A Hybrid Protection Method to Enhance Data Utility while Preserving the Privacy of Medical Patients Data Publishing

Shermina Jeba[1]
Department of Computing
Muscat College
Muscat, Oman

Mohd Arfian Ismail*[3]
Faculty of Computing
Universiti Malaysia Pahang
Kuantan, Pahang, Malaysia

Sarachandran Nair[5]
Department of Computing
Muscat College
Muscat, Oman

Mohammed BinJubier[2]
Faculty of Computing
Universiti Malaysia Pahang
Kuantan, Pahang, Malaysia

Reshmy Krishnan[4]
Department of Computing
Muscat College
Muscat, Oman

Girija Narasimhan[6]
Information Technology Department
University of Technology and Applied Sciences
Muscat, Oman

***Abstract*—Medical patient data need to be published and made available to researchers so that they can use, analyse, and evaluate the data effectively. However, publishing medical patient data raises privacy concerns regarding protecting sensitive data while preserving the utility of the released data. The privacy-preserving data publishing (PPDP) process attempts to keep public data useful without risking the medical patients' privacy. Through protection methods like perturbing, suppressing, or generalizing values, which lead to uncertainty in identity inference or sensitive value estimation, the PPDP aims to reduce the risks of patient data being disclosed and to preserve the potential use of published data. Although this method is helpful, information loss is inevitable when attempting to achieve a high level of privacy using protection methods. In addition, the privacy-preserving techniques may affect the use of data, resulting in imprecise or even impractical knowledge extraction. Thus, balancing privacy and utility in medical patient data is essential. This study proposed an innovative technique that used a hybrid protection method for utility enhancement while preserving medical patients' data privacy. The utilized technique could partition information horizontally and vertically, resulting in data being grouped into columns and equivalence classes. Then, the attributes assumed to be easily known by any attacker are determined by upper and lower protection levels (UPL and LPL). This work also depends on making the false matches and value swapping to make sure that the attribute disclosure is less likely to happen. The innovative technique makes data more useful. According to the results, the innovative technique delivers about 93.4% data utility when the percentage of exchange level is 5% using LPL and 95% using UPL with a 4.5K medical patient dataset. In conclusion, the innovative technique has minimized risk disclosure compared to other existing works.**

Keywords*—*Medical patients data publishing; anonymization; protection method for preserving the privacy

I. Introduction

Data publication is the simplest data-sharing method, allowing research institutions to conduct data mining operations on published medical-patient databases for knowledge extraction. This knowledge represents, interprets, or discovers new patterns [1] [2]. However, the potential of published data has yet to be explored because scholars face several challenges when extracting information from published medical-patient data. One of these challenges is the patient data privacy, which results in the exposure of individuals' identities, unauthorized access to information and private data, and use of personal information for unintended purposes [1] [3] [17]. Even if the identity attributes (IAs) (such as names and social security numbers) that identify users from the patient table are removed based on data protection, the remaining patient data can be used to re-identify the person in most cases. Furthermore, sensitive attributes (SA) may continue to flow due to linking attacks, in which sensitive data are revealed by linking the remaining attributes, such as those in published patient data, with other available data sources. This is referred to as a composition or intersection attack [3] [17]. Several anonymization techniques in PPDP have addressed data privacy concerns while preserving data utility. The goal of data anonymization is to reduce the threat of revealing personal information while preserving the possibility of using published data and causing uncertainty in identity inference or sensitive value estimation [50] [3]. However, information loss is inevitable when attempting to achieve a high level of privacy. In addition, the anonymization techniques may affect the use of patient data, resulting in imprecise or even impractical knowledge extraction via data mining. In data applications, balancing privacy and utility is crucial. In addition, nearly every technique leaves a question unanswered about whether anonymized data can be used effectively for data mining [1] [3].

The main contribution of this paper study is to propose an innovative technique that utilized a hybrid protection method to increase the utility of medical patients' data publishing while preserving privacy. This study aims to solve the problem of identity disclosure individuals or disclosing the sensitive value in medical patient's tables whilst preserving data utility. This research's accomplishments are summarised in the following points:

- The design of an innovative technique based on the UL technique in order to prevent attackers from identifying individuals or disclosing sensitive values in medical patient tables. In addition, the proposed technique stroked a better balance between utility, information loss, and privacy. The utilized technique could partition information into horizontal and vertical partitions. In the vertical partition, firstly, the attributes were separated into more than one. Then, the similar attributes were further grouped in a subset in a fashion that designated every attribute to a subset. Therefore, the subset of each attribute was called a cell (a pair of attributes), and the combination of these yielded the column. In the horizontal partition, the table was divided into different subsets so that each tuple could only be assigned to a single subset. Every subset of these tuples was referred to as a bucket or an equivalence class.

- A hybrid protection method can address the deficit in other existing works in determining the amount of protection required to prevent personal information disclosure. Instead of the random procedure employed in other works to break the correlations between the attribute values, the lower and upper protection levels (LPL and UPL) are then utilized in each equivalence class to determine the values of the unique and identical attributes for data privacy protection while keeping data utility. The LPL and UPL determine the level of protection surrounding the attribute values, ensuring that an attacker cannot get the sensitive information required to identify the record owner during such periods. This work also relied on value swapping within UPL and increasing the number of fake tuples within LPL as a safeguard against attribute exposure to any attack. The innovative technique protects published data from disclosure while increasing data value.

This paper falls into six sections: Section I provides the introduction and highlights the crux of the issue. Background and highlights the axes of privacy-preserving data publication are reviewed in Section II. Section III reviews relevant studies, Section IV describes the flow of research procedures in this paper, and Section V presents the evaluation of performance analysis. Finally, Section VI provide the conclusion and discussion of the study.

II. Background

One of the advantages of pervasive computing (ubiquitous computing) is the generation of large information volumes known as “big data”. The explosion of information has made retrieving private and public information on individuals a major part of everyday life [1] [30]. Typically, companies, such as health care, maps, and education acquire data to meet legitimate needs. Most data can be unstructured or intricate; a considerable portion of this data has been generated by a number of sources, like records of business sales, sensors in the use of Internet of Things, medical records of patients in hospitals, social media, and images and video archives. Big data processing using traditional data processing applications is becoming difficult [2]. Nevertheless, considering the Internet co-dependency and the IS, i.e., information systems, this data can be susceptible to corruption, theft or individual privacy violations [3].The ability to own data and extract new knowledge, which is known as data mining, is now considered a key competitive advantage [4].

Apart from the importance of extracting new knowledge (data mining), which is significant in many applications, an increasing concern has been focused on the privacy threats that emerge during data publishing for data mining operations, where numerous establishments need to publish data in different formats to extract new knowledge [4]. As a result, there has been a lot of focus on potential data privacy infractions as well as data exploitation; hence, effective data protection must be assured because failure may result in scenarios that can harm individuals and organizations [5] [6]. As a result, many organizations must choose between providing information and safeguarding their privacy in order to obtain this essential information [1]. This situation has motivated and prompted researchers to create a new research area known as privacy-preserving data publishing (PPDP) [7] [8]. This new research area is a sub-field of data mining that has gaining traction, which began as an encouraging model for providing first-hand solutions to address the current dilemma. Besides that, various methods have been developed for protecting information privacy, or wide-ranging policies have been imposed to safeguard sensitive data in PPDP [1] [3]. The primary goal of PPDP is to make a portion of these data available to all, from which it is utilized effectively for various tasks of data publication like application in future research and, at the same time, achieving the privacy of individuals’ information [3]. Although it is helpful, information loss through protection methods is inevitable when attempting to achieve a high level of privacy. In addition, these protection methods in PPDP may affect the use of data, resulting in imprecise or even impractical knowledge extraction. Furthermore, the privacy-preserving data publication (PPDP) is dependent on three axes: 1) data forms 2) Privacy Vs. Utility, and 3) methods of adversary knowledge (Fig. 1). These three axes will be described in the subsequent subsections in detail.

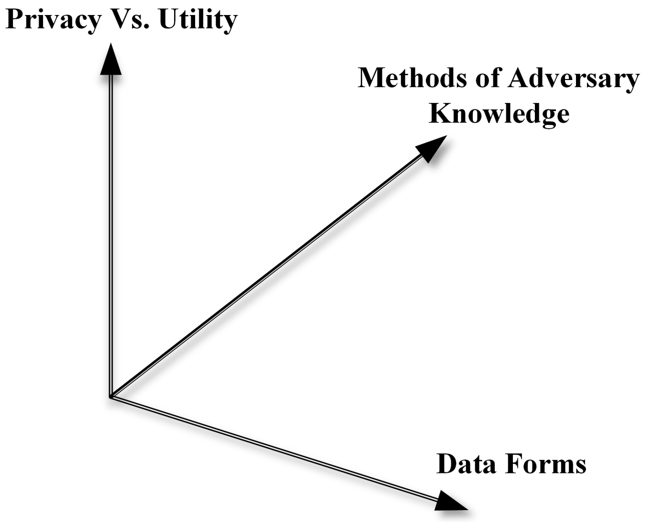

Fig. 1. Overview of Research Axes in PPDP.

A. Data Forms

Nowadays, the fast-growing number of Internet users and linked devices to ubiquitous computing have led to the generation of massive data volumes. It is essential to transform this gigantic, generated data into various forms for extracting valuable information embedded in massive data and providing many opportunities for advantages in different fields [2] [9]. Data forms, from which the types of data to be handled are selected, are seen to be a crucial component of the PPDP. The medical-patient database is one field where the deployment of big data can result in substantial changes. It can significantly raise the standard of patient care and provide insightful data that will help enhance patient outcomes, lower healthcare delivery costs, lower preventable diseases, and increase the overall quality of life.

Medical-patient databases are comprised of a wide range of records. Every record (i.e., row) signifies one client and includes a number of attributes specific to the selected client as shown in Table I [10] [11]. Such attributes are categorized into three categories [12] [13]: Identifying Attributes (IA), which distinguish explicitly the owner's records like his/her name, mobile number or the number of the driving license; Quasi-Identifier (QI) attributes, which denote the non-explicit attributes' sequence of individuals, including his/her age, gender, race, ZIP code, and date of birth as these attributes identify the record of the owner when these attributes are combined; and Sensitive Attributes (SA), which contain confidential data of individuals, such as diseases [14].

TABLE I. Medical Patient Database

Identifier (IAs)	Quasi-Identifier (QI)			Sensitive (SA)
Name	Age	Gender	Zipcode	Disease
Carl's	29	Female	462350	Disease
Abraham	22	Male	462351	Cancer
William	27	Male	150352	Flu
Linda	43	Female	462350	Heart Disease
Camila	52	Female	462350	Cancer

Much of the work done in PPDP is related to the static data publication form, with the idea of one record per person, and these records are presumed to be independent of various data publishers [10] [15] [16] [17]. However, the data form may be dynamic, i.e., it may be published consecutively while being changed by the same data publisher [18] [19] [20] [21]. The problem with static data publication forms is that it is possible for multiple organizations to publish a person's information at once [22]. In this situation, an attacker can use the published dataset and then, use a composition attack [23] [17] to change the dataset's privacy. An intruder commits a composition attack when he or she tries to find out who someone is by linking the different available attributes (QIs) in published available data to an external database to get sensitive information [23]. Therefore, the only way to protect people's privacy is to change these attributes in the available published data to hide the connection between a person and a specific value. This will stop these kinds of attacks and keep the public data usable [17].

B. Privacy vs. Utility

This aspect might be the most significant and fascinating aspect of the PPDP paradigm because any sanitized release of published data needs to address the trade-off between privacy and utility [1]. The common practice of sanitized data primarily depends upon certain guidelines and policies for restricting the publishable data types to achieve privacy. This common practice of sanitizing data is limited because it excessively falsifies data or demands a high level of trust in different scenarios of data-sharing impractically [24]. However, data privacy involves protecting private data from people who must not access this information and the individuals' capability to determine or interfere with the identity of individuals who can have access to their personal information. Developing protection tools and mechanisms for data publication is of is a major challenge so that the published data are practically usable and valid while the individuals' privacy is protected. A privacy policy is, therefore, a must to ensure that the sanitized release is safe from any attempts of intrusion while still being beneficial to the end-users. This means that there should be a balance conceived between these two notions, including utility and privacy [1].

One of the most prevalent and commonly utilised procedures in PPDP for providing privacy for individuals while retaining data utility is the anonymization technique of data before publication, which has previously been defined according to [25] as a group of certain protection methods, which aim to mitigate the risks of information disclosure for individuals, organizations or other businesses. The anonymization techniques use different protection methods and may be combined within the same technique, with the goal of causing uncertainty in identity inference or sensitive value estimation [25]. Some of these existing and most used protection methods are suppressing or generalizing [21] and perturbation [26] [27] [28] [29].

C. Methods of Adversary Knowledge

The assumption regarding the methods of adversary knowledge has led to the creation of alternative techniques and innovative protection methods. Although these techniques with protection methods provide a certain level of privacy preservation, it is crucial to examine the existing resources on adversaries regarding externally accessible data and different potential inferences for privacy-preserving of data. Privacy-preserving is classified into three categories based on the methods of adversary knowledge [24]:

The first category is that a certain privacy threat can occur when the attacker links the QI attributes within the published information with other available data sources. The attacker relies on the intersection of the datasets to exploit sensitive information since datasets are rarely isolated. This situation is known as a composition attack (i.e., the linking attacks), or an intersection attack [3] [17].

The second category is known as background knowledge attack. The attacker knows that the victim's record is in the released table or that the attacker has knowledge of some QIs for victims because the attacker is a close relative of the patient. The aim of the attacker is to link this knowledge for disclosure of sensitive attribute [1] [3]. The publishing table is regarded as privacy-preserving when it effectively prevents the attacker from performing linkages successfully.

An uninformative principle is the aim of the third category. This published table must give the attacker very little extra information outside the specified background knowledge. When the attacker possesses a larger distinction between prior and posterior beliefs, this attack is called a probabilistic attack. The inability to categorize many QIs attributes into a published table and to keep it without change (or modification) leads to a probabilistic attack and the possibility of accurately extracting the attributes of a person [1] [17] [24] [29].

III. Related Work

Data collectors collect a large amount of information from data owners and publish it so that data mining can provide a wide range of unprecedented potentials deemed necessary for providing meaningful information about data and improving the quality of medical patient data [2]. The data collected possibly hold the sensitive personal data of individuals. Thus, the goal of privacy-preserving data publishing (PPDP) is to release the data without publishing the private information about the owner of the data whilst preserving data utility in order to use them in any kind of medical or social analysis, etc. [48] [51] [52]. Differences in their concepts lead to differences in their protection methods. Binjubeir et al. [1] asserted that no general solutions could handle all privacy issues related to keeping sensitive information from unwanted disclosure. Hence, many techniques are used to provide privacy [53].

Despite the overlap between privacy and confidentiality in some contexts, when it comes to protecting people's privacy, there are certain ways they differ, especially related to their concepts and methods of protection. Confidentiality is seen as data-related, meaning that it is more about the data themselves, and it aims at protecting data from unauthorized access, alteration, or loss when transferred over a network [1] [54]. On the contrary, privacy has an additional "data owner-oriented" concept as it deals mainly with the data owners and aims to protect the private information of the data owners [55]. Hence, the current study uses PPDP as a way to keep sensitive data from being used illegally and to keep it safe from any threats. Consequently, there has been a lot of research on PPDP techniques over the past few decades.

The PPDP began with k-anonymity by Samarati and Sweeney [31]. Their work has been extended to cover various anonymization techniques like t-closeness [10], the l-diversity [32], (α, k)-anonymity [33], and Mondrian [34]. However, the aforementioned techniques have been susceptible to the composition attack [2] [23] because if two separately published tables satisfy a certain privacy principle, there will be no assurance that each pair's intersection for the equivalence classes of these two tables will be satisfying this same principle. Besides that, a higher dimensionality renders these techniques useless because the main record holders' identities are disclosed by combining the data with a public data (i.e., composition attack) [17] [35]. Readers can refer to [1] [36] [37] [38] for comprehensive studies on these techniques. Table II presents a summary of the PPDP protection techniques by offering the advantages and limitations of each technique.

Moreover, there are some techniques based on the dynamic data publication form [18] [20] [21] in addition to the static data publication form [10] [15] [16] [17]. The introduced techniques tackle composition attacks explicitly by utilizing an intersection of two or further sets of published data for uncovering individuals' sensitive information. In this anonymization process, which manages the dynamic data publication form, the data owner recognizes the entire versions of these published data and uses this information in these published versions for anonymizing the existing data set. For this study, the data owner has no knowledge of other datasets, which might be manipulated for the composition attack. Consequently, the data publications of various data custodians will be independent. However, the problem can be more challenging.

The most recent famous techniques for static and dynamic data publication form are the hybrid technique [35], e-differential privacy technique $(e-DP)$ [39], slicing technique [29], merging [17], and UL method [3]. All techniques have endeavoured to create privacy-preserving by using different protection methods.

As a protection method against sensitive value disclosure, the hybrid technique in [35] uses a combination of sampling, generalization, and perturbation. While the generalized data have been used as a method of protection in the $e-DP$ technique. The $e-DP$ technique in [39] probabilistically constructs a generalised contingency table and then adds noise to the counts. By using differential privacy-based data anonymization, the $e-DP$ can give robust privacy assurance for statistical query response and protect against the composition attack [23]. According to [17] [35] [40] [41], utilising the $e-DP$ for composition attack protection can result in significant data utility loss during anonymization.

The disclosed slicing technique [29], which is regarded as an innovative data anonymization technique in PPDP, has garnered a lot of attention. The authors [17] [29] [3] presented a non-generalizable risk disclosure preventative protection method. As a result, because the attribute values are not generalised, slicing protects data privacy while also preserving data utilities. These techniques [17] [29] [3] partition the data horizontally and vertically. In the vertical partitions (i.e., attribute grouping), the extremely associated attributes can be grouped into specific columns with every resulting column containing the attributes' subset. In horizontal partitions (i.e., the tuple partition), these tuples can be grouped in specified buckets or equivalent classes (Table III). The adoption of various protective methods causes the relationship between separate columns to be broken. This solution protects the privacy of public records from the hazards of attribute and membership exposure. Furthermore, slicing is preferred for high-dimensional data anonymisation since it preserves more data utility than attribute value generalisation. It is thus preferable to formalise slicing for a fuller comprehension. Consequently, the slicing formulation has been followed as suggested by Li et al. [29].

A. Attribute Grouping

The microdata table T consists of a set of t, $t \in T$ and n the number of a attributes, where t is a tuple of T and t is represented as $t = (t[a]_1, t[a]_2 ... t[a]_n)$, where

TABLE II. A Summary of Protection Methods for PPDP Techniques

Techniques	Protection Methods	Strength	Weakness
K-Anonymization	When the values of the QI attributes are modified, it is harder for an attacker to figure out who a person is. At the same time, the released data remains as helpful as possible, and the K value is used as a measure of privacy	This technique protects an individual's identity while releasing sensitive information	Indirect attacks on k-anonymity, like homogeneity attacks, background knowledge attacks, and composition attacks, make it possible to figure out exactly what a person looks like. Also, the high dimensionality renders this technique ineffective
L-diversity	This technique works to treat the values of a specific attribute similarly, regardless of its distribution in the data, thereby resulting in the sufficient representation of SAs within each equivalence class	This technique attempts to preserve privacy by a sufficient representation of SAs within each equivalence class	This technique is subjected to skewness attacks, similarity attacks, and composition attacks. Also, the high dimensionality renders this technique ineffective
T-closeness	The SA distribution in any equivalence class should be similar to the distribution of the attribute in an overall table	This technique works to distribute SAs in any equivalence class similar to the distribution of the attribute in an overall table, which leads to preserving privacy	This technique can't protect the critical values of the records from a composition attack reliably and constantly. In addition, The high dimensionality renders this technique ineffective
Mondrian	Partitioning the domain space recursively into several regions, each of which contains at least k records. A set of QI values are generalized in each equivalence class	Getting an anonymous dataset	This technique can't reliably and always protect the important values of the records from a composition attack. Also, most classification tools don't work well with overlapping intervals because they make it hard to classify things
(α, k)-anonymity	This technique integrated two novel concepts: (α, k)- anonymization by sampling and generalization for independent datasets to protect against composition attack	This technique effectively protects privacy and preserves data utility	There is still more data loss

TABLE III. A Published Data by Slicing

(Age, Gender)	(Zipcode, Disease)
(30, F)	(130350, ovarian cancer)
(23, M)	(130350, heart disease)
(28, F)	(130352, Flu)
(53, F)	(130350, heart disease)
(39, F)	(130352, Flu)
(60, M)	(130351, heart disease)

$t[a]_i \leq i \leq n$. In attribute partitioning, first, the attributes are separated into more than one; then, relevant attributes can be arranged in a specified subset, whereby each set can belong to a single subset only. Hence, the subset of each attribute is called a cell, and the combination of these yields the column. In the microdata table T, there are col columns, including $col_1, col_2...., colc$ satisfying $\bigcup_{i=1}^{c} col_i = a$ and for any $1 \leq i_1 \neq i_2 \leq col$, $col_{i,1} \cap col_{i,2} = \emptyset$. In these sensitive attributes, the sensitive attribute can be placed into the last-place position for an easy representation.

Definition 1 (cell): A cell represents one pair of attributes like (Gender, Age), where any cell $C_{col,E}$ is identified by the number of columns col_i and number of an equivalence class E_e. For example, in Table III, any cell in column (Age, Gender) is identified by col_i and E_j, where $1 \leq i \leq col$ and $1 \leq i \leq E$ and the first equivalence class is consisting of tuples $t = t_1, t_2, t_3, t_4$.

B. *Tuple Partition*

The goal of tuple partition is to generate different subsets of T in a manner that each tuple can only be assigned to only one subset. Each subset of tuples is known as a bucket or an equivalence class. Assume there are E equivalence classes, $E_1, E_2, ...E_e$ then, $\bigcup_{i=1}^{e} E_i = T$ for any $1 \leq i_1 \neq i_2 \leq e$, $E_{i,1} \cap E_{i,2} = \emptyset$.

C. *Problems of Slicing*

Slicing depends on the application of attribute grouping and tuple partition T. In Table III, by measuring the associations (similarity) among the attributes, the attribute group is applied, where very associated attribute values can be sorted into specified columns and uncorrelated attributes can be aggregated into other columns. The attribute partition can be characterized by Gender, Age, Disease, Zip Code, whereas the tuple partition is applied by grouping tuples into an equivalence class $\{t_1, t_2, t_3\}$, $\{t_4, t_5, t_6\}$. The central part of this tuple partition involves grouping all tuples that contain identical values in a similar equivalence class or it can be close to one another, thereby making it easier to breakdown uncorrelated attributes, and check whether an equivalence class satisfies I-diversity [1] [3] [34].

For slicing, the values of attributes are randomly permutated between the uncorrelated attributes for breaking the association among distinctive columns, whereas the attributes in columns that are highly correlated remain unchanged. However, the aspect of it remains an open question, i.e., “Does randomness always protect the identities of individuals from disclosure?” Slicing can have an impact on data utility and privacy, such as randomly permuting attribute values in each bucket, which increases the likelihood of creating bogus tuples, which reduces the utility of the released microdata. Furthermore, bogus tuples can easily cause various problems and wrong results in data mining process challenges. An attacker can learn about the implemented anonymization technique by analysing the spurious tuples in the published table, potentially breaching the privacy of public data [42] [3].

Hasan et al.[17] developed the merging technique to secure personal identification from disclosure. This been regarded as an extension of slicing. Merging’s primary purpose is to preserve privacy in many separate data releases by employing cell generalisation and random attribute value permutation to seperate connection between various columns. Regarding privacy risks and data utility, the merging technique conserved data usefulness while posing minor privacy hazards because of increased false matches in the released datasets. Nonetheless, the merging technique’s significant weaknesses are the randomised permutation way for the attribute values to breach the relationship between the columns and the increase in false matches for unique attributes. However, there will be a large number of matching buckets (more than the initial tuples), resulting in utility data loss, and could generate inaccurate and infeasible knowledge acquisition from data mining operations cite43 [44]. As a result, the main reasons for revealing people’s identities are unique attributes or the ability of some cells in the tuple to match with cells in other tuples in the same equivalence class, allowing precise extraction of a person’s attributes [1] [17] [29].

BinJubeir et al. [3] developed the UL technique as an efficient means of identifying the level of data protection required and selecting the best way to accomplish that level while keeping data utility. A lower level of protection, i.e., (LPL) and an upper level of protection, i.e., (UPL) can be employed to overcome these unique attributes with an identical data presence for data privacy protection. The unique attribute values are overcome by LPL, whilst the high identical attribute values are overcome by UPL. The LPL and UPL variables determine the level of protection surrounding the attribute values, ensuring that an attacker cannot access the sensitive information required to identify the record owner within such a time frame. The UL technique also makes use of value swapping to reduce the danger of attribute disclosure and increasing l-diverse slicing. Table IV illustrates the previous works which has been discussed of PPDP techniques for multiple independent data publishing. Some note that PPDP techniques are typically used to determine the level of privacy protection and information loss [56]. The two essential principles discussed here are privacy protection from any attack and data loss. The privacy preservation level refers to the degree of difficulty of estimating original data from perturbed data [57]. On the other hand, the information loss is a situation in which a significant portion of information of the original data set is lost after data anonymization.

The main contribution of this paper is to suggest an innovative technique that uses a hybrid protection method to enhance utility while protecting the privacy of medical patients’ data. The new technique proposed in this work is expected to keep data private while making patient data publishing more useful.

IV. Flow of Research Procedures

The flow of the research procedure is described in this section. It talks about the stages and methods that were used in this study to reach the research goals. As mentioned in the related work, many techniques have been proposed to address all privacy issues concerning protecting sensitive information from uninvited disclosure while preserving the utility of the data. However, there are still ways to enhance the utility of data while preserving user privacy. In essence, this study focuses on designing an innovative technique-based on UL technique that uses a hybrid protection method to enhance utility while protecting the privacy of medical patients’ data. Besides, this work used the UPL and LPL methods for anonymisation, which is more effective in determining the amount of protection required. UPL and LPL are choosing cell values that help identify disclosure and break the link between them by using a hybrid protection method (see Stage 3: protection methods) to keep data private while making patient data publishing more useful. This study aims to get a certain level of privacy while ensuring that as little information as possible is lost during data mining. So, this study aims to ensure that composition attacks and background knowledge attacks are less likely to happen when different independent hospitals release anonymous patient data while keeping the data intact. The flow of the research procedure consists of three main components, as depicted in Fig. 2. These three components are described in depth in the subsections that follow.

A. Research Gap

In related work, an analysis of how published data can be kept private is given. The main problem that has been found is when hospitals share patient information that could help them improve their efficiency and achieve their goals for the future. Sensitive Attribute (SAs) may still flow due to linking attacks wherein sensitive data may be revealed by linking the QI attribute in the published data with other available data sources. This situation is known as a “composition attack” or “intersection attack”. Also, many anonymization techniques fail to show a better balance of usefulness and privacy before any data product is made public. The criterion for evaluating the efficiency of the anonymization technique is the capability of data privacy preservation by decreasing the vulnerability of revealing people’s data and protecting the likelihood of the published data being used [3] [50] [17]. Previous techniques [29] [17] [34] [35] [39] recurrently resort to using protection methods, such as suppression and generalization, randomization, and/or combined. This work proposes designing an innovative technique based on the UL technique that uses a hybrid protection method to enhance utility while protecting

TABLE IV. A SUMMARY OF PPDP TECHNIQUES

Techniques	Protection Methods	Strength	Weakness
hybrid	This technique combines sampling, generalization, and perturbation by adding the Laplacian noise to the count of every SA value in each equivalence class	The proposed work reduces the risk of composition attacks and preserves data utility	There is still more data loss
$(e-DP)$	First probabilistically generates a generalized contingency table and then adds noise to the counts	$e-DP$ provides a strong privacy guarantee for statistical query answering and protection against a composition attack by differential privacy-based data anonymization	This technique is subjected to skewness attacks, similarity attacks, and composition attacks. Also, the high dimensionality renders this technique ineffective
Slicing	This technique uses vertical partitioning (attribute grouping), horizontal partitioning (tuple partition), and its sliced table should be randomly permutated	Slicing provides data privacy by randomly permutated of data and preserves data utilities that is devoid of generalization	Random permutate for attribute values are led to creating invalid tuples which will negatively affect the utility of the published microdata
Merging	The primary aim of merging approach is to preserve privacy by increasing the false matches in the published datasets and it uses vertical partitioning, horizontal partitioning, and its sliced table should be randomly permutated	Getting an anonymous dataset and preserves data utilities	The major drawback of merging is the random permutation procedure and increasing the false matches in the published datasets
UL method	The (LPL) and (UPL) can be used to determine the level of data protection needed and employed to overcome unique attributes and an identical data presence for data privacy protection whilst preserving data utility	This technique effectively protects privacy and preserves data utility	There are alternative protection strategies that may preserve data utility and privacy.

the privacy of medical patients' data. The goal of the hybrid protection method is to discover the peculiar features to swap between them rather than a random way of separating the relationship amongst the attribute values used in other existing works to enhance the privacy of published patients' data and keep more data utility.

B. Research Methodology (Design The Innovative Technique)

The overall methodology for designing the innovative technique, as shown in Fig. 2, comprises three stages for protecting the published patient data from unsolicited disclosure. Meanwhile, published patient data remains as useful as possible. Fig. 2 illustrates the proposed innovative technique for patient data protection, at the same time preserving the utility of the data. The following is the discussion of these three stages.

Stage 1 is preparing the dataset. The datasets stage aims to initialize the dataset and measure the correlation between attributes. To evaluate the experiments with other existing works, a medical patient database was used for the experiments [45]. This dataset is the standard machine learning dataset known as the "Adult" [46] has been changed and added one new column called disease. The main reason for adding this column is to mimic the medical patient dataset. Ronny Kohavi, together with Barry Becker extracted and congregated this dataset from the 1994 United States Census Bureau. Accordingly, this dataset is made up of fifteen QI attribute with 48,842 tuples.

In the dataset initialization process, independent patient datasets were required for the simulation of the existent medical patient data publishing case, particularly in a case in which such datasets are separately published by various medical organizations that have similar records. However, the pitfall of this proposition lies in the fact that the data of an individual is often published by many medical organizations [22]. Under such conditions, any intruder can initiate a composition attack or background knowledge attack [23] [17] on such published patient datasets just to alter the privacy of the dataset. In the process of dataset initialization, this independent dataset is taken from a database of medical patients.

After the initialization process, the correlation between attributes is measured, where the initialization of dataset generates various medical patient datasets to simulate the actual independent medical patient data publishing scenario. Each medical patient dataset should be treated as microdata table T. In a case in which the microdata table T applied possess a_i attributes, where $i = 1, 2, ...$ The strength of the correlations between pairs of attributes can be computed using several methods [17] [3]. Because most attributes are categorical, the most suitable method for the estimation of the correlations between pairs of attributes is the mean square contingency coefficient (MSCC). The MSCC is a chi-square useful measure of the correlation between two categorical attributes. The value of this coefficient r ranges from $[0, 1]$. If there is a perfect relationship between the two attributes, it would be preferable to have the measure of the association have a value of 1.

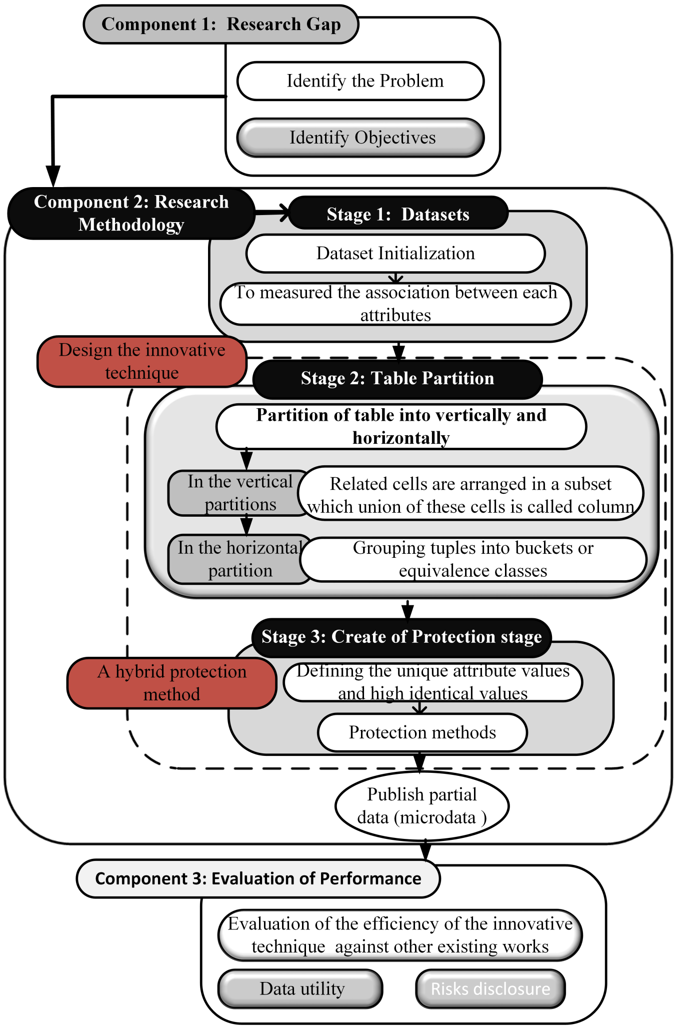
Component 1: Research Gap
Identify the Problem
Identify Objectives
Component 2: Research Methodology
Stage 1: Datasets
Dataset Initialization
To measured the association between each attributes
Design the innovative technique
Stage 2: Table Partition
Partition of table into vertically and horizontally
In the vertical partitions
Related cells are arranged in a subset which union of these cells is called column
In the horizontal partition
Grouping tuples into buckets or equivalence classes
Stage 3: Create of Protection stage
A hybrid protection method
Defining the unique attribute values and high identical values
Protection methods
Publish partial data (microdata)
Component 3: Evaluation of Performance
Evaluation of the efficiency of the innovative technique against other existing works
Data utility
Risks disclosure

Fig. 2. Flow of Research Procedures.

Otherwise, these measures differ in their maximum value. In case of no relationship between the two attributes, the measure of association has a value of 0. The MSCC between a_1 with value domain $\{v_{11}, v_{12}, ...v_{1d1}\}$ and a_2 with value domain $\{v_{21}, v_{22}, ...v_{2d2}\}$, and their domain sizes are d_1 and d_2, respectively. The MSCC between a_1 and a_2 is defined as follows [17] [3]:

$$r^2(a_1,a_2) = \frac{1}{min\{d_1,d_2\}} \sum_{i=1}^{d_1}\sum_{j=1}^{d_2} \frac{f_{ij} - f_i * f_j}{f_i * f_j} \qquad (1)$$

, where $r^2(a_1,a_2)$ is the MSCC between a_1 and a_2 attributes; $f_{i.}$ and $f_{.j}$ refer to the occurrence fractions of the v_{1i} and v_{2j} in the data, respectively; and f_{ij} is the fraction of cooccurrence of v_{1i} and v_{2j} in these data. Therefore $f_{i.}$ and $f_{.j}$ are the marginal totals of f_{ij} : $f_{i.}$ =$\sum_{j=1}^{d_2} f_{ij}$ and $f_{.j}$ =$\sum_{i=1}^{d_1} f_{ij}$. $0 \le r^2(a_1,a_2) \ge 1$.

Stage 2 deals with vertical and horizontal fractionalization of table. The dataset in the table is vertically and horizontally fractionalized, depending on the computation of correlation r for respective attributes pairs. This phase is aimed at categorizing similar attributes according to the degree of their inter-attribute connections that are suitable for utility and privacy. Concerning the utility of data, closely connected attributes are categorized in order to make warrant that their inter-attribute connections are maintained. Notwithstanding, as regards privacy, detection of vulnerabilities is greater as a result of the categorization of unrelated attributes in comparison with the categorization of more connected attribute values, which makes them highly identifiable. In order to guarantee a higher level of protection, it is important to dissolve the connection that exists among the unrelated attributes [29]. The microdata table T consists of a set of t, $t \in T$ and n the number of a attributes, where t is a tuple of T and t is represented as $t = (t[a]_1, t[a]_2...t[a]_n)$, where $t[a]_i \le i \le n$. In the vertical partition, firstly, the attributes are separated into more than one, then, the similar attributes are further presented in a subset in a fashion that designates every attribute to a subset. Therefore, the subset of each attribute is called a cell (a pair of attributes), and the combination of these yields the column. In the microdata table T, there are col columns, including $col_1, col_2...., col^c$ satisfying $\bigcup_{i=1}^{c} col_i = a$ and for any $1 \le i_1 \ne i_2 \le col$, $col_{i,1} \cap col_{i,2} = \emptyset$. Furthermore, QIs, SAs, and all other attributes presented in columns col_i,$1 \le i \le n$ They are clustered in n columns denoted by col_n, upon which the size of sensitive column col^c is not dependent. In some cases, the number of attributes a_s in the sensitive column col^c may be predetermined to be c.The col^c is determined in size by the use of parameter c, mathematically presented as $|col^c| = c$, in a case where $c = 1$, $col^c = 1$ also. That is, $col^c = \{S\}$. In a scenario where $c = 2$, the procedure is said to be equal to bucketization. In the case where $c > 1$, $|col^c| > 1$. QI attributes are contained in the sensitive column col^c. So as to ease the discussion in this study, the sensitive attribute a_s is focused on as one. Assuming the column in which a_s is contained is last column col^c. The column is as well referred to as sensitive column, as presented in Table V. In a case where several sensitive attributes are contained in the data, their individual or collective distribution may be employed [14]. Attributes (cells) that are highly related are put together in a column in vertical partitions, while unrelated are as well put in separate columns in a way that individual attributes a_i becomes designated to one subset. As shown in Table V, col_i columns $\{col_1, col_2...., coln\}$ contain all attributes a_i. In Table V, the three partitions for the Col_i columns are presented:

(1) T^* contains all columns with highly correlated attributes col^*, where col^*= $\{col_1^*, col_2^*, ...col_i^*\}$, and $col^* \in T^*$.

(2) T^{**} contains all columns with uncorrelated attributes col^{**}, where col^{**}= $\{col_1^{**}, col_2^{**}, ...col_i^{**}\}$, where $col^{**} \in T^{**}$.

(3) T^c contains columns with SA col^c when a single SA exists, and its SA is placed in the last position for easy representation, where $col^c \in T^c$ and $(col^* \cap col^{**}) \cap T^c = T$.

K-medoid clustering algorithm that is otherwise called partitioning around medoids algorithm (PAM) [47], is employed in the presentation of similar attributes in columns in a manner that designates each attribute to a column is used to organize the similar attributes into columns for each attribute to be part of a column. This algorithm guarantees the resolution of

TABLE V. Example of Partition the Table T into Three Columns

T^* contains all columns with highly correlated attributes.		T^{**} contains all columns with uncorrelated attributes.		T^c contains a column with sensitive attributes.
col_1^*	col_2^*	col_1^{**}	col_2^{**}	col^c
(a_1, a_2)	(a_3, a_4)	(a_5, a_6)	(a_7, a_8)	(a_s)

every attribute in the form of a point in the cluster space, while the inter-attribute disparity in the clustered space is represented thus: $d\,(a_1, a_2)$= $1 - r^2(a_1, a_2)$, that falls within the range of 0-1. However, the disparity amongst affiliated data points becomes less within the clustered space if two attributes are firmly correlated. After determining the disparity amongst affiliated between the related data points, the k-medoid method arranges related attributes in a subset called a cell, and the combination of these yields the column $(T^*, T^{**}$ and $T^c)$.

In horizontal Partition, the table is divided into different subsets so that each tuple can only be assigned to a single subset. Every subset of these tuples is referred to as a bucket or an equivalence class. Assume there are E equivalence classes, $E_1, E_2, ...E_e$ then, $\bigcup_{i=1}^{e} E_i = T$ for any $1 \leq i_1 \neq i_2 \leq e$, $E_{i,1} \cap E_{i,2} = \emptyset$. In addition, all tuples containing similar values are categorized into categories referred to as bucket or equivalence classes. Here, every individual is joined to 1 specific sensitive value in a way that makes it impossible for an attacker to penetrate the SA values of an individual where the probability is greater than 1/l. The tuples were categorized by the Mondrian [34] algorithm. They are separated in the equivalence classes, in the absence of generalization attributes, according to the top-down technique.

Stage 3 is the protection. The table partition stage generated the partition of microdata table T into partitions vertically and horizontally partitioned. The aim of the table partition stage is for all attributes to be clustered into columns (including both QIs and SAs) to prevent the unauthorized disclosure of an individual's identity by altering attributes (QI values) so that the connection that exist among the individual and specific values are hidden while ensuring that the data published is used through the application of protection methods. In this stage, the hybrid data protection method will provide robust patient data privacy while increasing medical data publishing utility for microdata table T partitioned using the innovative technique based on the UL technique via two steps, namely defining the unique attribute values, high identical values, and protection method.

1) Defining the Unique Attribute Values and High Identical Values: When it comes to a hybrid protection method, the magnitude of the connection among attributes is chiefly measured by the correlation coefficient r. Formula 2 illustrates the manner in which UPL and LPL attempt to define the unique attribute values and high identical values through the extraction of two kinds of cell values: (1) the values of the exclusive (unique) cells and (2) high identical cell values in T^{**}. Each cell that possesses unique values and such values are within the range of $0.0 < LPL \leq \Theta$ are determined by the LPL. For such attributes, the r value is usually hovering around 0 but not equivalent to 0. In a similar vein, the UPL determines those cells that possess numerous similar attributes with values that are in the range of $\Theta \leq UPL < 1.0$. r value for such attributes usually hover around 1 but is not equivalent to 1. Assuming those cells possess high r value in T^{**}, that will imply that the possibility of cells are in the same equivalence class. A hostile party has a higher degree of certainty around the SA when such cells are linked to other cells in T^*, thereby resulting in violation of privacy. In the remaining cells, the attributes and membership remain secured since they appear in multiple equivalence classes. In addition, the remaining cell categories need to be greater than a given limit, i.e., it should have a diversity value that is greater than or equal to 2 in respective equivalence classes, for the proposed privacy goal to be achieved.

$$(UPL\ and\ LPL) = \begin{cases} \overline{C_{col,E}} = \Theta \leq UPL < 1.0 \\ \underline{C_{col,E}} = 0.0 < LPL \leq \Theta \end{cases} \tag{2}$$

It is the goal of UPL and LPL to discover the collection of unique cells and high identical values for cells from T^{**}, that are assumed to be known to intruders into these periods: $0.0 < LPL \leq \Theta$ and $\Theta \leq UPL < 1.0$. The attributes that are for discovering within this period are referred to as the protection attributes. Protection rate, represented by $theta$, refers to values that have been initially tagged to be discovered. Typically $theta$ is in the range of $1\%-10\%$, implying that the fraction of protection qualities will be smaller than one.

Definition 2 (Matching Buckets): Assuming col^{**} represents the columns, and col^{**}= $\{col_1^{**}, col_2^{**}, ...col_n^{**}\}$, and $col^{**} \in T^{**}$. Let t^{**} represent a tuple, and $t^{**}|col_i^{**}|$ represent the col_i^{**} value of t^{**}, then let E^{**} represent an equivalence class in microdata table T^{**}, and $E^{**}|col_i^{**}|$ stand for the multiset of col^{**} values in equivalence class E^{**}. E^{**} denotes a matching bucket of t^{**} if for all $1 \leq i \leq col^{**}$, $t^{**}|col_i^{**}| \in E^{**}|col_i^{**}|$.

2) Protection Method: This study's hybrid protection method guarantees the privacy criteria in every equivalence class. In order to enhance utility as well as individual privacy in the suggested innovative technique, the connection amongst unique attributes as well as cells that possess similar identical values are eliminated through two steps.

- Creation of Fake Tuples
  During the protection method step, a random permutation in a bucket may not be shielded from attribute or membership disclosure because permutation increases the risk of attribute disclosure rather than ensuring privacy [42]. In addition, increasing the false matches as a protection method in both unique attributes and high identical values, this method may generate a slight fraction of these fake tuples; however, this can result in countless matching buckets (i.e., more than the number of original tuples), leading to a huge data utility loss, producing erroneous or unfeasible extraction of the knowledge via operations of data mining [43] [44] [3]. Therefore, the fake tuples are used in a bucket as a protection method for all unique attributes (LPL) only. A few fake tuples do not change how useful the published patient dataset is, but they make it more likely that false matches will be found during a composition attack on the published

table [17] [29]. As a protection method, n fake tuples with similar QI values are made and given sensitive values based on how sensitive values are spread out in the original dataset. The main goal of creation of fake tuples is to obtain the anonymized table T.

- Swapping or Generalization of Attributes
The protection method for UPL attributes is attribute swapping or generalisation. This study's innovative technique ensures the privacy criterion in each equivalence class. To boost diversity and personal privacy, rank swapping is utilised to break the relationship between attributes with high identical values. Attribute swapping modifies tuple data with high identical values (UPL) by switching the values of these characteristics across record pairs in a defined proportion of the original data. When it is not possible to swap attributes, the attributes must be generalised. The primary purpose of swapping or generalising attribute values is to create the anonymized table T, which has no nonsensical combinations in the record (invalid tuples) while satisfying the l-diverse slicing.
Definition 3 (Attribute Generalisation): Let T^{**} be part of microdata table T, and a_i^{**} be a QI attribute set in T^{**}. Generalisation replaces the QI attribute values with their generalised version. Let d_i^{**} and d_j^{**} be two domains with dimensional regions $\{d_{i1}^{**}, d_{i2}^{**}, ...d_{in}^{**}\}$ and $\{d_{j1}^{**}, d_{j2}^{**}, ...d_{jn}^{**}\}$, respectively, where $\bigcup_{d_{in}^{**}} = d_i^{**}$ and $d_i^{**} \cap d_j^{**} = \emptyset$. If the values of d_j^{**} are the generalisation of the values in domain d_i^{**}, we denote $d_i^{**} < d_j^{**}$ (a many-to-one value generalisation procedure). Generalisation works according to a domain generalisation hierarchy, which refers to a collection of domains that is ordered according to relationship $d_i^{**} < d_j^{**}$ (see Fig. 3).

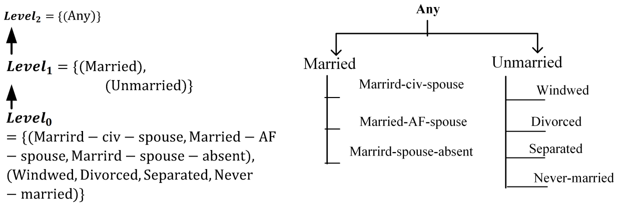
$Level_2$ = {(Any)}
$Level_1$ = {(Married), (Unmarried)}
$Level_0$ = {(Marrird − civ − spouse, Married − AF − spouse, Marrird − spouse − absent), (Windwed, Divorced, Separated, Never − married)}
Any
Married
Unmarried
Marrird-civ-spouse
Married-AF-spouse
Marrird-spouse-absent
Windwed
Divorced
Separated
Never-married

Fig. 3. Example of the Domain (Left) and Value (Right) Generalization Hierarchies for the Marital-Status(MS) Attributes.

In Fig. 3 (right), the likely domain generalisation hierarchy for marital-status (MS) attributes is described. At lower levels in the generalisation hierarchy for (MS) attributes, generalisation is not used. Nonetheless, at the top of the hierarchy levels, the MS tends to be more general. A singleton is a maximal domain level element that denotes the likelihood of values to be generalised in every domain to a single value.

C. Evaluation of Performance

This study describes maintaining privacy as minimizing disclosure of information on individuals. The usefulness of the data means to what extent we can use the sterile medical patient dataset for intensive analyses. For instance, by suppressing each QI, a medical patient dataset can be generalized. Maximum privacy is provided in this manner, but the information obtained is useless. Finding a good balance between privacy and utility is necessary because the published datasets (sanitized) must permit tasks related to data mining operations for search and analysis. As a result, the usefulness of data in the published medical patient dataset is assessed by how well statistical and aggregate data are used. The ability to protect data privacy by lowering the risk of disclosing personal information and maintaining the potential use of published data is the criterion for judging the effectiveness of the anonymization technique [3] [17] [1].

V. Comparison of Evaluation

Python was used to implement this experiment. The independent medical patient datasets were the experimental prerequisite to complete the experimentation of the actual independent data publishing setup. The independent datasets, known as the medical-patient-dataset, were pulled from the medical patient dataset [45], which contained eight QI attribute values: marital status (categorical, 7), relationship (categorical, 6), gender (categorical, 2), age (continuous, 74), work class (categorical, 8), salary (categorical, 2), disease (categorical, 16), as well as occupation (categorical, 14).

Each dataset contains 4K tuples chosen at random, with the remaining 8K tuples being used to generate an overlapping tuple pool and to check for potential composition attacks. By injecting 100, 200, 300, 400, and 500 tuples into the medical-patient-dataset, five copies of the remaining tuple pool were created, yielding datasets of the following sizes: (4.1K), (4.2K), (4.3K), (4.4K), and (4.5K) for the medical-patient-dataset.

In the experimental analysis, the static data publication form might be employed. There are two basic approaches for making the published medical patient dataset available. The first way is an interactive setting in which the data collector computes a function on the medical patient dataset in order to answer the data analyser's inquiries. The second way is a non-interactive setup in which the medical patient dataset is sanitized and then released [3]. The experiment was designed to evaluate non-interactive privacy settings; however, the majority of the work was done on differential privacy [39], which was consistent with interactive settings because medical patient datasets were commonly known to be made public. As a result, for the experiment on differential privacy, which was noted in [39], the noninteractive mode was chosen. Furthermore, the merging, $e-DP$, hybrid, UL, and Mondrian techniques yielded the quasi-identifier equivalence class as k-anonymity [16].k = 6 was chosen to build an equivalence class, where L-diversity was also provided as 6. The primary goal of L-diversity is to protect privacy by increasing the diversity of sensitive values. The Laplacian noise in a differential privacy equivalence class was added to the sensitive values' count [49] with e= 0.3 for the e-differential privacy budget.

The experiments on the medical patient database were performed in two parts. The first part was designed to evaluate the effectiveness of the innovative technique in data utility preservation by comparing it to other existing works. The second part was designed to assess the innovative technique to see how well it can fight and prevent composition attack occurrence. The innovative technique's effectiveness was tested by relating it to the effectiveness of similar techniques, like hybrid [35], merging [17], $e-DP$ [39], UL method [3] and Mondrian [34] techniques, in the non-interactive privacy settings. The experimental results showed that the innovative technique provided privacy protections against the considered attacks while preserving data utility.

A. Data Utility Comparison

This experiment measured the data utility obtained from the distortion ratio (DR). The DR in published medical patient dataset can be evaluated by different methodologies [21] to quantify the anonymisation outcome on the overall distortion data. The generalised distortion ratio (GDR) is a suitable measure for estimating the DR [44] used to quantify the anonymisation outcome on the overall distortion data.

The swap and generalise method was used to break the association of the attributes because the majority of these attributes were categorical. For any two categorical attributes $(a_1^{**}, a_2^{**} \in T)$, where t is its taxonomy tree and a node p in t is used to swap or generalise the attributes, the DR with p is defined as follows:

$$DR(a_1^{**}, a_2^{**}) = \begin{cases} 0, a_1^{**} = a_2^{**} \\ \frac{|common(a_1^{**}, a_2^{**})|}{|N|}, a_1^{**} \neq a_2^{**} \end{cases} \quad (3)$$

, where $|N|$ denotes the set of all the leaf nodes in t and $|common(a_1^{**}, a_2^{**})|$ is the set of leaf nodes in the lowest common tree of a_1^{**} and a_2^{**} in t.

Fig 3 denotes the taxonomy of the marital-status (MS) attribute; if the values of a_1^{**} and a_2^{**} are in the same rank group and have no nonsensical combinations, then their swap values are equal, and the DR is 0. Moreover, if the values of a_1^{**} and a_2^{**} are not in the same rank group or have any nonsensical combinations, then, their generalized values are equal to $\frac{|common(a_1^{**}, a_2^{**})|}{|N|}$, and the DR is equal to $\sum_{j=1,k=1}^{n,m} d_{j,k}$, where $d_{j,k}$ represents the attribute's distortion of a_j^{**} of the tuple t_k.

The distortion ratio (DR), also known as data utility, is a proportional measure that compares the amount of distortion in a generalised medical patient dataset to the amount of distortion in a fully generalised medical patient dataset. It is possible to determine the value of the data by subtracting the DR from Equation 4 shown below [21]:

$$Datautility = (100 - DR)\% \quad (4)$$

Fig. 4 displays the data utility experimental results based on data loss on the medical-patient database. The innovative technique in Fig. 4 had a protection rate (θ) of 5% using LPL and 95% using UPL. The assessment of the innovative technique, done through its comparison with hybrid [35], merging [17], $e-DP$ [39], UL method [3], and Mondrian [34] techniques revealed that the data utility obtained by the innovative technique was higher than that of all the known works. The merging technique had N fake tuples with the same QI values as in the original table. The sensitive values were assigned to them based on the sensitivity value distribution in the initial dataset. Hybrid, $e-DP$ and Mondrian techniques used the generalization procedure as a protection method. Therefore, these techniques resulted in more data loss than the innovative technique.The innovative technique employs selective generalization within the cell when satisfying the privacy requirements is essential; hence, more data utility is preserved.

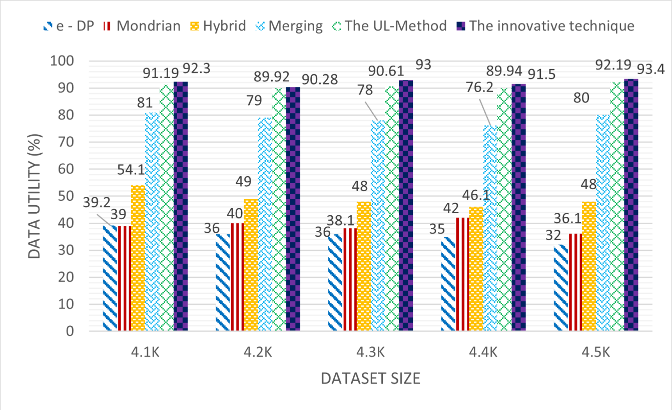

Fig. 4. Data Utility on the Medical-Patient Dataset (Protect Rate (θ) of 5% using LPL and 95% using UPL).

B. Measuring Risks

The measurement of disclosure risk in microdata during a composition attack is covered in this section. A composition attack occurs when an intrusive party, especially one knowledgeable about some of the QI values, attempts to identify a specific person in the microdata by linking several readily accessible records to an external database to disclose restricted information [8] [3]. As a result, gauging disclosure risk is essentially measuring the rareness of a cell in microdata publishing.

Medical patient dataset publishers should strive to measure the risk disclosure of PPDP outputs to ensure privacy preservation. This step is key in defining the level of protection needed. Therefore, differentiating the risk disclosure measures is important because the quantity to be measured must not depend on how the data representation method is selected. According to the works done previously, the risk disclosure can be quantified by determining the proportion of the true matches to the total matches, as expressed in Equation 5.

$$Disclosure\ risk\ ratio\ (DRR) = \frac{Matched\ records}{Total\ records} X100\% \quad (5)$$

Fig. 5 shows the experimental result for the disclosure risk ratio (DRR). DRR defines the adversary confidence level followed to elucidate the sensitive values in the medical-patient database. The $e-DP$ technique [39] revealed the least privacy risks compared to the innovative technique and other available approaches. The $e - DP$ technique achieved approximately 0.63% disclosure risk ratio privacy risk for medical-patient database when k = 6, l = 6 for size of 4.5K. Based on the proposed solution [39], it probabilistically generated a generalised possibility table and added noise to the total. The $e - DP$ offered high privacy assurance and protection opposed to composition attack by differential privacy grounded data anonymization [23] [3], as shown in the results. It was observed by [35] [40] [41] [17] [3] that using $e - DP$ to protect against composition attacks generates a significant amount of data utility losses during anonymization, confirming the result discussed in Fig. 4.

The hybrid technique [35] generated a lower probability of sourcing the end-user's private data than Mondrian technique [34] and merging technique [17]. The hybrid technique achieved approximately 1.55% disclosure risk ratio (privacy risk) for medical-patient database when the K= 6, l = 6 for size of 4.5K.

Compared to the innovative technique, the UL method and the innovative technique decreased the likelihood of composition attacks on the released medical patient datasets than the hybrid [35], merging [17], and Mondrian [34] techniques by disabling the unique attributes and high identical attribute values by UPL and LPL, and providing multiple matching cells in each equivalence class, leading to protection against identity disclosure. The UL method and the innovative techniques achieve approximately 1.5% disclosure risk ratio when the K= 6, l = 6 for size of 4.5K. Meanwhile, the innovative technique enhanced the data utility (Fig. 4) through the use of hybrid protection method. Increasing the false matches for a unique attribute in LPL and value swapping for high identical values in UPL helped to enhance the data utility and guarantee a lower risk of attribute disclosure.

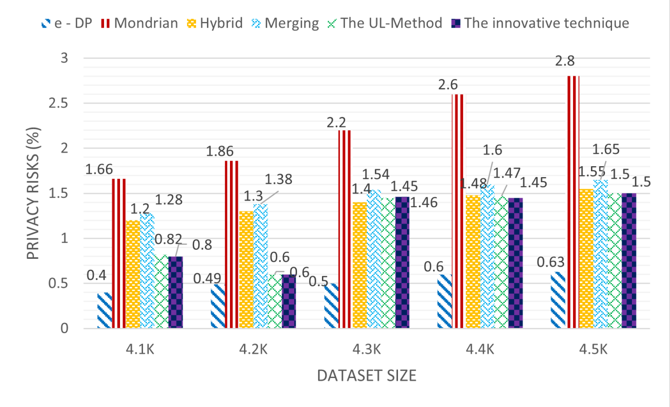

Fig. 5. Privacy Risk Medical-Patient-Database (k = 6, l = 6).

VI. Conclusion and Discussion

This paper presented an innovative technique using a hybrid protection method for utility enhancement while preserving data privacy for medical patient dataset; and for limiting the prospect of popular composition attack when the independent medical organizations cannot coordinate prior to medical patient dataset publication. The experiment showed that the innovative technique could satisfy the requirements of privacy after intersecting the separately published medical patient datasets. By contrast, many existing techniques reduced the published medical patient data utility due to the protection method used such as generalization and perturbation. This work, however, introduced a robust hybrid protection method by finding the unique attribute values (LPL) and high identical attribute values (UPL), then creating a fake tuple for the unique attribute and swapping values for the high identical attribute to decrease the attribute disclosure risk and ensure attainment of l-diverse in the published microdata table against the composition attacks. The model's effectiveness lies in the selection of specific attributes to enhance the privacy of published data and maintain data utility. The experimental findings revealed that the introduced method in this study could provide greater data utility compared with the state-of-the-art techniques. The achieved performance using our innovative technique helps researchers, decision-makers, and experts benefit from the published medical patient dataset to extract knowledge that may be used for disease prevention, medical decision-making, and many other areas of medical organizations. Similar to other scholarly research, this study leaves ample room for additional works to address its limitations and to expand upon its foundation to focus on adding or replacing another new protection methods to the innovative technique or extending some stages of the innovative technique to increase data utility and decrease risk disclosure. Moreover, the effectiveness of the innovative technique has been tested against composition attacks and background knowledge attacks, and by using the medical patient dataset; thus, it is important to test its performance against different attacks and by using different types of datasets.

Acknowledgments

This paper is part of a funded Project 'Developing an effective Privacy Protection Framework for Oman's current Healthcare Sector in Big data environment' from the Ministry of Higher Education, Research and Innovation (MoHERI/BFP/MC/01/2020), Sultanate of Oman in call 2020; and from UMP with vote no. UIC201518.

References

[1] M. Binjubeir, A. A. Ahmed, M. A. Bin Ismail, A. S. Sadiq, and M. Khurram Khan, "Comprehensive Survey on Big Data Privacy Protection," IEEE Access, vol. 8, pp. 20067–20079, 2020, doi: 10.1109/ACCESS.2019.2962368.

[2] M. Binjubeir, M. A. Ismail, S. Kasim, H. Amnur, and Defni, "Big healthcare data: Survey of challenges and privacy," International Journal on Informatics Visualization, vol. 4, no. 4, pp. 184–190, 2020, doi: 10.30630/joiv.4.4.246.

[3] M. BinJubier, M. Arfian Ismail, A. Ali Ahmed, and A. Safaa Sadiq, "Slicing-Based Enhanced Method for Privacy-Preserving in Publishing Big Data," Computers, Materials & Continua, vol. 72, no. 2, pp. 3665–3686, 2022, doi: 10.32604/cmc.2022.024663.

[4] J. M. Cavanillas, E. Curry, and W. Wahlster, New Horizons for a Data-Driven Economy. Cham: Springer International Publishing, 2016.

[5] Lei Xu, Chunxiao Jiang, Jian Wang, Jian Yuan, and Yong Ren, "Information Security in Big Data: Privacy and Data Mining," IEEE Access, vol. 2, pp. 1149–1176, 2014, doi: 10.1109/access.2014.2362522.

[6] T. Yu and S. Jajodia, "Secure Data Management in Decentralized Systems," Sushi1 Jajodia George Mason University US, 2007.

[7] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," Science, vol. 349, no. 6245, pp. 255–260, 2015.

[8] B.-C. Chen, D. Kifer, K. LeFevre, and A. Machanavajjhala, "Privacy-Preserving Data Publishing," Foundations and Trends® in Databases, vol. 2, no. 1–2, pp. 1–167, Jun. 2009, doi: 10.1561/1900000008.

[9] C. L. Philip Chen and C. Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," Information Sciences, vol. 275, pp. 314–347, 2014, doi: 10.1016/j.ins.2014.01.015.

[10] 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, 2007, pp. 106–115, doi: 10.1109/ICDE.2007.367856.

[11] Y. A. A. S. Aldeen, M. Salleh, and M. A. Razzaque, "A comprehensive review on privacy preserving data mining," SpringerPlus, vol. 4, no. 1, pp. 1–36, 2015, doi: 10.1186/s40064-015-1481-x.

[12] A. Sharma and N. Badal, "Literature Survey of Privacy Preserving Data Publishing (PPDP) Techniques," International Journal Of Engineering And Computer Science, vol. 6, no. 5, pp. 1–12, 2017, doi: 10.18535/ijecs/v6i4.12.

[13] P. R. Bhaladhare and D. C. Jinwala, "Novel approaches for privacy preserving data mining in k-anonymity model," Journal of Information Science and Engineering, vol. 32, no. 1, pp. 63–78, 2016.

[14] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," in Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on, 2006, p. 24.

[15] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "L -diversity," ACM Transactions on Knowledge Discovery from Data, vol. 1, no. 1, p. 3, Mar. 2007, doi: 10.1145/1217299.1217302.

[16] 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.

[17] A. Hasan, Q. Jiang, H. Chen, and S. Wang, "A New Approach to Privacy-Preserving Multiple Independent Data Publishing," Applied Sciences, vol. 8, no. 5, p. 783, May 2018, doi: 10.3390/app8050783.

[18] A. Gkoulalas-Divanis and G. Loukides, Medical Data Privacy Handbook. Cham: Springer International Publishing, 2015.

[19] B. C. M. Fung, K. Wang, A. W.-C. Fu, and J. Pei, "Anonymity for continuous data publishing," in Proceedings of the 11th international conference on Extending database technology: Advances in database technology, 2008, pp. 264–275.

[20] R. C.-W. Wong, A. W.-C. Fu, J. Liu, K. Wang, and Y. Xu, "Global privacy guarantee in serial data publishing," in 2010 IEEE 26th International Conference on Data Engineering (ICDE 2010), 2010, pp. 956–959.

[21] R. C.-W. Wong and A. W.-C. Fu, "Privacy-preserving data publishing: An overview," Synthesis Lectures on Data Management, vol. 2, no. 1, pp. 1–138, 2010, doi: https://doi.org/10.2200/S00237ED1V01Y201003DTM002.

[22] B. Malin and L. Sweeney, "How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems," Journal of Biomedical Informatics, vol. 37, no. 3, pp. 179–192, Jun. 2004, doi: 10.1016/j.jbi.2004.04.005.

[23] S. R. Ganta, S. P. Kasiviswanathan, and A. Smith, "Composition attacks and auxiliary information in data privacy," in Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08, 2008, p. 265, doi: 10.1145/1401890.1401926.

[24] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," ACM Computing Surveys (Csur), vol. 42, no. 4, pp. 1–53, 2010.

[25] T. A. Lasko and S. A. Vinterbo, "Spectral Anonymization of Data," IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 3, pp. 437–446, Mar. 2010, doi: 10.1109/TKDE.2009.88.

[26] S. E. Fienberg and J. McIntyre, "Data swapping: Variations on a theme by dalenius and reiss," in International Workshop on Privacy in Statistical Databases, 2004, pp. 14–29.

[27] R. Brand, "Microdata Protection through Noise Addition," in Inference control in statistical databases, Springer, 2002, pp. 97–116.

[28] C. K. Liew, U. J. Choi, and C. J. Liew, "A data distortion by probability distribution," ACM Transactions on Database Systems (TODS), vol. 10, no. 3, pp. 395–411, 1985.

[29] T. Li, N. Li, J. Zhang, and I. Molloy, "Slicing: A New Approach for Privacy Preserving Data Publishing," IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 3, pp. 561–574, Mar. 2012, doi: 10.1109/TKDE.2010.236.

[30] P. Mikalef, J. Krogstie, I. O. Pappas, and P. Pavlou, "Exploring the relationship between big data analytics capability and competitive performance: The mediating roles of dynamic and operational capabilities," Information & Management, vol. 57, no. 2, p. 103169, Mar. 2020, doi: 10.1016/j.im.2019.05.004.

[31] P. Samarati and L. Sweeney, "Generalizing data to provide anonymity when disclosing information," in PODS, 1998, vol. 98, p. 188, doi: 10.1145/275487.275508.

[32] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "L-Diversity: Privacy beyond k-Anonymity," ACM Trans. Knowl. Discov. Data, vol. 1, no. 1, pp. 3–es, Mar. 2007, doi: 10.1145/1217299.1217302.

[33] R. C.-W. Wong, J. Li, A. W.-C. Fu, and K. Wang, "(α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing," in Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, 2006, pp. 754–759, doi: https://doi.org/10.1145/1150402.1150499.

[34] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian Multidimensional K-Anonymity," in 22nd International Conference on Data Engineering (ICDE'06), 2006, pp. 25–25, doi: 10.1109/ICDE.2006.101.

[35] J. Li, M. M. Baig, A. H. M. Sarowar Sattar, X. Ding, J. Liu, and M. W. Vincent, "A hybrid approach to prevent composition attacks for independent data releases," Information Sciences, vol. 367–368, pp. 324–336, Nov. 2016, doi: 10.1016/j.ins.2016.05.009.

[36] R. Mendes and J. P. Vilela, "Privacy-Preserving Data Mining: Methods, Metrics, and Applications," IEEE Access, vol. 5, pp. 10562–10582, 2017, doi: 10.1109/ACCESS.2017.2706947.

[37] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, and Y. Theodoridis, "State-of-the-art in privacy preserving data mining," in SIGMOD Record, 2004, vol. 33, no. 1, pp. 50–57, doi: 10.1145/974121.974131.

[38] N. Zhang and W. Zhao, "Privacy-Preserving Data Mining Systems," ieee, vol. 40, no. 4, pp. 52–58, Apr. 2007, doi: 10.1109/MC.2007.142.

[39] N. Mohammed, R. Chen, B. C. M. Fung, and P. S. Yu, "Differentially private data release for data mining," in Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11, 2011, p. 493, doi: 10.1145/2020408.2020487.

[40] G. Cormode, C. M. Procopiuc, Entong Shen, D. Srivastava, and Ting Yu, "Empirical privacy and empirical utility of anonymized data," in 2013 IEEE 29th International Conference on Data Engineering Workshops (ICDEW), 2013, pp. 77–82, doi: 10.1109/ICDEW.2013.6547431.

[41] R. Sarathy and K. Muralidhar, "Evaluating Laplace noise addition to satisfy differential privacy for numeric data.," Trans. Data Priv., vol. 4, no. 1, pp. 1–17, 2011.

[42] A. S. M. T. Hasan, Q. Jiang, J. Luo, C. Li, and L. Chen, "An effective value swapping method for privacy preserving data publishing," Security and Communication Networks, vol. 9, no. 16, pp. 3219–3228, Nov. 2016, doi: 10.1002/sec.1527.

[43] A. Sharma, G. Singh, and S. Rehman, "A Review of Big Data Challenges and Preserving Privacy in Big Data," in Advances in Data and Information Sciences, Springer Nature Switzerland, 2020, pp. 57–65.

[44] S. Rohilla, "Privacy Preserving Data Publishing through Slicing," American Journal of Networks and Communications, vol. 4, no. 3, p. 45, 2015, doi: 10.11648/j.ajnc.s.2015040301.18.

[45] "mohd-akaber/Medical-Patient-Database." [Online]. Available: https://github.com/mohd-akaber/Medical-Patient-Database. [Accessed: 14-Jun-2022].

[46] R. Kohavi and B. Becker, "UMI Machine Learning Repository: Adult Data Set," Irvine, CA: University of California, School of Information and Computer Science., 2019. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Adult. [Accessed: 04-May-2020].

[47] L. Kaufman and P. J. Rousseeuw, Finding Groups in Data, vol. 344. Hoboken, NJ, USA: John Wiley & Sons, Inc., 1990.

[48] M.-Q. Tran et al., "Reliable Deep Learning and IoT-Based Monitoring System for Secure Computer Numerical Control Machines Against Cyber-Attacks With Experimental Verification," IEEE Access, vol. 10, pp. 23186–23197, 2022, doi: 10.1109/ACCESS.2022.3153471.

[49] A. H. M. S. Sattar, J. Li, J. Liu, R. Heatherly, and B. Malin, "A probabilistic approach to mitigate composition attacks on privacy in non-coordinated environments," Knowledge-Based Systems, vol. 67, pp. 361–372, Sep. 2014, doi: 10.1016/j.knosys.2014.04.019.

[50] A. Majeed and S. Lee, "Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey," IEEE Access, vol. 9, pp. 8512–8545, 2021, doi: 10.1109/ACCESS.2020.3045700.

[51] U. Sivarajah, M. M. Kamal, Z. Irani, and V. Weerakkody, "Critical analysis of Big Data challenges and analytical methods," Journal of Business Research, vol. 70, pp. 263–286, 2017, doi: 10.1016/j.jbusres.2016.08.001.

[52] C. Pu, H. Zerkle, A. Wall, S. Lim, K.-K. R. Choo, and I. Ahmed, "A Lightweight and Anonymous Authentication and Key Agreement Protocol for Wireless Body Area Networks," IEEE Internet of Things Journal, pp. 1–1, 2022, doi: 10.1109/JIOT.2022.3175756.

[53] A. Shah and R. Gulati, "Privacy Preserving Data Mining: Techniques, Classification and Implications-A Survey," International Journal of Computer Applications, vol. 137, no. 12, 2016.

[54] A. Senosi and G. Sibiya, "Classification and evaluation of privacy preserving data mining: a review," in 2017 IEEE AFRICON, 2017, pp. 849–855.

[55] Wang, Tao, Zhigao Zheng, Mubashir Husain Rehmani, Shihong Yao, and Zheng Huo. 2019. "Privacy Preservation in Big Data From the Communication Perspective—A Survey." IEEE Communications Surveys & Tutorials 21(1):753–78.

[56] C. C. Aggarwal and P. S. Yu, "A General Survey of Privacy-Preserving Data Mining Models and Algorithms," in Privacy-preserving data mining, Springer US, 2008, pp. 11–52.

[57] K. Chen and L. Liu, "Privacy preserving data classification with rotation perturbation," in Proceedings - IEEE International Conference on Data Mining, ICDM, 2005, pp. 589–592, doi: 10.1109/ICDM.2005.121.