in medical data, the correlation between disease and its symptoms needs to be accurately preserved in anonymized data. However, the application of anonymity changes the correlation in some cases, and the true knowledge either becomes hidden or fake knowledge emerges in anonymized data that does not exist in reality, leading to data with lower truthfulness.

1.2 Motivation behind this research

Andrew Ng introduced the concept of data-centric artificial intelligence (DC-AI) to develop transformative AI systems [5]. Based on this idea, a substantial number of developments have been made recently, and promising results have been achieved in domains involving lower- or poor-quality data. DC-AI-based developments can augment accuracy by up to 40%, and can significantly lower computing overhead [1]. Hegde [6] obtained 100% accuracy in anomaly detection scenarios by using time series data. Other promising applications of these concepts are training AI models with reduced computation complexity while sustaining similar accuracy [7], and training them with the least amount of data possible [8]. Data-centric practices significantly contribute to the customization of language models as well[2]. Based on these breakthroughs, it is fair to say that data-centric efforts are vital to advancing major technologies.

Motivated by the above developments in other domains, we explored the role of data-centric practices for the first time in the information privacy domain. We selected and analyzed some benchmark datasets from the University of California, Irvine (UCI), and Kaggle repositories that have been widely used to test the performance of privacy methods. We found various kinds of vulnerabilities in these datasets. Some can be fixed during pre-processing, but some cannot be fixed and can seriously impact the application of privacy models such as k-anonymity [9], ℓ-diversity [10], t-closeness [11], differential privacy (DP) [12], and their improved versions. We noticed a higher imbalance in sensitive attribute (SA) values in these datasets (the Adult dataset from UCI and the Stroke Prediction dataset from Kaggle), and many records can be at risk of disclosure. In these respective datasets, 38.14% and 90.25% of the records are at risk of disclosure owing to the distribution-related vulnerability. In these datasets, the cardinality of the SA column is two, which means that the maximum value of ℓ that can be applied to them is 2. The Adult dataset has two values of income under the SA column: $<= 50K$ and $> 50K$. The users having the first value are 24,720 and 7,841 users own the second value. Now, if the ℓ-diversity model is used to anonymize this dataset, only 7,841 classes out of 16,280 can be created with two diverse values while the rest 8,439 classes will have a single value of SA (e.g., no diversity). In this example, 16,280 stands for the total number of equivalence classes (or clusters), and this statistic is taken after the ℓ-diversity model at $\ell = 2$. In the second dataset, the count of people who had a stroke is just 249 while people who don't is 4,861. In this case, only 249 diverse classes out of 2,431 can be created with $\ell = 2$, and $\sim$ 2,182 classes can have zero diversity. When a lot of classes have no/less diversity, the privacy of some users can be easily leaked owing to background knowledge of attackers or linking attacks among auxiliary and published data. We believe that this situation is common in many real-world datasets when SAs have binary values. Specifically, we analyzed the ℓ-diversity [10] that explores diverse values in SA columns of the data. However, if the respective column does not encompass very many diverse values, the ℓ-diversity [10] can face two technical problems:

1) the model cannot be directly applied if there is lower equity in terms of SA value distribution (e.g., the distribution is highly uneven) in the data, and
2) if it is applied to poor-quality data, the SAs of users will be exposed to adversaries due to a lack of heterogeneity in the SA column of the anonymized data. Also, there is a risk that the ℓ-diversity model will not process many of the records owing to the lack of diversity.

1. https://landing.ai/data-centric-ai/

2. https://www.mosaicml.com/blog/introducing-pubmed-gpt

1.3 Contributions

This work's main contributions are listed as follows.

- A generic anonymity model, named data-centric ℓ-diversity, is proposed to protect personal data from contemporary privacy threats with significantly enhanced utility by extending the original ℓ-diversity model to operate on low-quality datasets. Specifically, we uncover a scenario where a naive ℓ-diversity model cannot be straightforwardly applied due to multiple vulnerabilities in the data, and we solve this crucial problem by identifying and addressing vulnerabilities in the data before anonymization.
- A synthetic data generation method is implemented to curate virtual records of high quality to increase the frequency of infrequent SA values in real data so that ℓ-diversity criteria can be consistently enforced in all classes whereas the original ℓ-diversity model and its ramifications often ignore this aspect, leading to two technical problems as cited in Sec. 1.2.
- A promising data partitioning strategy is developed that exploits homophily and diversity among records while assigning them to classes, satisfying k-anonymity and ℓ-diversity criteria. The classes established by our strategy are highly compact, diverse, and balanced w.r.t. k, compared to existing methods.
- A shallow anonymization method is developed to lessen anonymity in the data to preserve maximal knowledge in anonymized data. We introduced and implemented various refinements to restrict unnecessary changes by extracting common patterns that can be released (as is) due to fewer privacy risks [13]. We introduced lower-level anonymization for most parts of the data, which offers higher privacy guarantees and retains the same semantics as in real data.
- We performed thorough experiments on four complex benchmark datasets using various metrics, and compared the performance with eight state-of-the-art (SOTA) methods. Through experiments and comparisons, our model significantly outperformed baseline SOTA methods w.r.t. privacy and utility.
- To the best of our knowledge, this is the pioneering work in the syntactic privacy methods category that identifies and repairs vulnerabilities in data before anonymization to resolve the utility and privacy trade-off and rectify a popular privacy model.

The rest of this paper is organized as follows. Section 2 presents the background and SOTA methods used in data publishing. Section 3 presents the system model and formulates the problem. The proposed data-centric ℓ-diversity model and its main modules are presented in Section 4. Section 5 provides the privacy analysis of our model. Section 6 presents the results obtained from intricate experimentation with four benchmark datasets, and comparisons with SOTA methods. We wrap-up the paper in Section 7.

2 Background and Related Work

2.1 Privacy-preserving data publishing (PPDP)

PPDP is a process of releasing personal data to third parties for conducting analytics. The data release is vital for finding answers to research questions and advancing science. In the past decade, many anonymization methods have been proposed for making data more broadly available while addressing privacy issues [14]. Most of the anonymization methods change the structure of the data to protect privacy while permitting analysis of various kinds [15]. An overview of data in original form as well as anonymized form (with $k = 2$ and $\ell = 2$) is shown in Tables 1 and 2. From Table 2, we can see that the anonymized data are less specific but semantically consistent with the original data. Unfortunately, existing methods distort data too much, and many improvements have been introduced to prevent dilution of values [16]. To pinpoint existing developments, we classify them into five major categories: syntactic, semantic, clustering, AI-powered, and hybrid methods.

Syntactic methods: These methods partition data into classes, and generalization/suppression operations convert raw data into anonymized form. The famous syntactic methods is k-anonymity [9], and its extended versions (i.e., ℓ-diversity [10], t-closeness [11], etc.). These methods have been widely used in commercial and industrial domains owing to lower complexity and conceptual simplicity.

Semantic methods: Data are anonymized by using randomization operations and noise addition. DP [12] and its enhancements are among the semantic methods. In DP model, a randomized function ($\mathcal{F}$) ensures ϵ-DP if $\forall$ real datasets D_1 and D_2 differ by at most one tuple, and $\forall$ $S \subseteq Range(\mathcal{F})$,

$$Pr[\mathcal{F}(D_1) \in S] \leq exp(\epsilon) \times Pr[\mathcal{F}(D_2) \in S] \quad (1)$$

DP can be satisfied via Laplace (numerical data) and exponential (categorical data) mechanisms. DP has become a SOTA model for privacy preservation in many scenarios. However, DP is unable to effectively resolve the trade-off between utility and privacy, and selecting the optimal ϵ value is very difficult.

Clustering methods: These methods mostly employ the same parameters as syntactic methods (k, ℓ, etc.), but some other measures (Euclidian distance, similarity, groupings, etc.) are also used to augment utility. Recently, many clustering-based methods have been developed to address the flaws of syntactic methods, especially data utility [17], [18]. However, these methods have higher computing costs, and often ignore diversity in SA values, leading to explicit SA disclosure.

AI-powered methods: Besides the traditional methods, many AI-powered methods have been developed for PPDP. These methods can revamp various key parts of privacy methods [19]. We recently proposed an ML-based method to anonymize data with improved privacy and utility [20]. These methods can be used to select QIDs, to reduce dimensions in the data, and to find optimal partitions.

Hybrid methods: In some cases, more than one method is used to anonymize data. For example, k-anonymity and DP were combined to increase the utility of anonymized data [21]. Another study combined clustering with cryptography to boost utility in anonymized data [22]. Recently, a hybrid approach amalgamating a GAN with DP was proposed to preserve the privacy of industrial IoT data [23]. However, only one aspect (privacy or utility) was mostly improved in these hybrid methods.

2.2 Analysis of the state-of-the-art PPDP approaches

ℓ-diversity is the SOTA model that ensures each class must have a diverse value in an SA column to protect privacy. Various enhancements of this model have been proposed in the literature. Recently, a distributed version of this model was proposed using Apache Spark to reduce information loss (ILoss) in PPDP [24]. Another variant of ℓ-diversity called effective ℓ-diversity was proposed to lower ILoss by grouping similar records into classes [25]. Recently, the ℓ-diversity model was enhanced by an attribute-focused anonymization (AFA) scheme that combines fixed interval and diversity information [26]. A highly scalable and DP-like technique, called MRMondrian was proposed to anonymize multidimensional data [27]. The proposed approach has abilities to anonymize data in parallel form. A DP-based method called LoPub was proposed to provide privacy guarantees in crowdsourced data [28]. The proposed approach synthesizes real data and produces new data that have higher closeness to the real data. An improved three-phase approach named distributed hierarchical k-means for satisfying ℓ-diversity (DHkmeans-LD) was devised to accomplish the PPDP task [29]. A highly scalable and distributed anonymization technique that extends Mondrian, called SDDA was proposed to anonymize large datasets [30]. The proposed approach satisfies the k-anonymity and ℓ-diversity properties. An obfuscation and generalization-based approach, called HyObscure was recently proposed to anonymize tabular data while ensuring defense against inference attacks [31]. A divid-and-conquer-based approach, called DCA was recently proposed to anonymize high-dimensional dataset [32]. We affirm the contributions of each method; however, there are four major limitations to the above-cited SOTA algorithms.

1) Most SOTA methods do not explore ways to improve data quality beforehand, and therefore, they cannot be employed in realistic environments when privacy and utility requirements are relatively high.
2) Most SOTA methods do not identify common patterns from QIDs to limit heavier changes in anonymization. The highly distorted data can change the semantics/conclusions, and some minor values in the data are at higher risk of dilution/erasure during the anonymization process.
3) Most SOTA methods optimize one metric at the expense of another. They can optimize either privacy or utility. Also, when data have a skewed distribution, many records are left unprocessed, leading to a higher risk of privacy breaches.
4) Most SOTA methods cannot maintain an equilibrium between privacy and utility when the underlying data to be anonymized have poor quality w.r.t. SA distributions or big gaps in attribute values.

Our model solves the above-cited limitations in the prior methods by improving data quality as well as by introducing various optimizations in the anonymization process.

3 System Model and Problem Overview

3.1 System model

Our system model is given in Figure 1, where four actors (record owners, data holders, database publishers, and analysts) constitute a data publishing scenario. Record owners

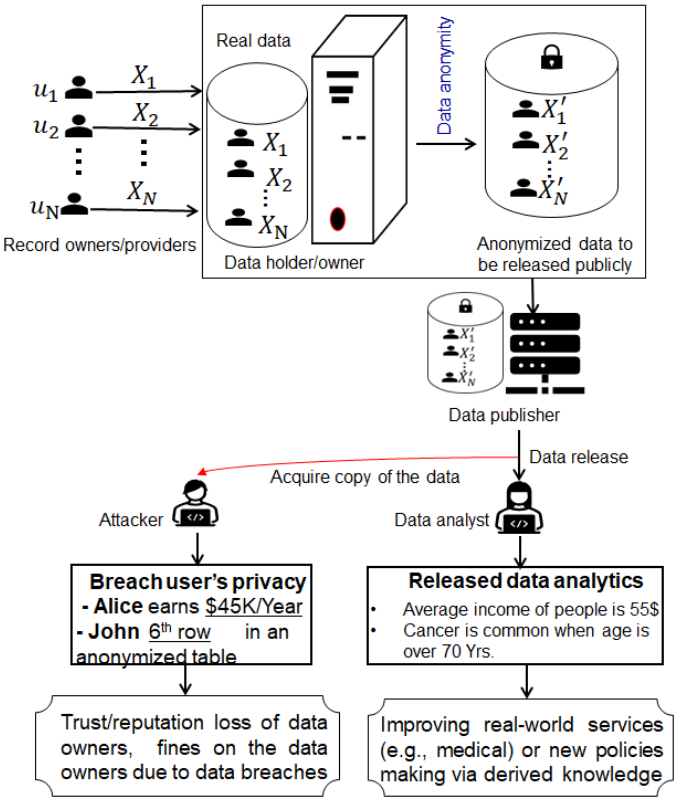


Fig. 1. Illustration of the system model of this paper.

offer their information to database owners. Subsequently, data holders aggregate the data from all N record owners. Later, the data in anonymized form are released to third parties or into the public domain for conducting analytics. The knowledge uncovered from the outsourced data is employed to enhance the performance of real-life services (e.g., healthcare). In this system model, few analysts might act as an adversary who intends to compromise the privacy of the data by linking it to auxiliary information available from external sources. Each record owner can have distinct (or the same) values for SAs w.r.t. other records in the data, and therefore, the distribution of SA values can be imbalanced if most records are from the same domain/community. We consider such a dataset of poor quality because it lacks diverse information. In this situation, the anonymity model can create many classes from this data that have the same SA for most of the records. The least diverse classes can pose a serious risk to privacy, and identity/SA disclosure can occur with high possibility. We aim to propose a secure model for PPDP that first repairs problematic parts of the data, and then applies the needed anonymity to effectively preserve privacy but provide enhanced utility.

3.2 Threat model

We assume most actors in our system model are honest, and carry out forthright actions. However, there is a chance that some data miners can act as an adversary and can compromise privacy. Our model removes all kinds of information that enables direct identification, QIDs can still be gathered from external sources (e.g., voter lists and online repositories) and can be matched to figure out people uniquely [33].

Hence, our model is susceptible to identity and associated SA disclosures in the following two ways.

1) Adversaries can have partial access to someone's QIDs already and might try to infer the rest of them. For instance, an adversary who has access to education and age might attempt to reveal ethnic race, leading to a unique identification of someone.
2) Adversaries may have access in advance to full records (i.e., entire QIDs) for a target person, and might also know with confidence that his/her information is contained in the data. Based on this information, the adversary attempts to extract the SAs (salary/diseases-contracted) of a target person.

Threat 1) is related to identity disclosure which assists the adversary in singling out the target person from released data based on background knowledge (BK) while threat 2) is related to SA disclosure/inference based on identity information. There exists a strong relationship between these two threats as identity disclosure often leads to SA inference/disclosure [34]. Considering the above threats, we aim to protect privacy against these two privacy threats that can stem from PPDP even if the data are anonymized. Both identity disclosure and SA disclosure are considered in experiments akin to most anonymity algorithms. The proposed model safeguarded the identity and SA revelation through the (k, ℓ) concept. The addition of synthetic records enhances the quantity of data and diversity in the SA column and contributes to lowering these threats. In addition, our model can limit other attacks on anonymized data, such as data reconstruction and dominant SA value exposure.

3.3 The privacy model and design goals

Our model uses generalization hierarchies to anonymize data. An overview of the privacy model, including its hierarchy, is in Figure 2. There are five levels in a hierarchy, denoted with $l_0 \sim l_4$. The selection of the optimal level from the hierarchy is desirable to balance privacy and utility.

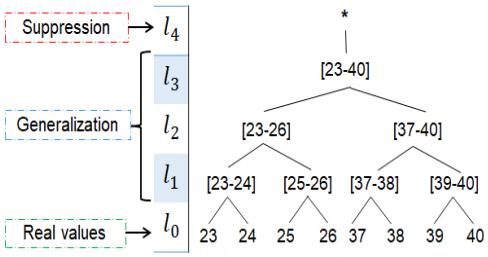


Fig. 2. The privacy model's domain generalization hierarchy for age.

The lower levels in the hierarchy are better for utility enhancement, whereas higher levels provide more privacy. Our model performs either lower-level generalization or no generalization by exploiting SA diversity and common pattern information to augment utility without leaking privacy. This paper aims to accomplish three major goals in PPDP.

Data Quality: Problematic parts of D are repaired beforehand in an automated way. The improved D has balanced distributions w.r.t. SA values, and can no longer pose serious threats to anonymization models' applicability and the conclusions drawn from $D^{'}$.

Utility: $D^{'}$ preserves sufficient knowledge to enhance real-world services. Particularly, it boosts accuracy and decreases information loss.

Privacy: Adversary possessing a large amount of external information cannot link/match a target user and figure out his/her SAs with confidence. Our model ensures that when an adversary attempts to link a record in $D^{'}$ by leveraging external data, a particular record matches diverse SA values. In short, it prevents inferring any one person's SAs from $D^{'}$.

3.4 Problem formulation

The key problem we pursue to address with our data-centric ℓ-diversity model is formally explained in Problem 1.

Problem 1. *We are given a real-world dataset, D, that includes various attributes (name, race, income, age, sex, diseases, residence type, etc.), where there can be multiple vulnerabilities in D w.r.t. data quality, such as fewer records, gaps in attribute values, outliers, missing values, duplicate records, imbalanced distributions in SA values, etc. How do we construct anonymized dataset $D^{'}$ where (a)$D \subseteq D^{'}$, (b) $D^{'}$ is k-anonymous, (c) $D^{'}$ is ℓ-diverse, and (d) $D^{'}$ offers superb quality (a.k.a. utility) for conducting mining/analytics (i.e., has low ILoss and significantly high accuracy)?*

4 Proposed Data-centric ℓ-diversity model

This section discusses the workflow of the proposed data-centric ℓ-diversity model in detail. We propose this model to anonymize low-quality datasets to provide strict privacy guarantees along with augmented utility. Figure 3 presents the conceptual overview of the proposed model. Table 3 describes the main notations used in our model.

TABLE 3
Key notations used in proposed data-centric ℓ-diversity model.

Symbols	Concise description
N	# of users/individuals in D, where $N = \|D\|$
$D, D^{'}$	Original data, anonymized/sanitized data
X_i, n	Data of the ith record, # of attributes in D
Q, Y	Set of QIDs, set of SA values
X_i^j, C	jth element of record X_i, Set of equivalence classes
$Sm(X_i, X_j)$	Similarity between two users (e.g., i and j)
ℓ, k	Diversity parameter, privacy parameter

There are two main modules in our proposed model.

- **Data quality enhancement:** In this module, the quality of D is significantly enhanced with the help of multiple operations listed in Module I of Figure 3. Specifically, we apply sophisticated five-step preprocessing to D to fix the basic vulnerabilities. Afterward, more rigorous steps are applied to identify the advanced vulnerabilities in D and to repair problematic parts of the data to produce a high-quality D for anonymization. In this work, the data with vulnerabilities is used interchangeably with low-quality data. For instance, if there are some vulnerabilities either basic or advanced in the data, then the respective data is regarded as low quality. Technical details of this module are given in Section 4.1.

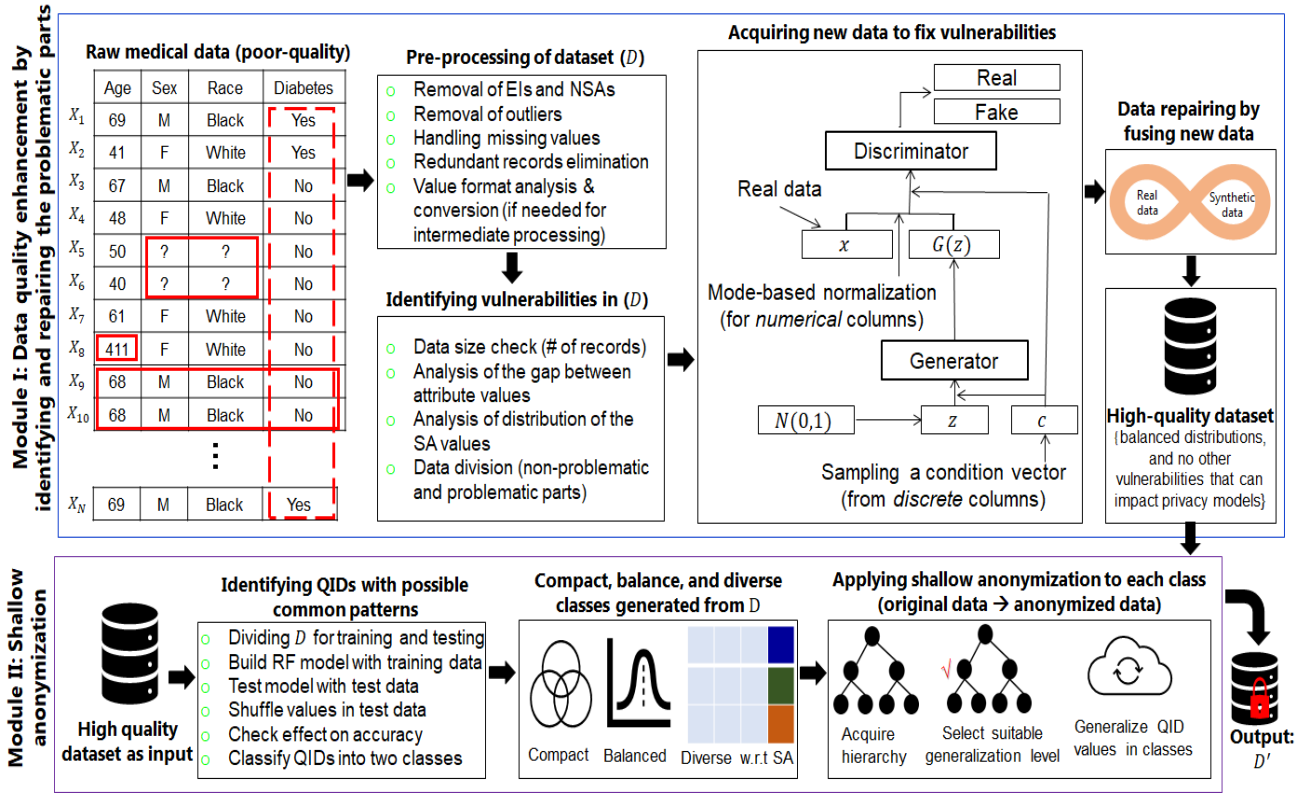


Fig. 3. Conceptual overview of the proposed data-centric ℓ-diversity model.

- **Shallow anonymization:** In this module, only shallow anonymization of D is performed (e.g., only a few small parts are anonymized), whereas existing methods mostly anonymize entire parts of D. In this module, we introduce various optimizations (e.g., identifying QIDs in which common patterns exist; creating balanced, compact, and diverse classes; performing lower-level generalization, or no generalization) as shown in Module II of Figure 3 to address the shortcomings of previous methods. Technical details of this module are in Section 4.2.

4.1 Quality Enhancement of D

4.1.1 The data model

Let $D = \{X_1, X_2, \ldots, X_N\}$ denote the collected dataset from the relevant users where X_i represents the entire ith record. We consider a scenario in which D encompasses n attributes that are represented as $A = \{a_1, a_2, \ldots, a_n\}$. Records in D are modeled $X_i = [x_i^1, x_i^2, \ldots, x_i^n]$, x_i^j, where x_i^j $(j = 1, 2, \ldots, n)$ represents the jth attribute of the ith record. For every attribute, we model $\omega_j = \{\lambda_j^1, \lambda_j^2, \ldots, \lambda_j^{|\omega_j|}\}$ as the domain of that particular attribute (i.e., A_j), where ω_j and $|\omega_j|$ denote the attribute value and its cardinality, respectively. In simple terms, D can be viewed as a matrix of $N \times n$ dimensions, where N shows the # of rows, and n represents the # of columns. The structure of D with some values is shown in Fig. 3.

4.1.2 Pre-processing D (fixing basic vulnerabilities)

In any real-world D, there can be multiple vulnerabilities that can impact the analyses, and therefore, pre-processing is applied to clean D. Usually, D can encompass four distinct types of attributes: QIDs, non-sensitive attributes (NSAs), SAs, and explicit identifiers (EIs). Depending upon the situation, some types of attributes (e.g., NSAs, EIs) may/may not be encompassed in D. After analyzing attributes' types, pre-processing is carried out on D. We propose and apply a number of techniques to fix common vulnerabilities. Table 4 demonstrates the pre-processing techniques in our proposed model. The structure of D after executing the steps in Table 4 becomes $D\{Q, Y\}$. In set Q, each QID has a multiple values and a label. For instance, q_1 can represent the country, and each record can have the country value (identical (or non-identical)) in the corresponding cell. In contrast, y_i represents the SA value in the respective record. A simplified layout for all tuples in D is expressed in Eq. 2:

$$T_{X,A} = \begin{pmatrix} X_i & q_1 & q_2 \cdots & q_p & Y \\ X_1 & x_1^1 & x_1^2 \cdots & x_1^p & y_1 \\ X_2 & x_2^1 & x_2^2 \cdots & x_2^p & y_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ X_N & x_N^1 & x_N^2 \cdots & x_N^p & y_1 \end{pmatrix} = \begin{pmatrix} X_i & q_1 = age & q_2 = gender & q_p = country & Y = income \\ 1 & 30 & M \cdots & Italy & \geq 50K \\ 2 & 35 & M \cdots & USA & < 50K \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 33,501 & 88 & F \cdots & Japan & \geq 50K \end{pmatrix} \quad (2)$$

After giving an understandable structure to D and fixing common vulnerabilities, D is further inspected for advanced vulnerabilities.

TABLE 4
Overview of pre-processing applied to D, particularly to fix basic vulnerabilities.

Step	Technique(s)	Objective(s)	Output
Attribute arrangement	Grouping QIDs	To yield simplified structure of D	D with well-separated QIDs and SAs
NSA and EI removal	Deletion from data	Limit identity disclosure	D with QIDs and SAs only
Outlier removal	$Min-max$ analysis, domain knowledge	Increase D's reliability w.r.t analytics	D with correct attribute values
Handling missing values	Imputation and removal	Increase D's quality w.r.t. analytics	D with complete information
Redundant record elimination	$if-else$ conditions, similarity checks	Reduction in computing overhead	D without duplicate records
Attribute type checks	$typeof(attributename)$ function	Prevent anonymity-model collapse	D without inconsistent information
Value formatting/enrichment	$< key, value >$ pairs, mapping	Simplify processing, reduce complexity	D with consistent information

4.1.3 Identifying advanced vulnerabilities

In the previous subsection, we implemented various ways to fix basic vulnerabilities in D. However, most of these vulnerabilities are common and impact privacy models minimally. However, in some cases, there are advanced vulnerabilities that can seriously impact privacy models (for instance, the distribution imbalance problem (DIP) in SA can impact ℓ-diversity criteria fulfillment). Algorithm 1 presents the method employed to find vulnerabilities in D w.r.t. SA distributions. In this algorithm $D, index[Y]$ is the input, and τ is the output. This algorithm finds unique values of SAs in their respective columns and computes their frequency utilizing all of D. Later, differences in the frequencies are computed to check for the distribution vulnerability in D. If there is no vulnerability in D, it is sent for anonymization.

Algorithm 1 Checking for the SA distribution vulnerability in D.

Require: $D, index[Y]$
Ensure: τ, where $\tau = 0/1$
$\tau \leftarrow 0$ ▷ Assuming no vulnerability in D.
$N \leftarrow |D|$ ▷ Determine data size.
$T_\tau \leftarrow {}^{N}/_{2}$
$Y_{uni} \leftarrow unique(index[Y])$ ▷ Identify unique SA values.
$\phi_1 \leftarrow Y_{uni}[0]$ ▷ Get first value of SA
$\phi_2 \leftarrow Y_{uni}[1]$ ▷ Get second value of SA
$f_{\phi_1} \leftarrow |D|$, where $index[Y] == Y_{uni}[0]$
$f_{\phi_2} \leftarrow |D|$, where $index[Y] == Y_{uni}[1]$
$d = |f_{\phi_1} - f_{\phi_2}|$
if $d > T_\tau$ **then**
    $\tau \leftarrow 1$
else if $d \leq T_\tau$ **then**
    $\tau \leftarrow 0$
Return τ

Algorithm 2 presents the method employed to classify data into problematic (faulty) and non-problematic (non-faulty) parts. In Algorithm 2, $D, index[Y], N, \phi_1, \phi_2, f_{\phi_1}, f_{\phi_2}$ constitute the input, and D_{pr} (faulty) and D_{npr} (non-faulty) are the output. This algorithm compares the frequencies of SAs and divides the data into two parts based on the f value. Later, D_{pr} is repaired with new data. In algorithms 1 and 2, the symbols f_{ϕ_1} and f_{ϕ_2} represent the frequency of major and minor SA values, respectively. $index[Y]$ denotes the colum in D which encompasses the SA values.

Algorithm 2 Dividing D into D_{pr} and D_{npr}.

Require: $D, index[Y], N, \phi_1, \phi_2, f_{\phi_1}, f_{\phi_2}$
Ensure: D_{pr}, D_{npr}
$D_{pr} \leftarrow \emptyset$
$D_{npr} \leftarrow \emptyset$
if $(f_{\phi_1} > f_{\phi_2})$ **then**
    for $i = 1$ *to* N **do**
        if $(\phi_1 == i(index[Y]))$ **then**
            $D = D - X_i$
            $T_{npr} \leftarrow T_{npr} \cup \{X_i\}$
        else if $(\phi_1 \neq i(index[Y]))$ **then**
            $D = D - X_i$
            $T_{pr} \leftarrow T_{pr} \cup \{X_i\}$
Return D_{pr}, D_{npr}

4.1.4 Acquiring new data to repair problems in D

After dividing D into D_{pr} and D_{npr}, D_{pr} is repaired using data augmentation. Specifically, we address the distribution imbalance problem (DIP) by curating more records of good quality, and it is one of the main contributions of this work. The main objective of getting more records is to solve DIP to ensure ℓ-diversity criteria satisfaction in most classes to reduce privacy risks. Also, the augmentation is only performed in minor classes and with good quality records, and therefore, the emergence of faked knowledge is well countered. The two crucial problems in ℓ-diversity (e.g., explicit privacy leakage from low diverse classes, and the possibility of processing only some parts of data where distributions are balanced) are successfully resolved with this step. To generate more samples of good quality, we used a conditional GAN (CGAN). The CGAN is an optimized implementation of a GAN that can generate an approximate representation of the data [35]. CGANs have been widely used for data augmentation and quality enhancement, specifically when the original data are not available on a large scale. A CGAN has a condition that prevents imbalanced learning (e.g., giving preference to the majority class only). The enforcement of a condition ensures that structural similarity between real and synthetic data remains high, and none of the unique values under categorical QIDs should lost during conversion. For instance, in the adult's dataset, the country (a categorical QID) has 41 different values, however, the representation of one value (e.g., USA) is very high (∼90%) compared to others. In contrast, the remaining 40 values have the very least representation compared to the USA, and they can be ignored (or minimally learned) when real data is converted to synthetic data by the CGAN

model. To prevent such things from happening, a condition is imposed on all categorical QIDs so that all values can get an equal chance of exploration in CGAN training and the count of unique values in categorical QIDs remains the same in real and synthetic data.

We used Xu eta al. [35] implementation as a baseline and further optimized it. Specifically, we used an improved implementation of a CGAN that correctly mirrors the properties of D, and prevents the vanishing gradient problem, model collapse, and imbalanced learning issues. Also, this implementation has the flexibility to generate needed samples only. In this method, D, Φ_G, and Φ_C (parameters of conditional G and critic C); batch size (m); and pac size (pac) are provided as input; modified parameters Φ_G and Φ_C, along with D_{new}, are gathered as output. In the intermediate steps, two neural networks are simultaneously trained, conditions are imposed on discrete columns (di), and loss metrics are updated accordingly. Finally, D_{new} with a higher structural similarity w.r.t. D is obtained. Although D_{new} has high structural similarity, some values in numerical columns are inconsistent (e.g., higher/lower than the real values). Therefore, we apply the bounding box technique to fix this issue, and D_{new} is made consistent for further processing. We pre-process D_{new} before fusing it with D_{pr}. We also analyze the distributions of newly generated data w.r.t. discrete columns. The impact of synthetic data on other anonymization techniques is given in **Appendix A**.

4.1.5 Good-quality D generation by fusing data

In this step, a good-quality D is generated by fusing three types of data: D_{pr}, D_{npr}, and D_{new}. Based on the experiments, we found that if the real data (D) is imbalanced, then the corresponding synthetic data (D_{new}) is also imbalanced w.r.t. SA values. However, since we require some records from D_{new} and particularly from the minor SA values to balance D, therefore, the imbalance in D_{new} has a negligible impact on our proposed model. The latest D from data fusion is free of most vulnerabilities. It has balanced distributions w.r.t. SA values, and does not pose threats to the applicability of privacy models. The concept employed in this work resembles the data augmentation concept (specifically, oversampling technique) used in machine learning, where the distribution of minor classes (minor SA value in our work) is enlarged until the number of samples in major and minor classes becomes equal. We adopt the same concept to balance the distribution of SA values in the SA column, particularly by increasing the # of samples in minor SA values. Once the distributions are balanced/equal in SA values of D, the diversity criteria are accomplished in all classes which is not the case with real imbalanced D.

The proposed model has been designed to accomplish the 2-diversity that scenario is commonly encountered. However, it is generic and can be extended to scenarios involving multiple SA values. To that end, the following modifications are needed: (i) imbalance needs to be checked between all SA values, (ii) more data needs to be curated to balance SA values, (iii) SA values and distribution require changes as per the ℓ value, and (iv) large ℓ value (e.g., $\ell > 2$) needs to set for better privacy protection.

4.2 Shallow Anonymization of D

After curating the high-quality D, we anonymize the data (i.e., real values $\rightarrow$ modified values) with the help of generalization hierarchies. We introduce three optimizations in the anonymization process to improve privacy and utility results. For the first time, we introduce the notions of no generalization (retaining common patterns as they are for informative analysis) and low generalization to yield a high-quality D' for information consumers.

4.2.1 Identification of QIDs with common patterns

As pointed out by a previous study in [3], there can be skewness in QIDs as well (i.e., one value can occupy more than 80% of D, and the remaining parts can be occupied by many other values). The frequently occurring value can be regarded as a common pattern, and can no longer pose a threat to privacy [13]. How to find QIDs from D that have such a pattern is challenging. In our recent work, we developed a new method to identify vulnerable QIDs using ML, specifically random forest [20]. Figure 4 presents the main idea employed to find the pattern-friendly (PF) and non-pattern-friendly QIDs. We adopt the same method

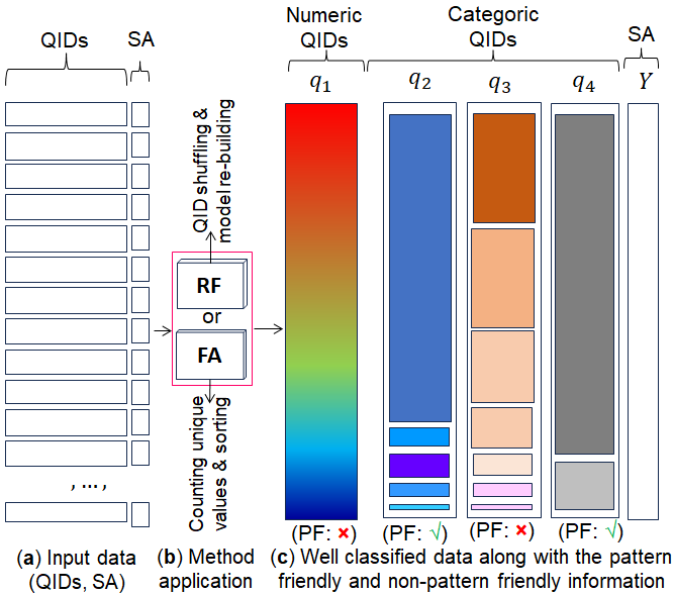


Fig. 4. Main idea employed to identify pattern-friendly QIDs from D.

[20] with slight modifications to identify the QIDs that have common patterns. In this model, training is performed once using training data, and testing is performed multiple times with shuffled columns (one at a time) to rank the QIDs. Specifically, if pattern-friendly QID values are shuffled there is the least impact on accuracy. Through experiments, we found that categorical QIDs have a much higher probability of encompassing common patterns compared to numerical QIDs. Hence, one can simply perform frequency analysis (FA) to find such QIDs, as shown in the middle block of Figure 4. We perform cross-validation to check the accuracy of the RF method. Our analysis confirmed the results. Specifically, we computed the weight (w) of each QID, and if the w value is lower, that QID is pattern-friendly and vice versa.

The information derived in this step is used in the anonymization process to prevent unnecessary changes. If a QID is PF, then one value may span across many classes/clusters, indicating a common pattern (population-level information) with the least probability of privacy breaches. Hence, such values can be skipped from the generalization process to boost data utility.

4.2.2 Forming compact, balanced, and diverse classes

We developed a strategy that exploits homophily w.r.t. QIDs and diversity w.r.t. SAs while compiling the classes. Furthermore, our strategy creates size-balanced classes with at least k records in each class. Specifically, each class possesses the following three key characteristics.

- *Compact:* Records in each class are highly similar based on QID values. Specifically, intra-class similarity is very high. Cosine similarity has been extensively employed in many real-world systems to estimate record-item, item-item, and record-record similarity [36]. Values between two records can be computed via Eq. 3:

$$Sm(X_1, X_2) = \frac{\sum_{i=1}^{p} X_{1_i} \times X_{2_i}}{\sqrt{\sum_{i=1}^{p} X_{1_i}^2} \times \sqrt{\sum_{i=1}^{p} X_{2_i}^2}} \quad (3)$$

  where i shows the QIDs of records X_1 and X_2, and p denotes the total number of QIDs.
- *Diverse:* The SA values in each class are diverse, meaning that no class has the same SA value for all records. The diversity, η, between records in class C_i can be computed using Eq. 4:

$$\eta(C_i) = -\sum_{i=1}^{|Y_{C_i}|} y_i log_2 y_i \quad (4)$$

  The value of η lies between 0 and 1, $\eta \in [0, 1]$. When $\eta = 0$, all records have the same value for an SA in C_i. In contrast, $\eta = 1$ is the ideal scenario due to higher diversity in SA values. Our model prevents η from being equal to 0, and therefore, there is no genuine risk to privacy. The implementation to prevent η from being equal to 0 is done by comparing the SA values of the cluster center and the records to be mapped in a cluster by using Eq. 4. If SA is the same between the cluster center and the record, then $log_2 y_i = log_2(1) = 0$, and then a new record with a dissimilar value than the cluster center is chosen to make $\eta(C_i) > 0$.
- *Balanced:* There are at least k records in each class, and the size of each class is no more than $2k$ in case of a residual record's addition. Also, the SA values are diverse. All three properties can be ensured using the following equation:

$$C_i = \left\{ \begin{array}{ll} |C_i|, & \text{for } k \leq |C_i| < 2k \\ \eta(C_i), & \eta(C_i) > 0 \\ Sm(C_i), & \text{for } (1\,to\,|C_i|) \sim 1 \end{array} \right\}$$

The pseudocode of the algorithm used to partition D into balanced classes is Algorithm 3. In this algorithm, D, k, N (dataset, privacy parameter, and number of records in D) are the input, and C (the set of classes) is returned as output. This algorithm ensures that all records in D are processed and that each class fulfills the three properties (i.e., compactness, diversity, and balancedness) cited above. In Algorithm 3, ℓ-diversity is performed when records are mapped to the relevant equivalence classes. The proposed model is (k, ℓ), meaning it simultaneously accomplishes both the k-anonymity and ℓ-diversity criteria. It fulfills the diversity criteria for more # of equivalence classes compared to the conventional (k, ℓ) methods. After obtaining set C, shallow anonymization is applied to each class to generalize QIDs. The complexity of algorithm 3 is $\mathcal{O}(C.(N + p).i) = \mathcal{O}(C.N.i)$, where C denotes data partition/clusters, N denotes total records in D whereas p (# of attributes/QIDs) has a constant upper bound, and i is the checks for cluster size w.r.t. k.

Algorithm 3 Partitioning D into balanced classes of size k.

Require: D, k, N
Ensure: C, where $C = \{C_1, C_2, \ldots, C_{nc}\}$
if ($N \leq$ privacy parameter k) **then**
  Return error, "fewer tuples than paramter$k$".
else if ($N > k$) **then**
  $C \leftarrow \emptyset$
  $nc \leftarrow {}^{N}/_{k}$ ▷ Determining total # of classes
  $X_j \leftarrow D - \{X_j\}$ ▷ Random selection of record X_j
  for $i = 1\,to\,nc$ **do**
    $C_i \leftarrow C_i \cup \{X_j\}$
    $X_{j+1} \leftarrow D - \{X_{j+1}\}$ ▷ Pick new record via $j + 1$
    $Sm(X_j, X_{j+1}) \leftarrow$ using Eq.3.
    $\eta(C_i)) \leftarrow$ using Eq.4.
    if ($Sm(X_j, X_{j+1}) > T_{Sm} \& \eta(C_i) > 0$) **then**
      $D \leftarrow D - (X_j, X_{j+1})$
      $C_i \leftarrow C_i \cup \{X_j, X_{j+1}\}$.
    else if ($Sm(X_j, X_{j+1}) \leq T_{Sm} \& \eta(C_i) = 0$) **then**
      $X_{j+2} \leftarrow D - \{X_{j+2}\}$ ▷ Pick new record via j
      While (no best match found) do
      Analyze X_{j+2} w.r.t Sm and η and add to C_i.
      Add X_j and X_{j+2} to C_i if criteria is met.
      End While
    While ($|C_i| < k$) do ▷ Check for balancedness.
    $X \leftarrow D - \{X\}$
    Repeat: steps 7 onward.
    $|C_i| == k$.
    $C \leftarrow C \cup \{C_i\}$
    End While
if ($|D| == 0$) **then** ▷ residual records' handling.
  Return success, "no residual record in D".
else if ($|D| \neq 0$) **then**
  Add residual records to C_i, where $0 \leq i < nc$ based on Sm and η.
  $C \leftarrow C \cup \{C_i\}$ ▷ Add updated classes in set C.
Return C

4.2.3 Applying shallow anonymization to convert D to D'

In the last step to convert D to D', shallow anonymization is applied to classes generated in the former steps. To preserve truthfulness in D', a careful approach is needed while converting QID values. In our model, similarity is ensured

between records, and therefore, truthfulness is maximally preserved in $D^{'}$. By introducing the no-generalization concept, some QID values that are PF will be released as they are. By performing lower-level generalization (or no generalization), our model can yield higher utility. In contrast, most existing methods use wide generalization intervals or symbols like $\leq$, $\geq$, $*$, etc., during anonymization, which can lead to poor utility. Our model injects external records of good quality, hence, the quality of $D^{'}$ is much better than from previous methods. Specifically, our model introduces two optimizations in the generalization process.

- *No generalization:* If a QID is PF, and all users in a class have the same value for that QID, then no changes are applied to QID values. It is worth noting that privacy is not leaked in this way because other QIDs in the records are generalized, and SAs have diverse values as well.
- *Lower level generalization:* Since most classes have a higher η w.r.t. SA values, most QID values can be generalized into lower levels of the generalization hierarchy to retain semantics as much as possible. It is worth noting that the ℓ value cannot exceed the # of unique SA values present in D, and if the distribution of all SA values is balanced, lower-level generalization is the most preferred solution. In our work, ℓ is small (e.g., $\ell = 2$), and distribution in SA values is balanced due to D_{new} addition, therefore, the lower level generalization does not compromise the desired privacy requirements. However, if ℓ is large enough (and the distribution of SA values is highly imbalanced) and we want very rigorous privacy protection, we have to choose high-level generalization.

The above two optimizations prevent greater changes in the anonymization process, leading to good quality $D^{'}$.

In conclusion, there are two main modules in our proposed model, data quality enhancement and shallow anonymization. The first module inspects data quality and improves it before anonymization so that required privacy parameters can be consistently enforced on all parts of data rather than a few good parts as most of the baselines do. The second module ensures that only parts of the data that are likely prone to privacy breaches are anonymized whereas most baselines anonymize every part of data, leading to higher changes in anonymized data and poor utility. This is the pioneering approach to rectify ℓ-diversity (a very popular privacy model) and its latest ramifications. The detailed comparison between our model and existing algorithms based on various features is given in **Appendix B**.

5 Privacy Analysis

This section formally proves and verifies that our proposed model satisfies ℓ-diversity by taking into account definitions of ℓ-diversity as well as its formal properties. First, we verify and prove the impact of virtual records created to satisfy the privacy criteria of ℓ-diversity without degrading D's quality.

Lemma 1. For the generated virtual records to be added to D to balance the distribution of least frequent SA values, contributes to enhancing utility in $D^{'}$ at the same time lowering the SA disclosure from $D^{'}$ by increasing the uncertainty for attackers in the SA column.

Proof. As discussed earlier, the representation of some values in the SA column can be very low, impacting the applicability of privacy models like ℓ-diversity. To overcome this issue, a bunch of virtual records was added to only rare SA values. These virtual records increase the frequency/representation of infrequent SA values, which can increase uncertainty for attackers in guessing the true SA value of target subjects, leading to better privacy protection. The SA inference/disclosure becomes $1/\ell$ from all QID groups/classes, which aligns with the criteria of ℓ-diversity. If $\ell = 2$, and D is imbalanced, there is a possibility that some QID groups can have one SA value (e.g., major), and the SA disclosure probability can be 1 (e.g., $1/\ell = 1/1 = 1$). To prevent this situation from happening and to keep SA disclosure less than $1/\ell$, we add synthetic records for the minor SA values. In our model, if $\ell = 2$, then the count of major and minor SA values are more balanced, so none of the groups can have one SA value thereby SA disclosure probability less/equal than $1/\ell$, leading to no loss of privacy guarantees. Since only a few records are added and the quality of virtual records is very close to records in real data, therefore, the possibility of fake knowledge induction in D is very low. Also, the newly added virtual records increase the possibility of general patterns in some QIDs, which can lower the generalization degree, leading to a better quality of $D^{'}$ for downstream tasks. By lowering the generalization, more knowledge is preserved in $D^{'}$, and knowledge disparity between D and $D^{'}$ is not lost. Based on the above analysis, it is fair to say that virtual records addition to D contributes to both privacy and utility enhancement in $D^{'}$. □

Theorem 1. *If $|Y| \geq \ell$, then the $D^{'}$ produced by the data-centric ℓ-diversity model satisfies ℓ-diversity across all QID groups.*

Proof. For any D, an anonymity model with two different privacy parameters k and ℓ is applied to create anonymized $D^{'}$, and it ensures that the entire $D^{'}$ satisfies ℓ-diversity considering the diversity requirement of SA values (e.g., at least ℓ distinct values in each QID group). The proposed model exploits the intrinsic characteristics of both QIDs (e.g., similarities) and SA (e.g., diversity) in D to create QID groups of size k. Since the proposed model increases diversity for least occurring values in the data refinement phase, each QID group has at least ℓ distinct values which prevent the SA disclosure even when the attacker has strong background knowledge or auxiliary information. The proposed model increases diversity by increasing the count of minor SA values via synthetic samples. The diversity is increased until the count of minor SA values becomes roughly equal to the count of major SA values (e.g., the frequency gap is minimal b/w major and minor SA values). In all QID groups, the proposed model ensures that the SA column has at least ℓ diverse values, and SA disclosure is $1/\ell$, which obeys the standard ℓ-diversity model. To accomplish ℓ-diversity in all QID groups, $|Y| \geq \ell$. Due to synthetically generated records, it is easier to append $\ell - 1$ records with dissimilar SA values in each QID group, thereby consistently ensuring $|Y| \geq \ell$. There are at least k records coupled with ℓ diverse values

of SA in each QID group in $D^{'}$, and therefore, identity or SA disclosure is effectively protected. Based on the above analysis, it can be concluded that our method satisfies ℓ-diversity criteria across all QID groups in $D^{'}$. □

Based on the aforementioned analysis and numerical results, it is fair to say that our model offers formal privacy guarantees in data publishing scenarios.

6 Experimental Evaluation and Analysis

To prove the efficacy and significance of our model, we performed extensive experiments on four real-life datasets. To benchmark our model, the results were also compared with eight SOTA and recent methods. In the next subsections, dataset details, the experimental setup, results for privacy and utility, and comparisons are presented in detail.

6.1 Dataset Description

In the experiments, we used a relational D encompassing both QIDs and SAs. The QIDs denoted identity-related information, such as age, sex, race, etc., of the record holder, whereas SAs denoted sensitive information like income. We performed rigorous experiments on the Adult, Stroke Prediction, Census-Income, and Diabetes 130-US hospitals datasets. The Adult dataset [37] includes both discrete and numerical types of QIDs, and the SAs are binary. This dataset has a substantial number of records and has been widely used in the anonymization literature to evaluate the strength of privacy models. The Stroke Prediction dataset [38] encompasses various demographics and clinical features used for predicting the onset of stroke. This dataset has various QIDs and a single SA. The Census-Income dataset [39] is obtained from the population surveys conducted by the U.S. Census Bureau. This dataset has mixed-type QIDs (i.e., discrete and numerical), and income is SA. This dataset is relatively bigger and has been widely used in testing the scalability and privacy strengths of anonymity models. Diabetes 130-US hospitals [40] is the clinical care data extracted from 130 US hospitals and delivery networks. It encompasses the clinical information and demographics. We utilized the demographics and some salient clinical features as QIDs, and DiabetesMed (diabetic medications prescribed with values: 'Yes' and 'No') was used as SA.

In all datasets, we ignored EIs and other NSAs before conducting the experiments. The numbers of QIDs used were seven, six, eight, and six from the Adult, Stroke Prediction, Census-Income, and Diabetes 130-US hospitals datasets, respectively. There were some vulnerabilities in four Ds, and therefore, pre-processing was applied to fix them. We present concise information on these datasets in Table 5. It is worth noting that all datasets were imbalanced, therefore, we present the details of the imbalance in SA values in the last column of Table 5.

The existing methods process a dataset as is, and there is a high chance of privacy breaches because many classes lack diversity, and minor values of some QIDs will be diluted or erased. In contrast, our model adds more records of good quality to D to balance the distributions. After augmentation, the number of records were $\approx$ 44K, 8.5K, 336K, and 140K for the adult, stroke, census, and diabetes datasets, respectively. The cardinalities for discrete columns were the same, and there was a slight change in the cardinality of the numerical QIDs. After augmentation, the distributions of the SA were more balanced than in the original D. The experimental analysis and distribution balance in D_{new} is given in **Appendix C.**

TABLE 5
Insight to datasets employed for performance evaluation.

Dataset	No. of records	QID name (Cardinality, Type, H levels)	SA label & counts
Adult [37]	32,561	Race (5, Discrete, 3) Age (74, Numerical, 7) Relationship (6,Discrete, 3) Gender (2, Discrete, 2) Work class (8,Discrete, 4) Education (17, Discrete, 5) Country (41, Discrete, 4)	Salary/Income ≤50K= 24,720 >50K=7,841
Stroke [38]	5,510	Gender (2, Discrete, 2) Age (82, Numerical, 7) M_status (2, Discrete, 2) Work_type (5, Discrete, 4) Residence_type (2, Discrete, 2) Smoking_status (4, Discrete, 3)	Stroke prob. 0= 5,261 1= 249
Census [39]	199,523	Sex (2, Discrete, 2) Age (91, Numerical, 7) Class of worker (9, Discrete, 4) Marital Status (7, Discrete, 4) Race (5, Discrete, 3) Education (16, Discrete, 3) Country_of_Origin (42, Discrete, 4) Working_hours (53, Numerical, 3)	Income levels 50000+= 12,382 -50000=187,141
Diabetes [40]	101,766	Race (5, Discrete, 3) Sex (2, Discrete, 2) Age (100, Numerical, 7) Time_in_hospital (14, Numerical, 4) #_lab_procedures (118, Numerical, 4) Insulin_dosage (4, Discrete, 3)	DiabetesMed No=23,403 Yes= 78,363

6.2 Implementation Setup

The experiments were carried out on a PC having an Intel Core i5-3320M CPU clocked at 2.60GHz and running Windows 10 Professional with 8GB RAM. The model was implemented using the Python language 3.9 (64-bit) version with built-in libraries support. A public implementation of CGAN with enhancement was used in generating a D_{new} of high quality. A customized RF implementation was used in identifying the QIDs having common patterns. There are four numerical (e.g., size of training data, size of testing data, # of the trees, # of QIDs required at the time of tree's split) and six non-numerical parameters (RF model type, splitting rule, QID's importance, whether to preserve forest, labels for QIDs, and SA label) that were used in RF model while identifying QIDs with common patterns. We find optimal values for these parameters through repeated tests while considering the characteristics of each D. We used default values for some parameters other than those listed above. For instance, node size=1, sample fraction= 0.8, and sampling scheme= bootstrapping. We determined these values via repeated tests under different settings. Training data size is $^2/_3 \times D$, and testing data size is $^1/_3 \times D$.

Through rigorous experiments with the RF and a few post-processing, country, gender, race, and work classes were identified as pattern-friendly QIDs in the Adult dataset. In these QIDs, one of the values occurred at very high frequencies (i.e., 87.81%, 83.22%, 79.06%, and 70.05% of the records, respectively) and those particular values no longer posed a threat to privacy because they are regarded as common patterns (population level information) [13]. In the Stroke Prediction dataset, marital status and smoking

habits were identified as pattern-friendly QIDs. One of those values occurred with very high frequency (64.36% and 57.75%, respectively) as well. We found that in the gender QID, one value had a frequency of up to 56.29%, but there were three values in this QID, so it was classified as a non-pattern-friendly QID. In the census datasets, country, race, gender, and class of worker were identified as pattern-friendly QIDs. In these QIDs, one of the values occurred at very high frequencies (i.e., 88.6%, 83.8%, 52.11%, and 50.24% of the records, respectively) and they can be regarded as general patterns. We found that in the marital status QID, two values had a combined frequency of up to 85.5%, and the other five had a frequency of 14.5%, so marital status QID was also classified as a pattern-friendly QID. The working hours QID has one of the values that occurred at high frequencies (e.g., 35.2%), but the chances of the appearance of this unique value in many classes/clusters are low, and therefore, it was also classified as a non-pattern-friendly QID. In the diabetes datasets, race, sex, and insulin dosage were identified as pattern-friendly QIDs. In these QIDs, one of the values occurred at very high frequencies (i.e., 74.7%, 53.8%, and 46.6% of the records, respectively) and were less risky in terms of privacy disclosures. The remaining QIDs in this dataset are numerical, and their values are not concentrated into one value.

Our analysis can greatly contribute to reducing over-anonymization issues by making use of these statistics during anonymization. We cross-validated the statistics using FA, and validation results ensured the correctness of the analysis. The 0.88 and 0.69 values were used for T_{Sm} and T_{η}, respectively. The eight (i.e., from 2 to 100) different k values were used, and the value of ℓ was set at 2 (the optimal limit for the chosen datasets). The optimal values satisfied the properties of the privacy-utility curve and ensured very small gaps between the results of utility and privacy.

Metrics and Results Evaluation Criteria: To evaluate the performance of the proposed model, five metrics were utilized: three metrics for measuring privacy strength, and two for utility estimation. To evaluate and contrast privacy, we used SA disclosure risk (e.g., QID-based re-identification first; inferring SAs later), the exposure of frequently occurring SA values within classes, and the data reconstruction risk. The disclosure risk, $\mathcal{D}$, is computed via Eq. 5:

$$\mathcal{D}(D^{'}) = \frac{\sum_{X \in D'} D(X^{'})}{|D'|} \tag{5}$$

where $D(X^{'})$ shows the probability of accurately finding SA of target $X^{'}$, which is expressed below.

$$D(X^{'}) = \begin{cases} 0, & \text{if } X^{'} \notin C \\ \frac{\sum_{y \in Y_C} max\{1/|C|, Pr_C(y)\}}{|Y_C|}, & \text{if } X^{'} \in C \end{cases} \tag{6}$$

where $Pr_C(y)$ is frequency of SA value y in $C \in D^{'}$, and Y_C is the total # of SA values in C. this function has two main concepts: user identity and SA revelation. The value of SA is exposed if an adversary can find the tuple related to the target, expressed as $Pr\ 1/|C|$. Also, an adversary can find SA value y having probability $Pr_C(y)$. So the probability (in total) is $max\{1/|C|, Pr_C(y)\}$. Further insight regarding the above equations can be learned from [41], [42]. The second metric (i.e., most occurring SA value y_i exposure $\mathcal{E}$) value can be measured via Eq. 7.

$$\mathcal{E}_{y_i} = \frac{\sum_{i|most\ occuring\ value\ of\ the\ SA} y_i}{\sum_{i=1}^{|C|} y_i} \tag{7}$$

In Eq. 7, the numerator represents the most occurring value in a sample/cluster, while the frequency of the most occurring value in D is given in the denominator. The value of data reconstruction $\mathcal{R}$ can be computed using Eq. 8:

$$\mathcal{R} = P_X(Q, Y) \approx P_{X^*}(Q, Y) \tag{8}$$

where P_X and P_{X^*} , respectively, denote data representation before and after anonymization.

To assess the $D^{'}$ utility, we used ILoss and accuracy ($\mathcal{A}$). To compute ILoss, we employed distortion measure ($\mathcal{DM}$), which is widely used [43]. $\mathcal{DM}$ can be computed by checking the H's level, on which QID values are mapped, divided by the total # of levels in H, as expressed in Eq. 9:

$$\mathcal{DM} = \sum_{Q=1}^{p} \frac{l_{act}}{l_{tot}} \times w_Q \tag{9}$$

The $\mathcal{DM}$ values from every QID and tuple were aggregated for assessment and comparison. If no generalization is applied, $\mathcal{DM} = 0$. The value of $\mathcal{A}$ is calculated via Eq.10:

$$\mathcal{A} = \frac{T_p + T_n}{|D'|} \tag{10}$$

where T_n and T_p are true negative and true positive.

Different scales (small and large) of k were employed to produce the different anonymized versions of all four datasets to fairly compare results with the baseline methods.

Baselines: To benchmark our model, we contrasted the results with eight prior SOTA and recent methods: ℓ-diversity (LD) [10], AFA [26], map-reduced-based anonymity called MRMondrian [27], DP-based method named LoPub [28], enhanced ℓ-diversity model named DHkmeans-LD [29], scalable and distributed anonymization technique and extension of Mondrian named SDDA [30], an obfuscation and generalization-based approach named HyObscure [31], and divide-and-conquer-based approach named DCA [32]. These baseline algorithms are representative and yield competitive results in terms of privacy and utility. We also conducted an ablation study to show the impact of two modules of our model and the results are in **Appendix D**.

6.3 Privacy results comparison with SOTA methods

This subsection provides the empirical results based on three privacy evaluation metrics: $\mathcal{D}$, $\mathcal{E}$, and $\mathcal{R}$.

(i) Comparisons of $\mathcal{D}$: First, we evaluated and compared the results of $\mathcal{D}$ by using the existing SOTA methods. Before reporting the $\mathcal{D}$ numerical value resulting from each version of $D^{'}$ with varying k, we show in Figure 5 the differences in anonymization results produced from a small subset of data via our model and an existing method. In this analysis, we demonstrate the key difference in anonymization between an existing model and our proposed model by using a sample of original records shown in Figure 5 (a).

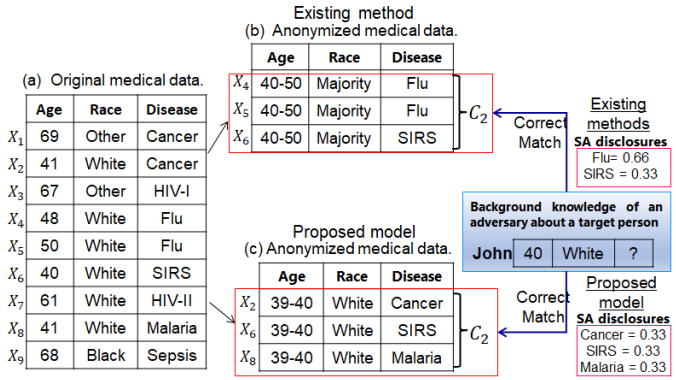


Fig. 5. Comparison of major differences in anonymization and attack settings in $D^{'}$.

Figures 5 (b) and 5 (c) show the $D^{'}$ sample created by the AFA algorithm (a recent enhancement of ℓ-diversity) and by our model, respectively. Existing methods mainly focus on ensuring diversity without paying due attention to homophily in QIDs, leading to overly anonymized data, as seen in C_2 of Figure 5 (b). An over-anonymized $D^{'}$ can make knowledge discovery harder (limiting analytics and data mining). Also, SA disclosure can be high when attention is paid solely to diversity rather than to distribution and balance. We demonstrate a background knowledge attack scenario in Figure 5, when the adversary has some QIDs of a user named John and wants to infer the SA. In this case, SA disclosure under the existing method is more than $1/\ell$ for flu. In contrast, our model ensures privacy (SA disclosure is less than $1/\ell$), and data are minimally generalized compared to the SOTA method. Our model applies various optimizations in the generalization process, and therefore, had a lower $\mathcal{D}$ than the previous method. The $\mathcal{D}$'s numerical values determined from two small scales D and with different k values are reported in Tables 6 and 7.

TABLE 6
Average $\mathcal{D}$ (Adult dataset): proposed model versus SOTA methods.

Algorithms	k value							
	2	5	10	20	30	50	75	100
LD [11]	0.90	0.81	0.78	0.71	0.69	0.76	0.82	0.72
AFA [26]	0.75	0.71	0.70	0.62	0.62	0.56	0.67	0.64
MRMondrian [27]	0.80	0.74	0.73	0.63	0.62	0.58	0.64	0.66
LoPub [28]	0.71	0.66	0.68	0.61	0.60	0.56	0.67	0.61
DHkmeans-LD [29]	0.84	0.77	0.74	0.62	0.64	0.60	0.72	0.69
SDDA [30]	0.69	0.65	0.66	0.6	0.59	0.55	0.65	0.6
HyObscure [31]	0.62	0.63	0.63	0.57	0.56	0.52	0.62	0.55
DCA [32]	0.65	0.64	0.64	0.57	0.57	0.53	0.64	0.57
Proposed model	**0.60**	**0.63**	**0.58**	**0.56**	**0.55**	**0.51**	**0.58**	**0.54**

In these experiments, we assumed the adversary had access to some record holder data and performed linking to infer SAs. From the results, we can see that our model had lower $\mathcal{D}$ values than the existing methods. There is no obvious increasing or decreasing trend with the k value because we used different combinations of QIDs to assess privacy leakage from $D^{'}$. The proposed model, on average, showed a 13.45% improvement compared to the existing methods on two small-scale datasets. The $\mathcal{D}$ obtained from two relatively large-scale datasets with eight different k values is given in Table 8 and 9.

TABLE 7
Average $\mathcal{D}$ (Stroke dataset): proposed model versus SOTA methods.

Algorithms	k value							
	2	5	10	20	30	50	75	100
LD [11]	0.80	0.78	0.74	0.76	0.75	0.69	0.67	0.73
AFA [26]	0.65	0.64	0.65	0.73	0.65	0.62	0.60	0.59
MRMondrian [27]	0.62	0.62	0.64	0.71	0.65	0.63	0.62	0.61
LoPub [28]	0.62	0.59	0.63	0.69	0.64	0.59	0.58	0.56
DHkmeans-LD [29]	0.70	0.65	0.68	0.73	0.68	0.65	0.65	0.67
SDDA [30]	0.60	0.58	0.61	0.67	0.62	0.57	0.57	0.59
HyObscure [31]	0.54	0.54	0.56	0.64	0.60	0.54	0.55	0.57
DCA [32]	0.57	0.56	0.58	0.66	0.61	0.56	0.56	0.59
Proposed model	**0.50**	**0.54**	**0.52**	**0.63**	**0.61**	**0.54**	**0.55**	**0.52**

TABLE 8
Average $\mathcal{D}$ (Census dataset): proposed model versus SOTA methods.

Algorithms	k value							
	2	5	10	20	30	50	75	100
LD [11]	0.97	0.81	0.80	0.69	0.65	0.86	0.79	0.97
AFA [26]	0.85	0.73	0.71	0.61	0.59	0.80	0.64	0.92
MRMondrian [27]	0.90	0.74	0.74	0.62	0.61	0.81	0.66	0.94
LoPub [28]	0.83	0.69	0.69	0.60	0.58	0.79	0.63	0.91
DHkmeans-LD [29]	0.88	0.79	0.75	0.65	0.64	0.82	0.74	0.95
SDDA [30]	0.81	0.67	0.68	0.59	0.56	0.76	0.60	0.89
HyObscure [31]	0.74	0.65	0.66	0.57	0.52	0.71	0.55	0.85
DCA [32]	0.77	0.65	0.67	0.58	0.54	0.74	0.57	0.88
Proposed model	**0.68**	**0.64**	**0.64**	**0.56**	**0.49**	**0.66**	**0.53**	**0.82**

TABLE 9
Average $\mathcal{D}$ (Diabetes dataset): proposed model versus SOTA methods.

Algorithms	k value							
	2	5	10	20	30	50	75	100
LD [11]	0.88	0.93	0.84	0.68	0.95	0.77	0.87	0.99
AFA [26]	0.71	0.87	0.75	0.60	0.90	0.72	0.77	0.88
MRMondrian [27]	0.80	0.88	0.79	0.61	0.88	0.73	0.80	0.90
LoPub [28]	0.69	0.86	0.74	0.59	0.90	0.69	0.76	0.80
DHkmeans-LD [29]	0.86	0.91	0.83	0.65	0.94	0.75	0.82	0.95
SDDA [30]	0.68	0.84	0.71	0.58	0.88	0.65	0.73	0.77
HyObscure [31]	0.64	0.81	0.67	0.53	0.84	0.60	0.67	0.76
DCA [32]	0.66	0.82	0.70	0.56	0.86	0.64	0.71	0.75
Proposed model	**0.62**	**0.78**	**0.64**	**0.57**	**0.82**	**0.57**	**0.64**	**0.85**

The lowest performance is shown by LD [11] baseline and comparable performance is shown by HyObscure [31]. The main reason for the high performance from HyObscure [31] is the joint use of obfuscation and generalization operations. In contrast, our model improves data quality and applies diversity-aware least generalization, leading to better privacy protection than most baselines. The proposed model, on average, showed a 10.39% improvement compared to the existing methods on two large-scale datasets. It is worth noting that all datasets have an imbalance, which means a significant portion of the data will have one SA value in the last column of the data, and if the adversary has some strong background knowledge, then $\mathcal{D}$ is 100% from all other baselines except HyObscure [31] and our model. HyObscure [31] is secure as it applies anonymity to non-sensitive parts of data as well, and our model is secure

as it adds synthetic records and enforces diversity criteria rigorously in all classes/clusters. These results fortify the significance of our model w.r.t. prevention of SA disclosure.

(ii) Comparisons of $\mathcal{E}$: Secondly, we evaluated and compared the results of $\mathcal{E}$ with the existing SOTA methods. As shown in Table 5, all datasets had one dominant value, and therefore, the privacy disclosure risk for that value can be higher in data publishing. We evaluated and compared the $\mathcal{E}$ value from the existing methods, and the results are shown in Table 10. To perform a fair assessment, we clustered $D^{'}$ based only on dominant SA values on the assumption that the SA is known to the adversary, and we performed a disclosure analysis. From the results, we can see that our model yielded better protection than the existing methods. It is important to note that HyObscure [31] also yielded competitive results. However, the placement of records is not changed much by HyObscure, so the dominant values can still be exposed to adversaries. These results verify the effectiveness of our model in terms of privacy preservation.

TABLE 10
Average $\mathcal{E}$ from our proposed model versus existing SOTA methods.

Dataset	SA value	Comparisons of $\mathcal{E}$ with SOTA algorithms.								
		Ours	LD [11]	AFA [26]	MRMondrian [27]	LoPub [28]	DHkmeans-LD [29]	SDDA [30]	HyObscure [31]	DCA [32]
Adult [37]	$\leq 50K$	22.21	50.24	38.14	33.23	29.45	38.65	31.72	27.67	25.46
Stroke [38]	0	25.48	83.70	52.12	35.12	33.12	62.13	35.98	31.76	29.34
Census [39]	-50000	12.73	74.71	56.03	49.36	37.35	72.83	41.23	34.19	32.32
Diabetes [40]	Yes	21.01	56.11	45.58	42.08	35.06	52.61	39.06	32.98	31.09

(iii) Comparisons of $\mathcal{R}$: Third, we evaluated and compared the results of $\mathcal{R}$ from the existing SOTA methods. Due to the significant rise in AI-powered tools, adversaries can reconstruct most parts of the data, and can subsequently launch QIDs and SA inference attacks. Protection against such a present-day attack is vital to securing personal data against misuse. To this end, we compared our model's results with eight existing methods from different anonymized versions of data, and the results are illustrated in Figure 6.

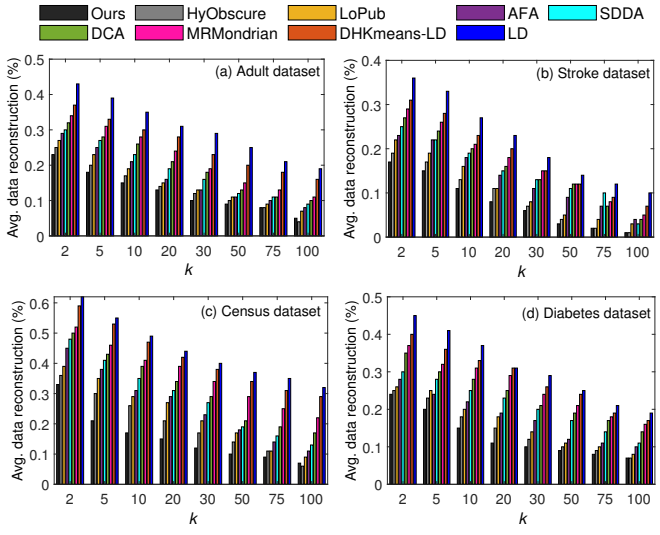


Fig. 6. Average $\mathcal{R}$: Proposed model versus existing algorithms.

As shown in Figure 6, $\mathcal{R}$ decreases with k, which means that $\mathcal{R}$ can be reduced by increasing k. However, the complexity increased for higher values of k, and most studies evaluate the privacy strengths for a reasonable k value. To perform a fair assessment, we performed experiments for $\mathcal{R}$ assuming the different proportions of D that can be available to the adversary. From the results, we can see that our model yielded better protection than the existing SOTA methods. On average, the proposed model reduced $\mathcal{R}$ by 43.75% on four benchmark datasets. Although HyObscure [31] has shown comparable performance and supersedes our model for some values of k, it incurs extra overheads due to the anonymity of non-sensitive data and at the expense of data utility. Hence, this baseline cannot be adopted in realistic scenarios, when both privacy and utility requirements are very high. These results verify the resilience of our model in terms of privacy preservation against present-day and robust privacy attacks.

Lastly, by improving data quality our model significantly reduced the number of records that risk exposure. As stated earlier, a substantial number of records can be exposed to adversaries if data are not enhanced before anonymization. Table 11 presents comparisons of our model with the existing methods in terms of a reduction in records at risk. The results showed our model had fewer, significantly reducing the number of records at risk in all four datasets. It is important to note that these comparisons are from a real D, and the number of risky records was further reduced via anonymization. Based on the experimental results and comparisons, our model has abilities to safeguard privacy more effectively than previous SOTA methods.

TABLE 11
Reduction in the number of records at risk: our model versus existing methods.

Dataset	Analysis	Existing methods	Our model
Adult [37]	# of records at risk	12,421	5,559
	% of records at risk	38.14 %	17.07 %
Stroke [38]	# of records at risk	4,612	1,404
	% of records at risk	90.25 %	16.41 %
Census [39]	# of records at risk	174,759	38,730
	% of records at risk	87.50 %	19.42 %
Diabetes [40]	# of records at risk	54,960	16,536
	% of records at risk	45.99 %	16.24 %

6.4 Utility results comparison with SOTA methods

In this subsection, we compare the results of our model with the existing SOTA methods in terms of $D^{'}$ quality. We used both special ($\mathcal{A}$) and general-purpose ($\mathcal{DM}$) metrics to quantify and compare the results. The results against the general-purpose metric (ILoss) computed with the help of $\mathcal{DM}$ metric are shown in Figure 7.

From the results, we see that ILoss increased with the k value due to more changes in the QID values. The relatively lower ILoss for the Adult dataset was due to the higher number of QIDs with common patterns, and from lower generalizations. In contrast, the Stroke dataset has mostly categorical QIDs, and generalizations are relatively higher than from the Adult dataset. The ILoss from two relatively large-scale datasets is higher than Adult and Stroke datasets, owing to higher # of records as well as QIDs. The similar

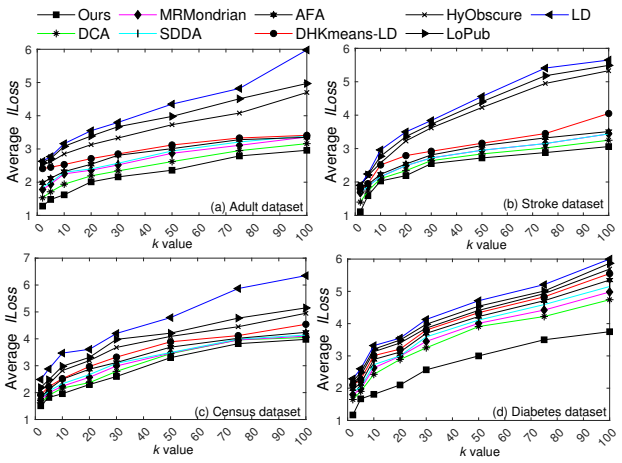
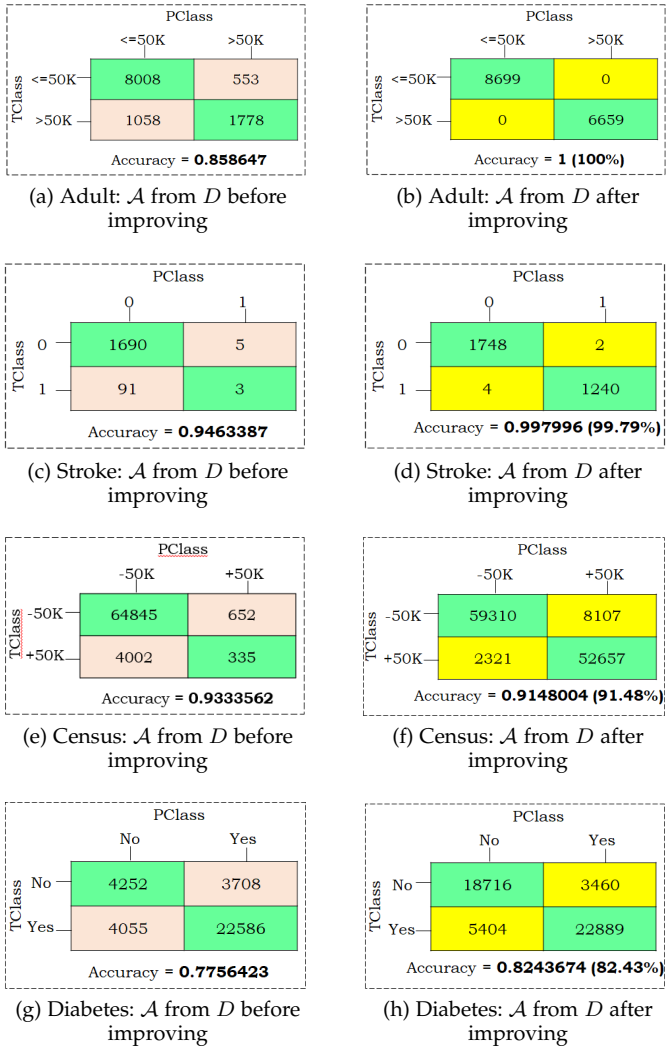


Fig. 7. Average ILoss: proposed model versus existing algorithms.



(a) Adult: $\mathcal{A}$ from D before improving
(b) Adult: $\mathcal{A}$ from D after improving
(c) Stroke: $\mathcal{A}$ from D before improving
(d) Stroke: $\mathcal{A}$ from D after improving
(e) Census: $\mathcal{A}$ from D before improving
(f) Census: $\mathcal{A}$ from D after improving
(g) Diabetes: $\mathcal{A}$ from D before improving
(h) Diabetes: $\mathcal{A}$ from D after improving

Fig. 8. Comparison of $\mathcal{A}$ and the confusion matrixes for the four datasets.

trend in the result is due to the dominance of categorical QIDs in each dataset. Through detailed assessment and fair comparisons, our model yielded a lower ILoss for all four datasets and outperformed all baselines for most values of k. The improvements in results are brought on by two optimizations (no generalization and lower-level generalization) that are simultaneously applied during the generalization process. These results show the significance of our model in terms of retaining higher truthfulness in $D^{'}$ compared to SOTA methods. The resulting $D^{'}$ can be highly useful for conducting analytics (mining useful information).

In the last set of experiments, we evaluated and compared the performance of our model in terms of $\mathcal{A}$, which belongs to the special purpose metrics category. In these experiments, we employed random forest as a prediction model because it is one of the extensively used models, and can yield higher $\mathcal{A}$. In the $\mathcal{A}$ calculation, the labels to be predicted were different for each dataset: Adult ($\leq$ 50K, $>$ 50K), Stroke (0,1), Census (-50000, 50000+), and Diabetes (yes, no). Before presenting results from $D^{'}$, we demonstrate a significant boost in the accuracy of D in Figure 8. The results show that our model significantly boosted $\mathcal{A}$, and the confusion matrix was more balanced than that from D, particularly for the minor SA value. Also, the false negative is lower than the true negative for all datasets. In census datasets, the $\mathcal{A}$ is slightly lower (i.e., -1.85%) than D, but the confusion metrics are still balanced compared to D. Despite low accuracy, it can still contribute to the satisfaction of ℓ-diversity parameters, which is the main assertion of this paper. To the best of our knowledge, none of the previous baseline methods have achieved similar accuracy on these datasets. Our model boosted $\mathcal{A}$ by 14.14%, 5.16%, and 4.87% on the benchmark Adult, Stroke, and Diabetes datasets, respectively.

The results for $\mathcal{A}$ from $D^{'}$ are in Figure 9. These results were produced from anonymized versions of the data created with distinct values of k. $\mathcal{A}$ decreases with k because data generalization decreases the diversity of QID's values, and transforms them from real to more general ones.

Referring to Figure 9, we can see that $\mathcal{A}$ resulting from our model was very close to D. The confusion matrix was more highly balanced than the SOTA algorithms. Further-

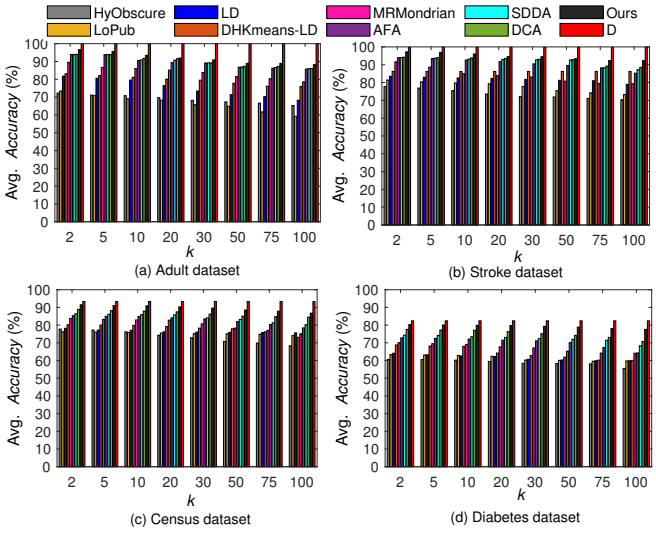


Fig. 9. Average $\mathcal{A}$: Proposed model versus existing algorithms and D.

more, our model provided better $\mathcal{A}$ values than the existing SOTA algorithms for most values of k. These results validated the supramacy of our model w.r.t. data mining and knowledge discovery. $D^{'}$ resulting from our model can be highly useful in data-driven services. Also, our model can protect minor values from dilution/erasure, which can prevent wrong conclusions from $D^{'}$.

Lastly, our model resulted in fewer changes in the data during anonymization, whereas existing SOTA algorithms anonymized entire sections of D. By applying higher generalization, the quality of $D^{'}$ can be seriously impacted. In some cases, the anonymized data can no longer provide any benefit to data miners except for frequency analysis. Our model separates the QIDs with common patterns, and applies no generalization to them, yielding a supreme quality $D^{'}$. The $D^{'}$ produced with our model can be extremely useful in data-hungry applications. Table 12 presents comparisons of the amount of generalization from our model versus the existing algorithms. Since all the algorithms we compared our results with apply generalization in the same manner, we combined the performance comparisons. The higher improvements in the Adult dataset are due to the existence of more pattern-friendly QIDs in it.

TABLE 12
Reduction in the amount of generalization: our model versus existing methods.

Dataset	Methods	# of generalizations	Improvement
Adults [37]	Existing algorithms Proposed model	307,167 167,185	45.57%
Stroke [38]	Existing algorithms Proposed model	51, 318 40,968	20.16%
Census [39]	Existing algorithms Proposed model	2,585,360 1,687,920	34.71%
Diabetes [40]	Existing algorithms Proposed model	841,140 612,282	27.21%

From detailed comparison and analysis, we found that the proposed model can maintain the balance between privacy and utility. Also, it yielded better results than SOTA algorithms in all settings. Based on the performance against recent SOTA algorithms, it is fair to say that the proposed model is a better candidate for PPDP, particularly when publishing high-quality data is imperative. It is the first step toward anonymization of a poor-quality D, and it can be highly useful in fulfilling data needs. The contrast between the proposed model and the DP model is provided in **Appendix E**.

6.5 Time complexity analysis

In this section, we analyze and compare the time overheads of the proposed model with the eight SOTA baseline methods. As discussed earlier, our proposed model has two main modules (data quality enhancement and shallow anonymization) whereas most baselines have three or more phases/modules/steps. In the first module, additional data is curated which may prolong the time in realistic scenarios. However, we reduce the vertical dimensions of the data (limit data to QID only) before curating new data, and therefore, the overheads are not very high. The training of CGAN is robust as the numerical and categorical QIDs are transformed into a vector representation that is suitable for neural networks. The condition vector and sampling-based training prevent the imbalanced learning issue. The mode collapse problem is resolved with the optimized value of the parameters and the PacGAN strategy. In the second module, clustering is applied to group similar users having diverse SA values, which takes relatively more time than the first module. On the other hand, at data generalization time, our proposed model reduces excessive lookup to the generalization hierarchies by identifying common patterns, which leads to reduced time overheads. Moreover, the addition of the virtual record makes the satisfaction of ℓ-diversity criteria easier, leading to reduced computing overheads by limiting redundant search operations for diverse values. In addition, k-anonymity and ℓ-diversity are simultaneously enforced rather than sequential application (apply k-anonymity first followed by ℓ-diversity) to reduce time. For a fair assessment of time complexity, we identified the main modules of each method and computed the time for comparative analysis. Also, some of the operations (e.g., basic pre-processing of data, clustering, k-anonymity and ℓ-diversity enforcement, and data generalization) were common among baselines and our model, therefore, we computed time once and assumed that it is the same across other baselines. Figure 10 presents the time overheads of our method and eight baseline methods on four benchmark datasets for eight different k values.

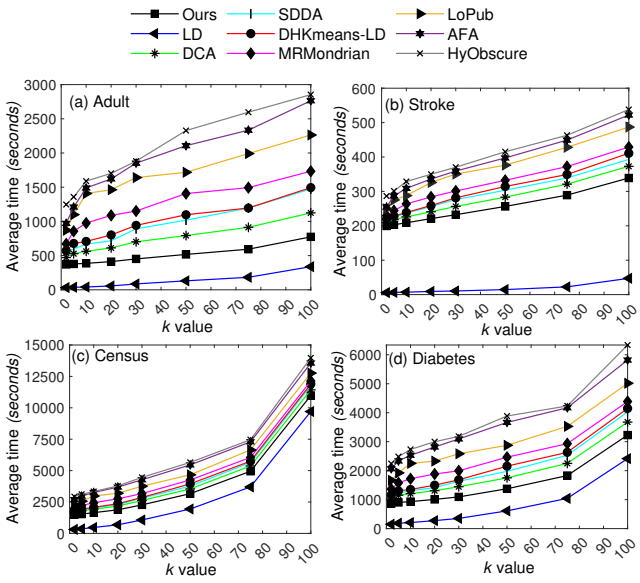


Fig. 10. Avg. time: proposed model versus existing SOTA algorithms.

From the results, it can be seen that time rises with the increase in k for all datasets. The time overheads from the HyObscure were very high due to the anonymization of non-sensitive attributes as well as QIDs. The higher time overheads occurred on the Census dataset due to more # of records as well as QIDs. In contrast, the time overheads were small for the stroke dataset, owing to the least # of records. The ℓ-diversity model is superior to all methods as it looks for fulfillment of ℓ criteria and makes the least adjustments to either lower IL or increase privacy when

data is imbalanced. In contrast, our method has higher overheads than ℓ-diversity, but privacy and utility results are significantly better on diverse datasets. Our model has a slightly lower time than some SOTA algorithms because most baselines use parallel architectures to reduce time. However, they do not explore ways to reduce overheads at data generalization time. In most baselines, there are additional steps for data partitioning and then combining results, therefore, the time complexity is high. In contrast, our model has few modules and due to a look-up reduction to the generalization hierarchy for a reasonable part of each dataset, our model has yielded lower overheads than most baselines. Lastly, the similar trend in time complexity result on four benchmark datasets is due to the identical structure (e.g., higher categorical QIDs and too few numerical QIDs) of each dataset. These results fortify the capability of our method in terms of lower overheads than SOTA baselines. The effect of the proposed pre-processing strategy on other scalable anonymization approaches is given in **Appendix F**.

The above-cited experimental evaluation of eight different types (e.g., 4 privacy, 2 utility, 1 generalization, and 1-time complexity) and comparisons with eight SOTA algorithms prove the superiority of the proposed data-centric ℓ-diverity model. We carried out performance evaluation on four datasets, all of which are benchmark datasets. The privacy results are better due to higher diversity in classes/clusters, and by paying ample attention to vulnerable parts of data. The chosen evaluation metrics can measure strengths against conventional and AI-powered attacks. The utility results are better as our model makes fewer changes in the data and restricts hefty generalization. The chosen metrics can measure the quality of $D^{'}$ from the perspective of both general (histogram analysis) and special purpose (building classifiers). Our model reduces the time complexity by lowering lookups into generalization hierarchies via the identification of PF QIDs, which remained unexplored in the literature. The overall improvements brought on by the proposed model on each dataset are given in **Appendix G**. The proposed model makes use of the generative AI and ML methods in the anonymization domain to solve performance bottlenecks and PUT, which aligns with the latest trends. Lastly, our model has applicability to poor-quality datasets which are prevalent nowadays whereas most of the existing methods only work well on balanced (or good-quality) datasets. Hence, our model is more practical and can help data owners outsource their data without risking users' privacy while providing higher utility. A concrete analysis of the proposed model's applications, limitations, and threats to the validity and their mitigation are given in **Appendix H**.

7 Conclusion and Future Work

In this paper, we proposed a novel anonymization model that significantly preserves privacy without degrading anonymous data quality in data-sharing scenarios. Specifically, our model applies to scenarios in which data quality is poor, whereas most of the existing methods lack robustness and cannot be applied, or put many individuals' privacy at risk if applied. Our model introduced three new perspectives in the privacy literature: (i) identifying vulnerabilities in data, and fixing them in an automated way before anonymization, (ii) creating balanced, diverse, and compact classes from D, and (iii) applying shallow generalization (or no generalization) by exploiting diversity and common-pattern information. To the best of our knowledge, the above three perspectives have not been simultaneously applied in any studies from the anonymization literature so far. We evaluated the efficacy of our model on benchmark datasets, and the results were carefully analyzed. Experimental results and analysis showed that the data-centric ℓ-diversity model a) is capable of accomplishing competitive results concerning privacy and anonymous data utility when the quality of original data is poor; b) offers less SA disclosure, especially when SA privacy preservation is the top priority; c) prevents disclosure of dominant values in SAs, and lowers data reconstruction risks, d) provides much higher utility in both general-purpose and special-purpose metrics, and, e) in boosting data quality, enhances accuracy by up to 14.37%. The computing overheads are also smaller than most baseline methods. The proposed model is highly suitable for domains involving limited or poor-quality data. In the future, we intend to extend the proposed model to detect more advanced types of vulnerabilities in data to further improve the privacy and utility results.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science and ICT) (RS-2024-00340882).

References

[1] S. Dash, S. K. Shakyawar, M. Sharma, and S. Kaushik, "Big data in healthcare: management, analysis and future prospects," *Journal of Big Data*, vol. 6, no. 1, pp. 1–25, 2019.

[2] J. Wieringa, P. Kannan, X. Ma, T. Reutterer, H. Risselada, and B. Skiera, "Data analytics in a privacy-concerned world," *Journal of Business Research*, vol. 122, pp. 915–925, 2021.

[3] M. Milani, Y. Huang, and F. Chiang, "Data anonymization with diversity constraints," *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[4] V. J. Reddi, G. Diamos, P. Warden, P. Mattson, and D. Kanter, "Data engineering for everyone," *arXiv preprint arXiv:2102.11447*, 2021.

[5] E. Strickland, "Andrew ng, ai minimalist: The machine-learning pioneer says small is the new big," *IEEE Spectrum*, vol. 59, no. 4, pp. 22–50, 2022.

[6] C. Hegde, "Anomaly detection in time series data using data-centric ai," in *2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*. IEEE, 2022, pp. 1–6.

[7] E. Jeczmionek and P. A. Kowalski, "Input reduction of convolutional neural networks with global sensitivity analysis as a data-centric approach," *Neurocomputing*, vol. 506, pp. 196–205, 2022.

[8] M. Motamedi, N. Sakharnykh, and T. Kaldewey, "A data-centric approach for training deep neural networks with less data," *arXiv preprint arXiv:2110.03613*, 2021.

[9] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.

[10] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, pp. 3–es, 2007.

[11] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *2007 IEEE 23rd International Conference on Data Engineering*. IEEE, 2007, pp. 106–115.

[12] C. Dwork, "Differential privacy: A survey of results," in *International conference on theory and applications of models of computation*. Springer, 2008, pp. 1–19.

[13] M. Strobel and R. Shokri, "Data privacy and trustworthy machine learning," *IEEE Security & Privacy*, vol. 20, no. 5, pp. 44–49, 2022.
[14] A. Sepas, A. H. Bangash, O. Alraoui, K. El Emam, and A. El-Hussuna, "Algorithms to anonymize structured medical and healthcare data: A systematic review," *Frontiers in Bioinformatics*, vol. 2, p. 112, 2022.
[15] B. C. Kara and C. Eyupoglu, "Anonymization methods for privacy-preserving data publishing," in *The International Conference on Artificial Intelligence and Applied Mathematics in Engineering*. Springer, 2023, pp. 145–159.
[16] M. E. Ferrão, P. Prata, and P. Fazendeiro, "Utility-driven assessment of anonymized data via clustering," *Scientific Data*, vol. 9, no. 1, pp. 1–11, 2022.
[17] Y. Yan, E. A. Herman, A. Mahmood, T. Feng, and P. Xie, "A weighted k-member clustering algorithm for k-anonymization," *Computing*, pp. 1–23, 2021.
[18] Y. Luo, Z. Wang, S. Zhang, and J. Liu, "Efficient-secure k-means clustering guaranteeing personalized local differential privacy," in *International Conference on Algorithms and Architectures for Parallel Processing*. Springer, 2023, pp. 660–675.
[19] H. Ranganatha *et al.*, "An enhanced data anonymization approach for privacy preserving data publishing in cloud computing based on genetic chimp optimization," *International Journal of Information Security and Privacy (IJISP)*, vol. 16, no. 1, pp. 1–20, 2022.
[20] A. Majeed and S. O. Hwang, "Quantifying the vulnerability of attributes for effective privacy preservation using machine learning," *IEEE Access*, 2023.
[21] Y.-T. Tsou, M. N. Alraja, L.-S. Chen, Y.-H. Chang, Y.-L. Hu, Y. Huang, C.-M. Yu, and P.-Y. Tsai, "(k, ϵ, δ)-anonymization: privacy-preserving data release based on k-anonymity and differential privacy," *Service Oriented Computing and Applications*, vol. 15, no. 3, pp. 175–185, 2021.
[22] N. Yuvaraj, K. Praghash, and T. Karthikeyan, "Data privacy preservation and trade-off balance between privacy and utility using deep adaptive clustering and elliptic curve digital signature algorithm," *Wireless Personal Communications*, vol. 124, no. 1, pp. 655–670, 2022.
[23] Y. S. Hindistan and E. F. Yetkin, "A hybrid approach with gan and dp for privacy preservation of iiot data," *IEEE Access*, 2023.
[24] F. Ashkouti, K. Khamforoosh, A. Sheikhahmadi, and H. Khamfroush, "Dhkmeans-ℓ-diversity: distributed hierarchical k-means for satisfaction of the ℓ-diversity privacy model using apache spark," *The Journal of Supercomputing*, pp. 1–35, 2021.
[25] W. Zheng, Y. Ma, Z. Wang, C. Jia, and P. Li, "Effective l-diversity anonymization algorithm based on improved clustering," in *International Symposium on Cyberspace Safety and Security*. Springer, 2019, pp. 318–329.
[26] J. A. Onesimu, J. Karthikeyan, J. Eunice, M. Pomplun, and H. Dang, "Privacy preserving attribute-focused anonymization scheme for healthcare data publishing," *IEEE Access*, vol. 10, pp. 86 979–86 997, 2022.
[27] X. Zhang, L. Qi, W. Dou, Q. He, C. Leckie, R. Kotagiri, and Z. Salcic, "Mrmondrian: scalable multidimensional anonymisation for big data privacy preservation," *IEEE Transactions on Big Data*, vol. 8, no. 1, pp. 125–139, 2017.
[28] X. Ren, C.-M. Yu, W. Yu, S. Yang, X. Yang, J. A. McCann, and S. Y. Philip, "Lopub: high-dimensional crowdsourced data publication with local differential privacy," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 9, pp. 2151–2166, 2018.
[29] F. Ashkouti, K. Khamforoosh, A. Sheikhahmadi, and H. Khamfroush, "Dhkmeans-ℓdiversity: distributed hierarchical k-means for satisfaction of the ℓ-diversity privacy model using apache spark," *The Journal of Supercomputing*, vol. 78, no. 2, pp. 2616–2650, 2022.
[30] S. De Capitani di Vimercati, D. Facchinetti, S. Foresti, G. Livraga, G. Oldani, S. Paraboschi, M. Rossi, and P. Samarati, "Scalable distributed data anonymization for large datasets," *IEEE Transactions on Big Data*, vol. 9, no. 3, pp. 818–831, 2023.
[31] X. Han, Y. Yang, J. Wu, and H. Xiong, "Hyobscure: Hybrid obscuring for privacy-preserving data publishing," *IEEE Transactions on Knowledge and Data Engineering*, 2023.
[32] R. Wang, J. Liang, S. Wang, and C.-C. Chang, "A divide-and-conquer approach to privacy-preserving high-dimensional big data release," *Journal of Information Security and Applications*, vol. 83, p. 103756, 2024.
[33] B. C. Fung, K. Wang, and S. Y. Philip, "Anonymizing classification data for privacy preservation," *IEEE transactions on knowledge and data engineering*, vol. 19, no. 5, pp. 711–725, 2007.
[34] V. Torra and G. Navarro-Arribas, "Attribute disclosure risk for k-anonymity: the case of numerical data," *International Journal of Information Security*, vol. 22, no. 6, pp. 2015–2024, 2023.
[35] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional gan," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
[36] F. Fkih, "Similarity measures for collaborative filtering-based recommender systems: Review and experimental comparison," *Journal of King Saud University-Computer and Information Sciences*, 2021.
[37] D. Newman, "Uci repository of machine learning databases, university of california, irvine," *http://www. ics. uci. edu/mlearn/MLRepository. html*, 1998.
[38] Kaggle, "Stroke prediction dataset," *Kaggle repository*, 2022. [Online]. Available: https://www.kaggle.com/fedesoriano/stroke-prediction-dataset
[39] M. Lichman *et al.*, "Uci machine learning repository," 2013.
[40] K. Sujatha and V. Udayarani, "Chaotic geometric data perturbed and ensemble gradient homomorphic privacy preservation over big healthcare data," *International Journal of System Assurance Engineering and Management*, pp. 1–13, 2021.
[41] Y.-Y. Wu, Z.-X. Shen, and W.-Y. Lin, "Anonymizing periodical releases of srs data by fusing differential privacy," *arXiv preprint arXiv:2211.10648*, 2022.
[42] A. Majeed, S. Khan, and S. O. Hwang, "Towards optimization of privacy-utility trade-off using similarity and diversity based clustering," *IEEE Transactions on Emerging Topics in Computing*, vol. 12, no. 1, pp. 368–385, 2023.
[43] B. C. Fung, K. Wang, A. W.-C. Fu, and S. Y. Philip, *Introduction to privacy-preserving data publishing: Concepts and techniques*. CRC Press, 2010.

Abdul Majeed received the Ph.D. degree in Computer Information Systems & Networks from the Korea Aerospace University, Korea, in 2021. He worked as a Security Analyst with Trillium Information Security Systems (TISS), Rawalpindi, Pakistan, from 2015 to 2016. He is currently working as an Assistant Professor with the Department of Computer Engineering, Gachon University, Korea. His research interests include privacy-preserving data publishing, statistical disclosure control, social network analysis and mining, and machine learning. Contact him at ab09@gachon.ac.kr.

Seong Oun Hwang received the Ph.D. degree in computer science from the Korea Advanced Institute of Science and Technology, in 2004, South Korea. He worked as a Software Engineer with LG-CNS Systems,Inc., from 1994 to 1996. He also worked as a Senior Researcher with the Electronics and Telecommunications Research Institute (ETRI), from 1998 to 2007. He worked as a Professor at the Department of Software and Communications Engineering, Hongik University, from 2008 to 2019. He is currently working as a full Professor with the Department of Computer Engineering, Gachon University, Korea. His research interests include cryptography, cybersecurity, and artificial intelligence. Contact him at sohwang@gachon.ac.kr.