IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 24, 2022, accepted March 6, 2022, date of publication March 10, 2022, date of current version March 18, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3158312

Heap Bucketization Anonymity–An Efficient Privacy-Preserving Data Publishing Model for Multiple Sensitive Attributes

J. JAYAPRADHA[1], (Student Member, IEEE), M. PRAKASH[2], YOUSEEF ALOTAIBI[3], OSAMAH IBRAHIM KHALAF[4], AND SALEH AHMED ALGHAMDI[5]

[1]Department of Computing Technologies, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu 603203, India
[2]Department of Data Science and Business Systems, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu 603203, India
[3]Department of Computer Science, College of Computer and Information Systems, Umm Al-Qura University, Makkah 21955, Saudi Arabia
[4]Al-Nahrain Nanorenewable Energy Research Center, Al-Nahrain University, Baghdad 10001, Iraq
[5]Department of Information Technology, College of Computers and Information Technology, Taif University, Taif 21944, Saudi Arabia

Corresponding author: J. Jayapradha (jayapraj@srmist.edu.in)

This work was supported by Taif University through the Taif University Researchers Supporting Project, Taif University, Taif, Saudi Arabia, under Grant TURSP-2020/313.

ABSTRACT The publication of a patient's dataset is essential for various medical investigations and decision-making. Currently, significant focus has been established to protect privacy during data publishing. The existing privacy models for multiple sensitive attributes do not concentrate on the correlation among the attributes, which in turn leads to much utility loss. An efficient model Heap Bucketization-anonymity (HBA) has been proposed to balance privacy and utility with multiple sensitive attributes. The Heap Bucketization-anonymity model used anatomization to vertically partition the dataset into 1. Quasi-identifier table and 2. Sensitive attribute table. The quasi-identifier is anonymized by implementing k-anonymity and slicing and the sensitive attributes are anonymized by applying slicing and Heap Bucketization. The metrics Normalized Certainty Penalty and KL-divergence have been used to compute the utility loss in the patient dataset. The experimental results show that the HB-anonymity can significantly achieve high privacy with less utility loss than other existing models. The HB-anonymity model not only balances the utility and privacy also eradicates the i) background knowledge attack, ii) quasi-identifier attack iii) membership attack, iv) non-membership attack and v) fingerprint correlation attack.

INDEX TERMS Privacy-preserving, anatomization, heap bucketization, Pearson correlation, k-anonymity, slicing, normalized certainty penalty and KL-divergence.

I. INTRODUCTION

Information is significant to the various innovations. To discover information, the data are retrieved and analyzed by the research community [1]. Public and private sectors examine the human behavior patterns to enhance their services. In the process of extracting knowledge, the individual's information is leaked and leads to privacy breaches. An adversary may use publically available data to gather individual information. Privacy is the foremost concern in all applications and sectors. Data is used for various purposes such as statistical analysis, knowledge discovery, policy-making, etc. Various organization, pharmacies and health sectors share their employee details and patient details to third parties for various analysis purposes. As the data grow tremendously, the analysis of the data becomes tedious. Thus to deal with big data, various approaches have been proposed [2, 3]. The lifecycle of data has different stages i) data creation ii) data storage, iii) pre-processing of data, iv) data archival and v) data purging. Existing techniques of privacy preservation are still in evolving stage and achieving the balance between privacy and utility is still an open issue in the research area.

Currently, the healthcare industry collects information about patients for a better, accurate diagnosis and treatment for the patients. Since the dataset consists of sensitive attributes, it needs to be anonymized. The healthcare

The associate editor coordinating the review of this manuscript and approving it for publication was Vivek Kumar Sehgal.

industry is the largest and currently developing area in research. It is shifting towards disease-oriented to patient-oriented approaches. Information and Communication Technology (ICT) is incorporated in health care practices. The volume of data in the health care industry is growing rapidly and the data are used for various analysis purposes. To achieve the best results from the data, the utility of the data needs to be maintained. There are many researches done in preserving privacy viz. privacy of the big data in the health care industry, privacy-preserving in the Internet of things (IoT), maintaining privacy in the cloud, Artificial intelligence in healthcare for maintaining privacy.

Various technologies have been used in the health care industry such as Machine learning [4]–[6], IoT systems [7]–[12], data analytics [13], [14], and cloud system [15]–[17]. For continuous monitoring of patients, wearable equipment has been introduced. The data recorded in the equipment is being continuously monitored, streamed, shared and analyzed to enable various health services to the patients [18]. Due to continuous monitoring, the patient can be diagnosed earlier and the proper action can be taken. Though all this technology improvises patient care, there arises a question "What about patient's privacy?"

Publishing the raw data will cause privacy breaches, which could lead to annoyance or deceit. With the released data, the intruder can cause heavy damage even to the status and life of individuals. Privacy-preserving has become an essential one when sharing data with researchers and third parties. Privacy-preserving data publishing (PPDP) delivers various methods, models and tools for protecting the breaches while publishing the data to third parties or analysts. Security and privacy are the major concerns in today's digital world. Recently, the PPDP has gained a lot of attention from the research community [19].

Earlier, researchers removed the explicit identifiers (e.g., name) considering the dataset is well protected. However, the intruders can easily infer the individual's sensitive attributes and complete details [20]. Thus, such measures seem to be insufficient because the individual record can be identified by relating it with the other external sources. Later, various anonymization techniques were proposed to mask the individual data viz. Generalization, suppression, permutation, encryption, role-built access control, etc.

In the proposed work, the privacy of the individual is protected from five breaches: (1) background knowledge attack (*bka*); (2) quasi-identifier attack (*qia*); (3) membership disclosure attack (*mda*); (4) non-membership disclosure attack (*n-mda*); and (5) fingerprint correlation attack (*fca*). The background knowledge of an individual can lead to identifying the pattern of that particular individual. Identifying the pattern of a particular individual by possessing background knowledge can lead to a fingerprint correlation attack. The multiple sensitive values of an individual are grouped to form the individual fingerprint. The correlation of the fingerprint among the various groups of k-anonymized can help the adversary to gain the insights of other individuals also in the dataset. The background knowledge paves a path for all the attacks such as quasi-identifier attack (*qia*) membership disclosure attack (*mda*) non-membership disclosure attack (*n-mda*) and fingerprint correlation attack (*fca*). The linking of individual quasi-identifier values helps the adversary to gain insights into personal information. Through the background knowledge and linking of qid values, if an adversary can find out the existence and non-existence of an individual in the microdata. Then, the membership and non-membership disclosure attacks persist. The fingerprint correlation attack is a strong privacy breach as it could snoop all the individual information in the dataset.

In the paper, Heap Bucketization –Anonymity model has been proposed and compared with two existing approaches (p,k)-angelization and (c,k) anonymization. In (p,k)-angelization, the sensitivity levels were fixed and that is represented as p and k represent the k-anonymous groups. The (p,k)-angelization is a strong approach and eradicated non-membership attack and membership attack. However, it could not resist *fca*. In (c,k)-anonymization, the *fca* was eradicated. However, the execution time is relatively high compared to HBA model.

The paper is systematized as below. Section II discusses the various literature works in privacy-preserving data publishing with 1:1 single sensitive attribute, 1:1 multiple sensitive attributes and 1:M micro data. Section III summarizes the problem definition and related preliminaries and their definitions. Section IV presents the details of the work contributed to the paper. Section V explains the motivation of proposed HB-anonymity which is an extension of (p,k)-Angelization and (c,k)-Anonymization. The (p,k)-Angelization and (c,k)-Anonymization definitions and the complete working of the models are elaborately discussed. Section VI discusses 1:1 microdata with multiple sensitive attribute attacks and their scenarios. Section VII discusses the proposed model Heap Bucketization-anonymity. Under this section, various steps involved in the model are clearly explained. Section VIII explains the implementation of slicing on the sensitive attributes table and the merging part of the quasi-identifier and sensitive attributes table. In addition, the workflow and framework of the Heap Bucketization-anonymity model are depicted clearly. Section IX gives a detailed step-wise algorithm for the Heap Bucketization-anonymity model. Section X describes the experimental details and results of the examination. The complete setup of the experimental, the outcomes and various utility metrics used in the experiments are discussed clearly. Furthermore, the experimental results are shown as graphs for a better understanding of the work. Finally, section XI concludes the work along with the future direction.

II. RELATED WORKS

The rapid growth of electronic health care systems and sharing of the data increases the need for privacy [21]. Due to sharing of data to third parties the protection of individual identity becomes a major challenge. However, the privacy of the microdata is set based on the well-defined procedures and policies for sharing the individual's health data. The Health Insurance Portability and Accountability Act proposed two methodologies to achieve de-duplication. The recent available privacy-preserving technologies have been analyzed and discussed [22].

A. 1:1 MICRO DATA WITH SINGLE SENSITIVE ATTRIBUTE

Samarati and Sweeney proposed k-anonymity. k-anonymity protects the dataset from record linkage and not every record of the table should be distinguishable from at least k-1 records. In k-anonymity, anonymization methods such as generalization and suppression are applied. It makes sure that the probability of re-identifying a person in the disclosed data must not be more than 1/k [23]. Another model called l-diversity was proposed which is an extended model of k-anonymity. It prevents attribute linkage l-diversity makes sure that there should be at least l different values for the sensitive attributes in every equivalence class [24].

In l-diversity, the skewness attack occurs due to the skewness of sensitive attributes in the overall distribution, so a model t-closeness was proposed. T-closeness ensures that sensitive attributes distribution in each class must be closer to the dissemination of sensitive attributes in the entire table [25]. Many researchers have proposed the extended version of k-anonymity, l-diversity and t-closeness. An extended version of the k-anonymity model was proposed with minimum data distortion for record suppression. The scalable l-diversity (ImSLD), an extended version of improved scalable k-anonymity (ImSKA) was proposed to handle a large amount of data. MapReduce has been used as a programming paradigm. The usage of MapReduce iteration reduced the running time consistently. To compute the utility loss in the anonymized dataset, a metric called Normalized Certainty Penalty (NCP) was used [26].

A new versatile publishing method was proposed with a set of privacy rules for the quasi-identifier and sensitive attributes. A Guardian Normal Form (GNF) was introduced for publishing each sub-table along with the existing publishing approach. When the sub tables are merged while publishing the entire table. Then, the privacy rules should be able to guarantee the utmost privacy. Two different algorithms: (1) Guardian Decomposition; and (2) Utility-aware decomposition was proposed to anonymize the microdata. The main focus of the work is to concentrate on the versatility problem of the privacy-preserving data publishing with various privacy rules incorporated for the anonymization of data [27].

Two privacy models: (1) enhanced identity-reserved l-diversity; and (2) enhanced identity (α, δ)-anonymity has been proposed. To implement the above two privacy models, the DAnonyIR generalization algorithm has been designed with a clustering technique to reduce information loss. The EIR l-diversity and EIR (α, β)-anonymity works well for the multiple sensitive attributes in static relational data [28]. The Sensitive Label Privacy Preservation with Anatomization (SLPPA) scheme has been proposed to protect the microdata. The scheme adopts two techniques i) table division and ii) group division. In the table division procedure, the mean-square contingency coefficient and entropy have been adopted for anonymization. In group division, non-overlapping groups have been framed to satisfy the $(\alpha, \beta, \gamma, \delta)$ model [29].

B. 1:1 MICRO DATA WITH MULTIPLE SENSITIVE ATTRIBUTE

The (α,l)-model was implemented to achieve proper diversity requirements for the dataset with multiple sensitive attributes. The two variables α and l confine the values of a sensitive attribute in the equivalence classes. The (α,l)-model is designed with k-anonymity as a foundation. The (α,l)-model has less running time and utility loss [30]. An addictive noise approach was proposed by satisfying the conditions of l-diversity [31]. A privacy-preserving data publishing method known as MNSACM was proposed to handle numerical attributes. The MNSACM method comprises of two approaches which are clustering and Multi-sensitive bucketization. A Two-dimensional bucket has been formed to anonymize the sensitive attributes. The MNSACM aims to publish a one-time static relational table [32].

A distribution model was proposed to fix the values of sensitive attributes. For multiple sensitive attribute values, a threshold p is set for minimizing the sensitive attributes disclosure probability [33]. A new framework (k, p) - anonymity was proposed to resolve the sensitive attributes disclosure problems in k-anonymity and l-diversity models [34]. A proficient approach (p, k)-Angelization has been proposed for anonymizing the dataset with MSA. The (p, k)-Angelization not only preserves privacy but also enhances the utility of the disclosed dataset [35].

Quasi-identifier-Multiple heterogeneous sensitive attribute (QI-MHSA) generalization algorithm was proposed to protect the privacy of the dataset with multiple sensitive attributes. k-anonymity has been applied for the quasi-identifier bucket and l-diversity on the sensitive attributes bucket. In addition, a flag has been set to generalize the sensitive attributes according to their sensitivity requirements [36].

Most of the researchers have separated the quasi-identifier and sensitive attribute from the microdata. Yuichi Sei [37] has adopted new privacy models $(l1, \ldots, lq)$-diversity and $(t1, \ldots, tq)$-closeness and stated that each attribute has a sensitive value in it. Thus, he categorized the quasi-identifier as sensitive QID. Two algorithms such as (1) anonymization; and (2) reconstruction algorithms were proposed to anonymize the sensitive QID's to achieve great privacy.

A PPDP for dynamic data with MSA was proposed and named KC slice [38]. An improvised version of the KC slice named KCi–Slice was proposed to balance the privacy and utility while publishing the dataset with MSA [39]. A model used an anonymization technique called slicing. It uses the fuzzy method for numerical sensitive attributes and the generalization method for categorical sensitive attributes [40].

C. 1:M MICRO DATA

Various models have been proposed for the 1: M dataset with MSA. A novel method called "MSAs generalization correlation attacks" was proposed for 1:M microdata for multiple sensitive attributes. An approach called (p,l)-angelization was proposed to anonymize the 1:M MSA dataset[41]. To preserve privacy in data publishing for 1:M microdata, a model known as G-model was proposed. G-model provides a proper balance between utility and privacy. It protects the 1:M microdata from gender precise sensitive attribute attacks [42]. An *f*-slip model has been proposed for 1:M microdata. It eradicates various attacks such as bk attacks, MSAcorr attacks, QIcorr, NMcorr and Mcorr attacks. A unique approach frequency-slip was adopted to preserve privacy [43]. Various methods have been adopted in relational data [44], [45].

III. PROBLEM DESCRIPTION AND PRELIMINARIES

A. PROBLEM DEFINITION

Let the dataset T_P, consist of multiple sensitive attributes. Can the dataset be anonymized in such a way that the intruder should not get any clue about the individual? The anonymization of data should ensure the optimal balance between utility and privacy. For the eases of consequent discussion, the basic notions and descriptions of the paper are presented briefly.

B. BASIC NOTIONS AND DESCRIPTIONS

The patient microdata is presented in a relational table. The table T_{r*c} consists of r which represents the rows and c that represents the columns. The microdata can be categorized as below:

Direct identifier, is a unique identifier such as social security number, driving license and name. The direct identifiers will be encrypted or removed before disclosing them to the third party.

Quasi-identifier (T_P^{QI}), is a group of attributes used to detect the individual by relating it with external sources. The QI are age, sex, height and weight in Table 1.

Sensitive attribute (T_P^{SA}), possesses secretive information of the individuals, which needs to be secured during the disclosure of microdata such as Disease, Pulse rate, etc in Table 1. The focus of the paper is to protect the sensitive attribute from being revealed with less utility loss and high privacy.

Definition 1 (Equivalence Class [46]): In the 1:1 microdata T_P, the records of the same quasi-identifier values constitute the equivalence class (i.e) the *n* subset of T_P comprises records that correspond to each other. If the tuple tp ϵ $T_{P,}$ then the generalized form of table T_P that comprises of tuple tp is represented in the form:

$$(Gf_i[1], Gf_i[2], Gf_i[3], \ldots\ldots..Gf_i[1], tp[T_P^{SA}]),$$

where $Gf_i (1 \le i \le n)$ is the unique quasi-identifier subset including tp. $Gf_i[j]$ $(1 \le j \le r)$ is the generalized value of the record on T_P for all the records in Gf_i and T_P^{SA} the sensitive attributes in the table $T_{P.}$

Definition 2 (k-Anonymity): The dataset T_P satisfy k-anonymity if the record of every individual should not be eminent from at least k-1 individual records whose record also exists in the dataset T_P.

Definition 3 (Slicing [47]): The dataset T_P is partitioned both vertically and horizontally. Vertical partition groups the attributes based on the high correlations among the attributes. Each column comprises a subset of highly correlated attributes.

There may be columns $a_1, a_2, \ldots, a_n$ (i.e.) $\bigcup_{i=1}^{a} a_i = C$ so for any $1 \le j_1 \ne j_2 \le a, a_{j1} \cap a_{j2} \equiv \emptyset$.

The horizontal partition groups the tuples into different buckets. Each tuple can belong to the only bucket. Consider the bucket B^{id} and the number of buckets $b_1^{id,}$ $b_2^{id}, \ldots\ldots.., b_m^{id}$ then $\bigcup_{i=1}^{b^{id}} B_i^{id} = T^P$ so $1 \le j_1 \ne j_2 \le b^{id}$,

$$B_{j1}^{id} \cap B_{j2}^{id} = \emptyset$$

Definition 4 (Bucketization [46]): The dataset T_P has been partitioned into *n* quasi-identifier groups and m groups of sensitive attributes. The subset of tuples in the partitioned table is called a bucket and represented in the form:

T_P^{QI} (QI, B^{id}) and T_P^{SA} (SA, B^{id})

The QI and SA are the quasi-identifier of the table T_P and the sensitive attributes of the table T_P. The B^{id} represents the bucket id.

Definition 5 (Heap Bucketization): Heap Bucketization is an advancement of bucketization. The dataset T_P has been partitioned into *n* quasi-identifier groups and m groups of sensitive attributes. The sensitive attributes of each record in the same bucket are cumulated and represented as the records of the same bucket.

C. WHY THE ADVANCEMENT OF BUCKETIZATION IS NEEDED?

Angel [48] and Anatomy [49] have implemented bucketization to preserve privacy in data publishing. However, Angel and Anatomy have been implemented on the single sensitive attribute. In (p,k)-angelization and (c,k)-anonymization, bucketization has been adopted for MSA. The (p,k)-angelization lead to high utility loss and (c,k)-anonymization lead to high execution time. So, an advancement of bucketization named "Heap Bucketization" is proposed to prevent higher utility loss and privacy loss.

IV. CONTRIBUTION

An efficient Heap Bucketization-Anonymity (HB) model was proposed to protect privacy in data publishing with MSA. The table is anonymized using the HB-anonymity approach to achieve an optimal balance between privacy and utility.

1. A unique privacy-preserving data-publishing model "Heap Bucketization"-anonymity has been proposed which can have an optimal balance between privacy and utility. The HB-Anonymity is framed for multiple sensitive attributes to achieve high privacy with less loss of information.

TABLE 1. Sample of patient data T_P.

	Quasi-identifier				Sensitive Attributes				
Pid	Age	Sex	Height (Ht)cm	Weight (Wt)	Temperature	Pulse Rate(PR)	Respiratory Rate(RT)	Blood Pressure	Disease
0	54	M	170	87	97.2	92	23	111/67	Allergic Rhinitis
1	51	F	174	85	98	95	22	130/60	Anaemia
2	49	M	176	84	98	93	28	130/80	Eye Disorders
3	60	M	168	72	98	100	22	120/80	Back pain
4	37	F	151	68	98.6	76	22	110/70	Costochondritis
5	24	F	162	71.4	97.6	82	18	120/70	Gastritis
6	40	F	154	85	98	84	19	210/100	Diabetes
7	58	F	156	55	97.8	86	18	140/80	Allergic Rhinitis
8	62	M	160	60	97	94	22	110/80	Fissure In Anal
9	19	F	156	46	97	88	22	114/80	Hair Fall
10	42	F	158	45	98	89	24	144/94	Diabetes
11	40	M	168	79	99	91	20	80/50	Hemorrhoids

An algorithm has also been framed for Heap Bucketzation-Anonymity.

2. The method has been evaluated both theoretically and experimentally to validate the proposed model. The proposed HB-Anonymity model prevents privacy under i) background knowledge attack, ii) quasi-identifier attack iii) membership attack, iv) non-membership attack and v) fingerprint correlation attack.

V. MOTIVATION OF PROPOSING HB-ANONYMITY—AN EXTENSION OF (P,K)—ANGELIZATION AND (C.K)—ANONYMIZATION

Earlier many models dealt with a single sensitive attribute. However, in real case scenarios, the health records might have multiple sensitive attributes. The health record comprises various attributes that are sensitive such as disease, temperature, etc. as shown in Table 1. Handling those sensitive attributes and maintaining the privacy of the individuals is not an easy task. In PPDP, the privacy and utility need to be balanced so that the researchers can make analysis and decision-making. If the utility is not preserved along with privacy, then the researchers would not be able to analyze and extract valuable information.

Definition 6 ((p,k)-Angelization [35]): The relation data Tp is said to be (p,k)- angelization, if the table is partitioned into category table, quasi-identifier table and sensitive attribute table. The anonymized table is published in two different batches i.e quasi-identifier table and sensitive attribute table. The p represents the category of sensitivity levels and k represents the group of k-anonymous data. The maximum weighted attribute is calculated using a weighted function as the (p,k)- angelization considers the maximum weighted attribute as the most sensitive attribute.

In (p,k)- angelization, the privacy breach is initiated with a highly weighted attribute and the values of the quasi-identifier table and the sensitive attribute table are correlated with the batch id. The (p,k)- angelization is an iterative process that failed in preventing the record re-identification of the individual with his/her complete details. Through an iterative process, the adversary can be able to obtain the details of the other individual as well in the dataset. In the (p,k)-angelization, the intersection of attribute values in two different buckets results in single sensitive attribute values against each sensitive attribute. The intersection value of an individual for all the attributes has been carried out to find the complete details of an individual.

Due to the iterative process, though the identification of the individual is complex, it leads to a privacy breach with the intersection of attribute values between two buckets. As the (p,k)- angelization did not completely utilize the angelization mechanism, the splitting of the patient table into two sub-tables: (1) Quasi-identifier table; and (2) Sensitive attribute table is useless. When the intruder finds the batch id of an individual in the generalized table he can easily infer the sensitive details of the individuals in SBT by correlating with the batch id, thus the splitting of the table into two is useless. In (p,k)- angelization, the weight of attributes is calculated to identify the highly sensitive attribute which is more likely to cause a privacy breach. To compare the (p,k)-angelization with HBA model, the attributes age, sex, height and weight from the generalized table, the attributes temperature, pulse rate, respiratory rate, blood pressure and disease forms the sensitive batch table. In our experiment work, the sensitivity level p = 4 (i.e) (Very High, High, Medium and Less). The value of k = 3 to anonymize the generalized table (i.e) quasi-identifier table.

Definition 7 ((c.k)-Anonymization [50]): A table Tp is said to be (c,k)-anonymization if the table comprises a generalized table and fingerprint bucket of sensitive attributes. The generalized table comprises quasi-identifier and k-anonymized with bucket id to prevent the linking attack. The bucket id in the quasi-identifier table is linked with the bucket id in the sensitive table. However, the sensitive table comprises of c varied records to avoid fingerprint correlation attacks. The c represents the category of the table. In the proposed work, c = 4(Very High, High, Medium and Less) and k = 3(i.e) anonymized group of data. In (p,k)- angelization and (c,k)-anonymization, the weights of the sensitive attributes are calculated to find out the highly sensitive attribute.

In (c,k)-anonymization, the finger bucket is created that satisfies the c-diversity to prevent an attack such as fingerprint correlation. The c-diversity is in the form of l-diversity in (c,k)-anonymization. The disadvantage in (p,k)-angelization has been overcome in (c,k)-anonymization by considering the two factors: 1. Minimizing the linking of records between two fingerprint buckets, 2. Un correlating the records between the fingerprint buckets. A linkability control factor (c_f) has been introduced to minimize the repetition of the same value of the attribute in the fingerprint bucket.

The goal of HB-anonymity is to provide sustainable privacy and less utility loss. In HB-anonymity, the correlation between the QI and the SA is computed using the Pearson correlation coefficient. The association of the QI and the SA of the table is calculated for slicing the highly correlated attributes. The purpose of slicing highly correlated attributes is to minimize information loss. If the correlated attributes are not connected, then the information distribution will be scattered and the researchers from the data that is anonymized cannot gain valuable information.

In the HB-anonymity model, the two tables are produced: (1) Quasi-identifier table ($\mathrm{TP^{QI}}$); and (2) Sensitive attribute table ($\mathrm{TP^{SA}}$). The attribute generalization is carried only in the quasi-identifier table and not in the sensitive attribute table to minimize the utility loss. The QI table is split into two tables $\mathrm{TP^{QI1}}$ and $\mathrm{TP^{QI2}}$ based on the correlation. As the k-anonymity cannot prevent attribute disclosure, it cannot protect the sensitive attribute effectively so, only the quasi-identifier is k-anonymized. The slicing is performed on the dataset to preserve the data utility and to protect the dataset against the membership disclosure and attribute disclosure attack as the generalization and bucketization may lead to membership disclosure attacks. The weights of the sensitive attributes and the iterative processes are not performed in HB- anonymity that increases the time complexity. In HB-anonymity, the buckets are formed with the increasing order of disease. (i.e.) alphabetically sorted.

Due to the slicing of highly correlated attributes, the data utility is highly preserved in the sensitive attribute table and the quasi-identifier table. Finally, Heap Bucketization is performed on the bucketized table to preserve privacy. As the HB-anonymity releases the single anonymized table, the linking of batch id is avoided. In the process of Heap Bucketization, all the records of a single bucket are combined to form the heap bucket. The heap bucket consists of sensitive details of the individuals.

In the proposed work, the privacy loss was checked by varying the bucket size ($\mathcal{B}_s$). If the bucket size is large, then the records of the heap bucket will be high and due to that, the loss of utility is also high. To minimize the utility loss, the $\mathcal{B}_s$ should be less. As the heap bucket consists of all the individual records of each bucket(i.e,)all the records of bucket 1 are comprised to form heap bucket1, the possibility of identifying an individual is almost zero. In Heap Bucketization, the probability of the distribution of the records will be high. Even the intruder knows the background details of the individual; the probability of identifying his record is close to zero.

TABLE 2. Quasi-identifier table of (p,k)-angelization.

Pid	Age	Sex	Height (Ht)cm	Weight (Wt)	Batch id
4,5,9	19-37	Person Person Person	150-162	75-88	1
6,10,11	40-45	Person Person Person	154-168	45-85	2
0,1,2	48-55	Person Person Person	170-176	80-90	3
3,7,8	55-65	Person Person Person	160-170	55-75	4

The proposed HB-Anonymity model prevents privacy under i) background knowledge attack, ii) quasi-identifier attack iii) membership disclosure attack, iv) non-membership disclosure attack and v) fingerprint correlation attack. In the heap bucketized anonymized data, the background knowledge attack (*bka*) cannot be accomplished since the intruder cannot gain any individual details even if the intruder has strong background knowledge of the individual. The linking of quasi-identifier cannot provide any information to the intruder as the QID are k-anoymized. The membership disclosure attack (*mda*) and non-membership disclosure attack (*n-mda*) cannot be accomplished as the existence and non-existence of an individual cannot be recognized in the proposed model due to the heap records of the buckets. The individual records from the buckets of the sensitive attributes cannot be identified at any cost, so the fingerprint correlation attack *(fca)* is also eradicated.

VI. 1:1 MICRODATA WITH MULTIPLE SENSITIVE ATTRIBUTE ATTACKS AND THEIR SCENARIOS

The Heap-Bucketization-anonymity anonymizes the dataset to protect it from five attacks to achieve high privacy and less information loss. The five attacks: (1) background knowledge attack (*bka*); (2) quasi-identifier attack (*qia*); (3) membership disclosure attack(*mda*); (4) non-membership disclosure attack (*n-mda*); and (5) fingerprint correlation attack *(fca)*. In the paper, the HB-Anonymity has been compared with two models (p,k)-Angelization and (c,k)-anonymization. For explaining the scenarios of all five attacks, the original patient table has been anonymized using (p,k)-angelization in Tables 2 and 3. The case scenario discusses that (p,k)-angelization could not resist the five attacks.

TABLE 3. Sensitive table of (p,k)-angelization.

Pid	Temperature	Pulse Rate(PR)	Respiratory Rate(RT)	Blood Pressure	Disease	Batch Id
4,5,9	98.6, 97.6,97	76,82,88	22,18,22	110/70,120/70,114/80	Costochondritis, Gastritis, Hair Fall	1
6,10,11	98,98,99	84,89,91	19,24,20	210/100,144/94,80/50	Diabetes, Diabetes, Hemorrhoids	2
0,1,2	97.2,98,98	92,95,93	23,22,28	111/67,130/60,130/80	Allergic Rhinitis, Anaemia, Eye Disorders	3
3,7,8	98,97.8,97	100,86,94	22,18,22	120/80,140/80,110/80	Back pain, Allergic Rhinitis, Fissure In Anal	4

A. SCENARIO 1

Scenario 1 discusses the background knowledge attack (*bka*). If an intruder can infer the sensitive information of individuals by possessing strong background knowledge. Then, *bka* can be accomplished. If the intruder knows that individual pid2, is a male, age < 50, with medium height and weight is suffering from some eye problem. Then, he can easily infer that the individual pid2 falls in bucket 3 in Table 2 and 3. If the intruder has strong background knowledge about pid2, then he can also conclude that individual pid2 does not suffer from allergy and anemic so the intruder confirms that the individual suffers from an eye disorder.

B. SCENARIO 2

Scenario 2 discusses the quasi-identifier attack (*qia).* If the intruder has strong background knowledge about the quasi-identifier values of the individual, then the intruder can correlate the quasi-identifier values to identify the sensitive attribute values. If the intruder knows that individual pid1, is a Female age > 50, with height > 170 and overweight, then the intruder can find the record in bucket 3 in Table 2 and3. If the intruder has strong background knowledge that the individual falls sick often, the intruder can conclude that the individual falls in bucket 3 and the disease is Anaemia.

C. SCENARIO 3

If the intruder possesses the individual background knowledge and quasi-identifier values. Then, the intruder can easily infer whether the individual is present in the dataset. If the intruder knows that pid7 is a female, age around 60, with height and weight around 155 and 50 respectively, and has enough knowledge that the individual does not suffer from severe disease. However, often sneezes and nose block, the intruder can conclude that pid7 falls in bucket 4 in Table 2 and 3.

D. SCENARIO 4

If the intruder possesses the background knowledge and quasi-identifier values of an individual. Then, the intruder can infer whether the individual exists in the dataset or not. The main aim of this non-membership disclosure attack is to find the non-existence of the individual. If the individual pid12, is male age > 75, suffering from severe pandemic disease. Then, the intruder can easily infer the non-existence of the individual as it does not lie in any of the buckets in Table 2 and 3.

E. SCENARIO 5

Scenario 5 discusses the fingerprint correlation attack. When two buckets are intersected, the unique sensitive values are derived from them and that helps in identifying the individuals. For example, if buckets 3 and 4 are intersected, temperature = 98, pulse rate = 22, Disease = Allergic Rhinitis are the sensitive values that can be uniquely identified. Because of this privacy breach, not only pid0 and pid7 are identified, the sensitive values of individual's pid1.pid2, pid3 and pid8 can also be identified. The definitions of the five attacks have been discussed in Table 4.

VII. HEAP BUCKETIZATION-ANONYMITY

Heap Bucketization-anonymity model has been proposed by designing architecture and algorithm. Various privacy-preserving models have been designed and proposed to carry out anonymization for multiple sensitive attributes in a 1:1 dataset. However, achieving the optimal balance between the privacy and utility challenge remains open. The proposed HB-anonymity model resists various attacks such as i) background knowledge attack, ii) quasi-identifier attack iii) membership attack, iv) non-membership attack and v) fingerprint correlation attack.

The goal of the proposed model is to achieve intensified privacy with less information loss. The HB-anonymity model performs the below steps i) pre-processing of the data, ii) anatomization of the table into T_P^{QI} and T_P^{SA} iii) calculating the correlation separately for both T_P^{QI} and T_P^{SA} iv) employing k-anonymity on T_P^{QI} and slicing v) implementation of slicing on T_P^{SA} vi) merging of T_P^{QI} and T_P^{SA} and vii) Heap Bucketization.

A. PRE-PROCESSING AND ANATOMIZATION OF THE TABLE

The real-time and unique dataset is used in the experimental work. As the dataset is received from the Interdisciplinary Institute of Indian System of Medicine, Ayurveda, the data is already pre-processed and in relational format. Few missing values of the attributes are filled by taking the average of the column values. In the HB-anonymity, the patient table is anatomized into two different tables i) quasi-identifier table T_P^{QI} and ii) sensitive attribute sub-table T_P^{SA}. The anatomization is performed to disconnect the relationship between sensitive attributes and quasi-identifier. The goal of anatomization is

TABLE 4. **1:1 Multiple sensitive attribute attacks.**

1:1 Multiple Sensitive Attribute Attacks	Description	Scenario
Background Knowledge Attack (*bka*)	If an intruder has strong background knowledge of an individual, then the background knowledge can be accomplished.	1
Quasi-Identifier Attack(*qia*)	If an intruder has the knowledge of the individual's quasi-identifier and has background knowledge of the individual, then the intruder can easily accomplish the quasi-identifier attack.	1 and 2
Membership Disclosure Attack (*mda*)	If an intruder can find the presence of an individual in the dataset, then the intruder can perform a membership disclosure attack.	1,2 and 3.
Non-Membership Disclosure Attack(*n-mda*)	If an intruder can find that the particular individual does not exist in the data set, then the intruder can perform a non-membership disclosure attack.	1,2 and 4
Fingerprint Correlation Attack (*fca*)	An intruder can find a distinct individual by intersecting the sensitive values of two buckets.	5

to apply different methods to the partitioned sub-table. Both $\mathrm{T_P^{QI}}$ and $\mathrm{T_P^{SA}}$ are allocated with a pid just for future reference. The pid will be eradicated during the publication of the table.

B. CORRELATION AMONG THE ATTRIBUTES

In the HB-anonymity model, the correlation of the attributes in both $\mathrm{T_P^{QI}}$ and $\mathrm{T_P^{SA}}$ is calculated. The purpose of finding correlation among the attributes in the HB-anonymity is to perform slicing. If slicing of the attributes is done randomly, then the linking relationship between the attributes will be broken and thus lead to utility loss. In Table 1, age, sex, height and weight are the quasi-identifiers. Temperature, pulse rate, respiratory rate, BP (further broken into systolic and diastolic) and disease are the sensitive attributes. Pearson correlation coefficient metric is used for computing the correlation among the sensitive attributes and quasi-identifier. A correlation matrix was generated to find the highest correlated attributes.

$$CorrMat\,(A,B) = \begin{cases} \rho cc_{A,B} & if\ A \neq B \\ 1, & otherwise \end{cases} \tag{1}$$

CorrMat = Correlation Matrix.

CorrMat(A, B) denotes the correlation coefficient among the two attribute A,B.

Pcc = Pearson Correlation Coefficient.

The Pearson correlation coefficient between the attribute A and B are calculated as below:

$$\rho cc\,(A,B) = \frac{covariance(A,B)}{sd\,(A)\,sd(B)} \tag{2}$$

$$Covariance\,(A,B) = EXP[(A - EXP\,[A])\,(B - EXP\,[B])] \tag{3}$$

where EXP [A] = Expected values of A and EXP [B] = Expected values of B.

$$std\,(A) = \sqrt{\frac{1}{N}\sum\nolimits_{x=1}^{N}(a_i - M_n)^2} \tag{4}$$

$$std\,(B) = \sqrt{\frac{1}{N}\sum\nolimits_{x=1}^{N}(b_i - M_n)^2} \tag{5}$$

M_n = mean value.

$$M_n = \frac{\sum x_i}{n} \tag{6}$$

As per the correlation metrics, the (age, sex) and (height, weight) are highly correlated in quasi-identifier. The (sys, dys), (Pulse rate, disease) and (respiratory rate, temperature) are highly correlated in the sensitive attribute table.

C. K-ANONYMITY AND SLICING ON THE QUASI-IDENTIFIER TABLE

The raw patient microdata is anatomized into two sub-tables i) quasi-identifier table $\mathrm{T_P^{QI}}$ and ii) Sensitive attribute table $\mathrm{T_P^{SA}}$. The quasi-identifier table is further divided into two tables based on the correlation. As per the Pearson correlation coefficient, age and sex form the first sub-table and the height and weight form the second sub-table. The quasi-identifier is anatomized into two tables to reduce the utility loss. After anatomization, k-anonymity (def.2) was implemented. 3-anonymity has been implemented separately on $\mathrm{T_P^{QI1}}$ and $\mathrm{T_P^{QI2}}$ as shown in Tables 5 and 6. The numerical attributes height, weight and age are substituted with the mean value of the particular equivalence class (def.1)[51] as shown in equation 7.

$$Mean = \frac{\mathrm{QI}_{11} + \mathrm{QI}_{12} + \mathrm{QI}_{13}}{n} \tag{7}$$

TABLE 5. 3-Anonymity on T_P^{QI1}.

PId	(Ht,Wt)	(Age, Sex)
0	(173.3, 85.3)	(51,M)
1	(173.3, 85.3)	(51,F)
2	(173.3, 85.3)	(51,M)
3	(160.3, 70.4)	(40.3,M)
4	(160.3, 70.4)	(40.3,F)
5	(160.3, 70.4)	(40.3,F)
6	(156.6, 66.6)	(53.3,F)
7	(156.6, 66.6)	(53.3,F)
8	(156.6, 66.6)	(53.3,M)
9	160.6,56.6	(33.6,F)
10	160.6,56.6	(33.6,F)
11	160.6,56.6	(33.6,M)

The QI_{11}, QI_{12} and QI_{13} are the quasi-identifier values of the attributes in each equivalence class and n is the total number of values in each equivalence class. For example, the values of the age attribute in the first equivalence class (ec) are 54, 51 and 49 then the first equivalence class is replaced with the mean value.

$$Mean\left(EC1_{Age}\right) = \frac{54+51+49}{3} = 51.3 \tag{8}$$

The mean $(EC1_{Age})$ is the mean value of the attribute age in the first equivalence class. Similarly, other equivalence class values are replaced with the mean value.

$$Mean\left(EC1_{Height}\right) = \frac{170+174+176}{3} = 173.3 \tag{9}$$

The mean $(EC1_{Height})$ represents the mean value of the attribute height in the first equivalence class.

$$Mean\left(EC1_{Weight}\right) = \frac{87+85+84}{3} = 85.3 \tag{10}$$

The mean $(EC1_{Weight})$ represents the mean value of the attribute weight in the first equivalence class.

Equations 8, 9 and 10 show the sample calculation of the mean value for the attributes age, height and weight (i.e.) data Perturbation [52]. The attribute sex is not generalized. In the HB-anonymity, generalization hierarchy trees are not adopted in the process of anonymization. Hence, the amount of utility loss is very less.

After implementing k-anonymity on the T_P^{QI1} and T_P^{QI2}, both the tables are merged to implement slicing as shown in Tables 7 and 8. The anatomization of tables, based on correlation reduces the loss of utility and the slicing helps in preserving the utility and correlation among the attributes. The slicing with the principle of k-anonymity prevents various attacks such as non-membership disclosure, membership disclosure, quasi-identifier attack and background knowledge attack.

TABLE 6. 3-Anonymity on T_P^{QI2}.

PId	Age	Sex
0	51.3	M
1	51.3	F
2	51.3	M
3	40.3	M
4	40.3	F
5	40.3	F
6	53.3	F
7	53.3	F
8	53.3	M
9	33.6	F
10	33.6	F
11	33.6	M

TABLE 7. Merging of T_P^{QI1} and T_P^{QI2}.

PId	Height (Ht)cm	Weight (Wt)	Age	Sex
0	173.3	85.3	51	M
1	173.3	85.3	51	F
2	173.3	85.3	51	M
3	160.3	70.4	40.3	M
4	160.3	70.4	40.3	F
5	160.3	70.4	40.3	F
6	156.6	66.6	53.3	F
7	156.6	66.6	53.3	F
8	156.6	66.6	53.3	M
9	160.6	56.6	33.6	F
10	160.6	56.6	33.6	F
11	160.6	56.6	33.6	M

TABLE 8. Slicing of highly correlated attributes.

PId	Height (Ht)cm	Weight (Wt)
0	173.3	85.3
1	173.3	85.3
2	173.3	85.3
3	160.3	70.4
4	160.3	70.4
5	160.3	70.4
6	156.6	66.6
7	156.6	66.6
8	156.6	66.6
9	160.6	56.6
10	160.6	56.6
11	160.6	56.6

VIII. IMPLEMENTATION OF SLICING ON TP^{SA} AND MERGING OF TP^{QI} AND TP^{SA}

Considering the utility loss caused by the anonymization process, the HB-anonymity does not adopt any type of hierarchy generalization or suppression. The patient dataset has six sensitive attributes temperature, pulse rate, respiratory rate, blood pressure and disease. The blood pressure comprises systolic and diastolic so the attribute BP is broken into two parts as shown in Table 9.

To anonymize the sensitive attributes of the dataset, HB-anonymity forms the buckets by sorting the values of diseases (i.e.) alphabetically sorted. Four buckets are formed in the sample dataset comprising of three records. After the formation of buckets, slicing has been performed on T_P^{SA} as shown in Table 10. The vertical slicing is performed based on the highly correlated attributes. The correlation matrix has been computed using the Pearson correlation coefficient to perform slicing. As per the correlation matrix, Pulse Rate (PR) and disease belongs to sl_1, Respiratory Rate (RR).

Temperature belong to sl_2, and Sys and Dys belongs to sl_3 ([PR, Disease ϵ sl_1], [RR, Temperature ϵ sl_2] and [Sys, Dys ϵ sl_3]). To anonymize the sensitive attributes by forming Heap

TABLE 9. Sensitive attribute table T_P^{SA}.

PId	Temperature	Pulse Rate (PR)	Respiratory Rate(RT)	Sys	Dys	Disease
0	97.2	92	23	111	67	Allergic Rhinitis
1	98	95	22	130	60	Anaemia
2	98	93	28	130	80	Eye Disorders
3	98	100	22	120	80	Back pain
4	98.6	76	22	110	70	Costochondritis
5	97.6	82	18	120	70	Gastritis
6	98	84	19	210	100	Diabetes
7	97.8	86	18	140	80	Allergic Rhinitis
8	97	94	22	110	80	Fissure In Anal
9	97	88	22	114	80	Hair Fall
10	98	89	24	144	94	Diabetes
11	99	91	20	80	50	Hemorrhoids

TABLE 10. Performing slicing on T_P^{SA}.

PId	(PR, Disease)	(RR, Temp)	(Sys,Dys)	Bucket
0	(92, Allergic Rhinitis)	(23, 97.2)	(111,67)	1
7	(86, Allergic Rhinitis)	(18,97.8)	(140,80)	1
1	(95, Anemia)	(22,98)	(130,60)	1
3	(100, Backpain)	(22,98)	(120,80)	2
4	(76, Costochondritis)	(22,98.6)	(110,70)	2
6	(84,Diabetes)	(19,98)	(210,100)	2
10	(89, Diabetes)	(24,98)	(144,94)	3
2	(93, Eye Disorders)	(28,98)	(130,80)	3
8	(94, Fissure In Ano)	(22,97)	(110,80)	3
5	(82, Gastritis)	(18,97.6)	(120,70)	4
9	(88, Hair Fall)	(22,97)	(114,80)	4
11	(91, Helmorrhoils)	(20,99)	(80,50)	4

Bucketization, the merging of T_P^{QI} and T_P^{SA} is done as shown in Table.11.

Heap Bucketization is formed by combining the records of each bucket as shown in Table 12. All three tuples comprise three individual records from bucket 1 itself. When an intruder tries to infer an individual record, he would not be able to identify even the buckets where the record is located as the generalized quasi-identifier is also distributed and the bucket is formed based on the disease. Finally, the anonymized data is released by sorting it according to the id of the patient.

The main goal of the proposed HB-Anonymity model is to perform Heap Bucketization. In bucketization, the quasi-identifier and the sensitive attributes are separated and the sensitive attribute values are randomly anonymized. In bucketization, the quasi-identifier values are published in the original form and thus it fails to protect the membership disclosure. Bucketization needs a clear parting of QI and SA values that might lead to the breaking of correlation among the quasi-identifier and sensitive attributes. Due to this breaking of the linking relationship, the utility loss will be high [53].

TABLE 11. Merging of T_P^{QI} and T_P^{SA}.

PId	(Age, Sex)	(Ht,Wt)	(PR, Disease)	(RR, Temp)	(Sys,Dys)	Bucket
0	(51,M)	(173.3, 85.3)	(92, Allergic Rhinitis)	(23, 97.2)	(111,67)	1
1	(51,F)	(173.3, 85.3)	(95, Anaemia)	(22,98)	(130,60)	1
2	(51,M)	(173.3, 85.3)	(93, Eye Disorders)	(28,98)	(130,80)	3
3	(40.3,M)	(160.3, 70.4)	(100, Back pain)	(22,98)	(120,80)	2
4	(40.3,F)	(160.3, 70.4)	(76, Costochondritis)	(22,98.6)	(110,70)	2
5	(40.3,F)	(160.3, 70.4)	(82, Gastritis)	(18,97.6)	(120,70)	4
6	(53.3,F)	(156.6, 66.6)	(84,Diabetes)	(19,98)	(210,100)	2
7	(53.3,F)	(156.6, 66.6)	(86, Allergic Rhinitis)	(18,97.8)	(140,80)	1
8	(53.3,M)	(156.6, 66.6)	(94, Fissure In Anal)	(22,97)	(110,80)	3
9	(33.6,F)	(160.6,56.6)	(88, Hair Fall)	(22,97)	(114,80)	4
10	(33.6,F)	(160.6,56.6)	(89, Diabetes)	(24,98)	(144,94)	3
11	(33.6,M)	(160.6,56.6)	(91, Hemorrhoids)	(20,99)	(80,50)	4

To overcome the disadvantages of bucketization such as membership disclosure and improper anatomization, HB-anonymity model is proposed. In HB-anonymity, the quasi-identifier and sensitive attributes are identified and further, the QI is broken into two sub-tables based on the correlation among the attributes. As the QI attributes are separated based on the correlation coefficient, the breaking of the linking relationship is prevented. k-anonymity is applied on the quasi-identifier and the QI is generalized by replacing it with mean values of the equivalence class. The sensitive attributes are anonymized with heap bucketization approach and slicing which in turn prevent the non-membership attack and fingerprint correlation attack.

The proposed HB-Anonymity model prevents privacy under: (1) background knowledge attack; (2) quasi-identifier attack; (3) membership disclosure attack; (4) non-membership disclosure attack; and (5) fingerprint correlation attack. If the intruder knows that pid0 is male, age > 50, the intruder can infer that the record falls in bucket 1. However, each record in bucket 1 comprises of all the three record values. Thus, the exact values of pid0 cannot be inferred. Even if the intruder knows the quasi-identifier values of an individual pid3, the intruder can correlate the values of qid and conclude the record falls in bucket 2. However, exact values for any attribute cannot be retrieved. Likewise, the existence (*mda*) and non-existence (*n-mda*) cannot be inferred precisely in Table 12. If buckets 2 and 3 are intersected, only the sensitive attribute disease = Diabetes is a common value that can be retrieved. In Bucket 2 and 3, there are total of 6 records and thus the probability of finding the individual is 0.1, which is very negligible. Thus the heap bucketization anonymity model protects the dataset from the fca also. An exhaustive evaluation of anonymization approaches on privacy-preserving data publishing has been studied and summarized in Table 13. The complete workflow of HB-anonymity and the framework of the HB-anonymity are depicted in Figure 1 and Figure 2.

TABLE 12. Heap bucketization.

PId	(Age, Sex)	(Ht,Wt)	(PR, Disease)	(RR, Temp)	(Sys,Dys)	Bucket
0	(51,M)	(173.3, 85.3)	[(92, Allergic Rhinitis), (86, Allergic Rhinitis)] (95, Anaemia)]	[(23, 97.2), (18,97.8) ,(22,98)]	[(111,67), (140,80), (130,60)]	1
1	(51,F)	(173.3, 85.3)	[(92, Allergic Rhinitis), (86, Allergic Rhinitis)] (95, Anaemia)]	[(23, 97.2), (18,97.8) ,(22,98)]	[(111,67), (140,80), (130,60)]	1
7	(53.3,F)	(156.6, 66.6)	[(92, Allergic Rhinitis), (86, Allergic Rhinitis)] (95, Anaemia)]	[(23, 97.2), (18,97.8) ,(22,98)]	[(111,67), (140,80) ,(130,60)]	1
3	(40.3,M)	(160.3, 70.4)	[(100, Back pain), (76, Costochondritis), (84,Diabetes)]	[(22,98), (22,98.6), (19,98)]	[(120,80), (110,70), (210,100)]	2
4	(40.3,F)	(160.3, 70.4)	[(100, Back pain), (76, Costochondritis), (84,Diabetes)]	[(22,98), (22,98.6), (19,98)]	[(120,80), (110,70), (210,100)]	2
6	(53.3,F)	(156.6, 66.6)	[(100, Back pain), (76, Costochondritis), (84,Diabetes)]	[(22,98), (22,98.6), (19,98)]	[(120,80), (110,70), (210,100)]	2
2	(51,M)	(173.3, 85.3)	[(93, Eye Disorders), (94, Fissure In Anal), (89, Diabetes)]	[(28,98), (22,97), (24,98)]	[(130,80), (110,80), (144,94)]	3
8	(53.3,M)	(156.6, 66.6)	[(93, Eye Disorders), (94, Fissure In Anal), (89, Diabetes)]	[(28,98), (22,97), (24,98)]	[(130,80), (110,80), (144,94)]	3
10	(33.6,F)	(160.6,56.6)	[(93, Eye Disorders), (94, Fissure In Anal), (89, Diabetes)]	[(28,98), (22,97), (24,98)]	[(130,80), (110,80), (144,94)]	3
5	(40.3,F)	(160.3, 70.4)	[(82, Gastritis), (88, Hair Fall), (91, Hemorrhoids)]	[(18,97.6), (22,97), (20,99)]	[(120,70), (114,80), (80,50)]	4
9	(33.6,F)	(160.6,56.6)	[(82, Gastritis), (88, Hair Fall), (91, Hemorrhoids)]	[(18,97.6), (22,97), (20,99)]	[(120,70), (114,80), (80,50)]	4
11	(33.6,M)	(160.6,56.6)	[(82, Gastritis), (88, Hair Fall), (91, Hemorrhoids)]	[(18,97.6), (22,97), (20,99)]	[(120,70), (114,80), (80,50)]	4

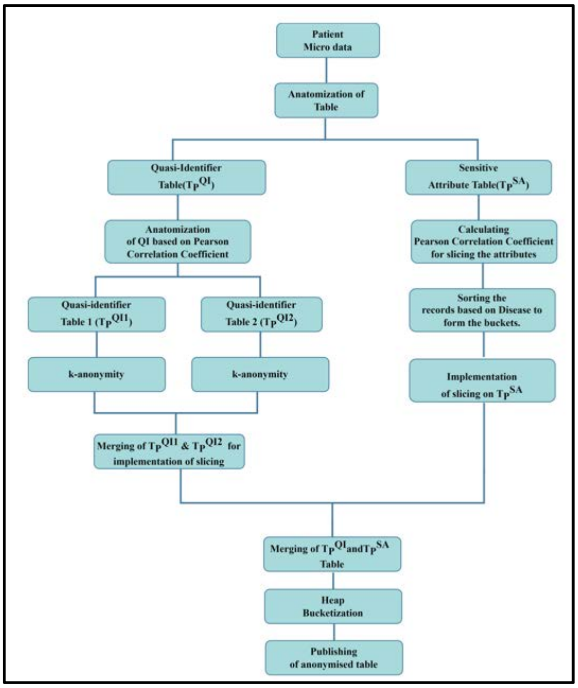

Patient Micro data
Anatomization of Table
Quasi-Identifier Table(Tp^{QI})
Sensitive Attribute Table(Tp^{SA})
Anatomization of QI based on Pearson Correlation Coefficient
Calculating Pearson Correlation Coefficient for slicing the attributes
Quasi-identifier Table 1 (Tp^{QI1})
Quasi-identifier Table 2 (Tp^{QI2})
Sorting the records based on Disease to form the buckets.
k-anonymity
k-anonymity
Implementation of slicing on Tp^{SA}
Merging of Tp^{QI1} & Tp^{QI2} for implementation of slicing
Merging of Tp^{QI}andTp^{SA} Table
Heap Bucketization
Publishing of anonymised table

FIGURE 1. Workflow of the HB-anonymity.

IX. HEAP BUCKETIZATION-ANONYMITY ALGORITHM

The primary aim of the HB-anonymity algorithm is to achieve a balance between privacy and utility. The Heap Bucketization is designed to overcome the limitations of bucketization. The generalization, slicing and bucketization are together implemented in HB-anonymity model. The complete process of the HB-anonymity model is explained in the HB-anonymity algorithm for better understanding purposes. In the HB-anonymity algorithm, the patient table, k variable is sent as an input argument in line 1. The output of the table is heap bucketized. The patient table is anatomized into two tables 1. Quasi-identifier and 2. Sensitive attribute table in lines 3 and 4. The correlation among the quasi-identifier attributes is calculated using the Pearson correlation coefficient in lines 5 and 6. In line 7, the k variable is passed to anonymize quasi-identifier and the correlation table of quasi-identifier D_1. The quasi-identifier table is further anatomized in line 8. The two quasi-identifier tables T_P^{QI1} and T_P^{QI2} are anonymized by implementing k-anonymity in lines 9 and 10. After anonymization of tables by k-anonymity, the tables are merged and slicing is applied on the highly correlated attributes in lines 11 and 12. The correlation among the sensitive attributes is calculated by the Pearson correlation coefficient in lines 13 and 14. In line 15, the sensitive attribute table is anonymized based on the correlation table D2. In the sensitive attribute table, the blood pressure is divided into two fields' sys and dys in line 16. The sensitive attribute table comprises six attributes Temp, Pulse Rate (PR), Respiratory Rate, Sys, Dys, and Disease in line 17. To form the buckets, the attributes are sorted with respect to disease in line 18. From lines 19 to 23, buckets have been formed with three records in each bucket. In line 24, the slicing has been performed on the bucketized table. In lines 25 and 26, the quasi-identifier and sensitive attribute tables are merged and the records are sorted based on the bucket id to implement Heap Bucketization. From lines 27 to 29, the values of the highly correlated attributes are grouped in each bucket to perform Heap Bucketization. Finally, the records are sorted based on the id of the records for data publishing.

X. EXPERIMENTAL DETAILS AND RESULT

A. EXPERIMENTAL SETUP

The experimental setup used for the proposed model is a windows 10 operating system with 8 GB memory, 1TB hard disk. We experimented with the work in Python 3. A novel dataset has been used in our work. The dataset is received from the Interdisciplinary Institute of Indian System of Medicine, Ayurveda. The total number of instances is 22,527. The dataset consists of information of the patients such as age, sex, height, weight, temperature, pulse rate, respiratory rate,

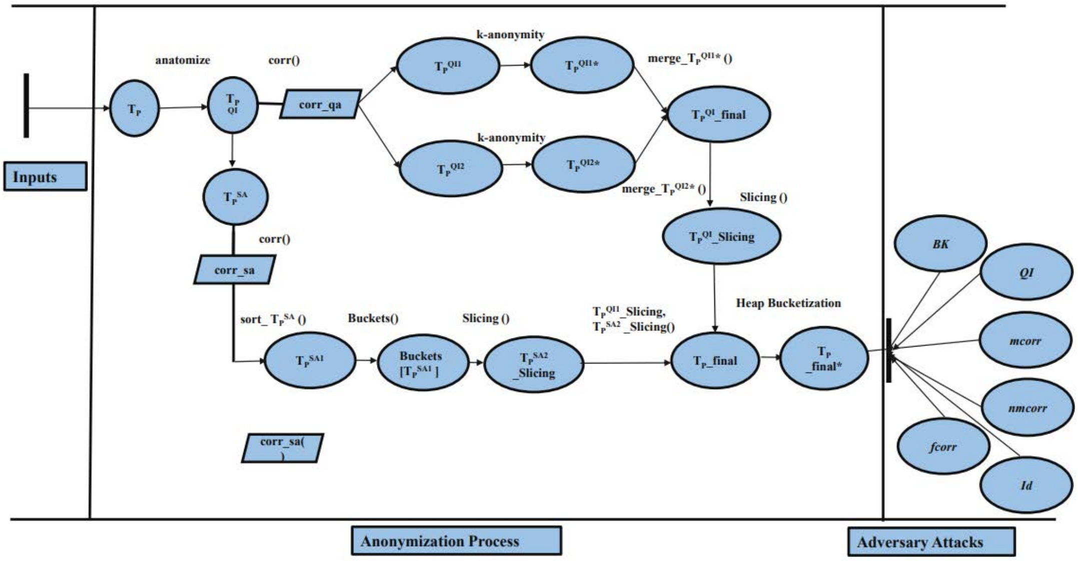


FIGURE 2. The framework of the heap bucketization- anonymity model.

blood pressure and disease. Age, sex, height and weight are categorized as quasi-identifier and temperature, pulse rate, respiratory rate, blood pressure and disease are categorized as the sensitive attributes. As age, sex, height, weight are the general information, they are categorized as quasi-identifier attributes.

B. RESULTS AND DISCUSSION

The proposed model objective is to improve the privacy of the data and to maintain the utility of the data. During preprocessing of the data, the missing field values are filled with the mean value of the column and the duplicate records are removed. After the removal of the duplications, the total number of instances is 22,043. In the proposed model, the generalization is carried out only in quasi-identifier. The sensitive attributes are not generalized or suppressed. Only the slicing of the highly correlated attributes is implemented as an anonymization process. The utility loss is measured for the quasi-identifiers using the metric Normalized Certainty Penalty (NCP) [44].

C. NCP

The utility loss for the anonymized attribute is measured using NCP as per equ.11. In the proposed model, the metric NCP is used to measure the anonymized quasi-identifier attribute.

Let ‘a’ be the attribute value of X. The NCP is defined as follows:

$$NCP\left(a\right)=f\left(x\right)=\begin{cases}0 & |a|=1\\ \dfrac{|a|}{|X|} & otherwise\end{cases} \tag{11}$$

Let |a| be the number of nodes enclosed by ‘a’ corresponding to generalized node and |X| be the total number of nodes in attribute X. The original value of the height, weight and age are taken as the old value and the generalized values are taken as the new value of the attributes.

$$\begin{aligned}Infoloss_{height} &= abs(abs\left(infoloss\left[ht_{new}\right]\right)\\ &\quad - abs\left(infoloss\left[ht_{old}\right]\right))\end{aligned} \tag{12}$$

The *infoloss* $[ht_{new}]$ represents the information loss of generalized value of attribute height and *infoloss* $[ht_{old}]$ represents the information loss of the original value of the attribute height in the patient table as per equ.12.

$$\begin{aligned}Infoloss_{weight} &= abs(abs\left(infoloss\left[wt_{new}\right]\right)\\ &\quad - abs\left(infoloss\left[wt_{old}\right]\right))\end{aligned} \tag{13}$$

The *infoloss* $[wt_{new}]$ represents the information loss of generalized value of attribute weight and *infoloss* $[wt_{old}]$ represents the information loss of the original value of the attribute weight in the patient table as per equ.13.

$$\begin{aligned}Infoloss_{age} &= abs(abs\left(infoloss\left[age_{new}\right]\right)\\ &\quad - abs\left(infoloss\left[age_{old}\right]\right))\end{aligned} \tag{14}$$

The *infoloss* $\left[age_{new}\right]$ represents the information loss of generalized value of attribute age and *infoloss* $\left[age_{old}\right]$ represents the information loss of the original value of the attribute age in the patient table as per equ.14.

The total unique records of the attributes such as height, weight and age are measured to find the mean deviation of the attributes across the unique values. The total unique records of the attribute height are 136, the unique records of the weight are 311, and the unique records of the attribute age are 89. The mean deviation of the attribute height across the 136 unique values is calculated as per equ.15 and the value of the mean deviation is 0.09.

$$\begin{aligned}&Mean\ deviation\ of\ ht\ across\ unique\ values(Md_{ht})\\ &\quad = \frac{Infoloss_{height}.mean()}{len(unique\left(QI\left[ht\right]\right))} * 100\end{aligned} \tag{15}$$

The total unique record of the attribute weight is 311. The mean deviation of the attributes weight across the 311 unique

```
Algorithm : Heap Bucketization
    Input: Patient_data $T_P$ (1:1 micro data), k, D
    Output: Heap bucketized anonymised table.
    # Splits 1:1micro data into $T_P^{QI}$ and $T_P^{SA}$
    anatomize (Patient_data($T_P$))
    split $T_P = T_P^{QI}, T_P^{SA}$
    # Finding the correlation between the quasi-identifier
    corr_qa ($T_P^{QI}$):
    pcc<-pearsoncorrcoeff($T_P^{QI}$)
    # Anonymizing the quasi-identifier table
    anony_$T_P^{QI}$ ($T_P^{QI}$, k, $D_1$):
        # Splits $T_P^{QI}$ into $\boldsymbol{T_P^{QI1}}$, $\boldsymbol{T_P^{QI2}}$
        anatomize($T_P^{QI}$)
        # Applying k-anonymity $T_P^{QI1}$, $T_P^{QI2}$
        $T_P^{QI1} * =$ k_anonymity ($T_P^{QI1}$, k);
        $T_P^{QI2} * =$ k_anonymity ($T_P^{Q2}$, k);
        # Merging $T_P^{QI1*}$, $T_P^{QI2*}$
        $T_P^{QI}$_final = merge ($T_P^{QI1*}$, $T_P^{QI2*}$);
        $T_P^{QI}$ _Slicing=slic($T_P^{QI}$_final,$D_1$)
    # Computing the correlation among the sensitive attributes
    corr_sa ($T_P^{SA}$):
    pcc<-pearsoncorrcoeff($T_P^{SA}$)
    # Sensitive attribute table anonymization
    anony_$T_P^{SA}$($T_P^{SA}$, $D_2$):
        # Dividing the blood pressure into Sys and Dys.
        Bloodpressure<-Dys, Sys
        $T_P^{SA}$ <- Temp, Pulse Rate (PR), Respiratory Rate, Sys, Dys, Disease.
        # Sort the $T_P^{SA}$ based on disease.
        $T_P^{SA1}$ <- sort_ $T_P^{SA}$ ("Disease")
        # Forming the buckets in the $T_P^{SA1}$
        $T_P^{SA1}$ ['Bucket']= 0
        b= 1
        for i in range(0,len($T_P^{SA1}$),3):
        $T_P^{SA1}$.loc[i:i+3,['Bucket']] = b
        b+=1
        # Implemention of slicing based on correlation
        $T_P^{SA2}$ _Slicing=slic($T_P^{SA1}$,$D_2$)
        # Merging the quasi-identifier and sensitive attribute table.
        $T_P$_final<- merge ($T_P^{QI1}$ _Slicing, $T_P^{SA2}$ _Slicing)
        $T_P$_final<- $T_P$_final.sort_values(['Bucket'])
        # Heap Bucketization
        $T_P$ _final*[(Sys,Dys)]<- $T_P$ _final.groupby([bucket])[(Sys,Dys)].transform(lamda x : ' '.join(x))
        $T_P$ _final*[(PR, disease)]<- $T_P$ _final.groupby([bucket])[( PR, disease)].transform(lamda x : ' '.join(x))
        $T_P$ _final*[(RR, Temp)]<- $T_P$ _final.groupby([bucket])[( RR, Temp)].transform(lamda x : ' '.join(x))
        $T_P$ _final*.sort_values('id')
        Return $T_P$ _final*
```

values is calculated as per equ.16 and the value of the mean deviation is 0.06

$$Mean\ deviation\ of\ \ wt\ across\ unique\ values(Md_{wt}) = \frac{Infoloss_{weight}.mean()}{len(unique\ (QI\ [wt]))} * 100 \tag{16}$$

The total unique record of the attribute age is 89. The mean deviation of the attributes age across the 89 unique values is calculated as per equ.17 and the value of the mean deviation is 0.1. Among the three attributes age, height and weight the weight has very less information loss compared to height

TABLE 13. An exhaustive evaluation of recent anonymization approaches on privacy-preserving data publishing.

Privacy Model/Technique	Category	Dataset used	Methodology	Problem Answered	Merits	Limitations
Multi-variable privacy characterization and quantification model [54].	1:1 relational dataset, multiple sensitive attributes (MSA).	US census data	Generalization and Suppression.	Privacy quantification problem in PPDP.	Eradicated the privacy misjudgement.	Quasi-identifier optimization is needed to minimize entropy and distribution leakage.
EQI-partitioning[55]	Set valued data.	POS, WV1, WV2.	Clustering of data.	Resist the counting, linear and batch linear query privacy breaches.	Disassociates the records in classifying combinations.	Over partitioning increases the overhead of data construction.
Enhanced rescued with neural networks.[56]	Time series data.	Taxi Trajectory Prediction and World cup	Adaptive Sampling, adaptive budget allocation, perturbation and filtering	Privacy-preservation of Spatio-temporal data in the social network.	Reduces the total group error to improve the utility.	The complexity of the overall structure.
High dimensional privacy preservation approach [57].	1:1 relational dataset	Musk, Census	Vertical partitioning, generalization, local recoding.	Anonymization of high dimensional data.	Progressive anonymization to attain a balance between both utility and privacy.	No.of fragments formed during vertical partitioning are random.
IDF-OPT method[58]	Trajectory dataset with multiple values for an individual.	D4D	Partition, Suppression.	Protection on the individual level of user dimension with multiple trajectories	Global utility loss was minimized.	Comparative analysis was not performed on the same trajectory dataset.
K-decomposition algorithm [59].	Graph dataset	Standard Network Analysis Project (SNAP2), Graph dataset.	K-anonymity mechanism.	Effective privacy-preserving methodology for large-scale graph datasets to support graph mining and analytic process.	Proposed a new matrix to measure the utility loss in the anonymized dataset.	Runtime increases as the dataset increases so careful partition are required.
An approach based on differential privacy and generalization [60].	Trajectory dataset	T-drive dataset, Chicago dataset, Taxi GPS, Beijing taxi trip dataset.	Generalization	An approach based on differential privacy and generalization. Supports possible order analysis.	Provides good data feasibility and privacy.	Implemented on the continuous static dataset.
1:M MSA-(p,l)-diversity[61]	1:M relational dataset	YouTube and informs	Mask the sensitive attribute with sensitivity based code	1: M-MSA resist attribute and identity disclosure, sfp-id, Sa-equ, SA-corr attacks.	Identity and attribute disclosures are completely eradicated using ZP3 and HLPN solver.	Due to category equalling diversity based on weight dependence, the execution time may vary according to the datasets.
Trajectory projection tree (TP-tree), IGSUP, IGSUP*, ILSUP, and ILSUP* algorithms [62].	Trajectory dataset	Oldenburg and Gowalla dataset.	Pruning strategies, suppression techniques	Inferring of unknown location with partial knowledge is eliminated.	The execution time of the algorithms are much lesser than the existing one.	Tp-tree needs high storage requirement to store the indexing structure of the original dataset.
Overlapped slicing method[63]	1:1 relational dataset, MSA	Adult	Overlapped slicing, bucketize technique and permutation.	The distribution model was proposed to resist relational attacks.	The distribution model for MSA without generalizing and suppressing the value.	Anatomy implementation could have reduced the size of the dataset for processing.
F-classify privacy model [64].	1:1 relational dataset, MSA.	Heart disease dataset, Adults	Permutation	Fuzzy logic privacy model has been proposed for privacy-preserving with MSA	First proposed rule-based fuzzy privacy model for anonymization.	There is no much difference in execution time when compared with existing.
Lsl-diversity mode [65].	1:1 relational dataset, MSA.	Irvine	Anonymized based on the security level, bucketize.	Data are anonymized according to the sensitivity level requirement of the attributes.	For MSA, the various security level for sensitive attributes and a combination of sensitive attributes were chosen.	There is a small increase in the runtime.
(p,aisg)-Sensitive - anonymity model, (p+,aisg)-sensitive - anonymity model and	1:1 relational dataset	Adult and census income	Hierarchical clustering method	Personalized privacy-preservation on individual sensitive attributes.	It computes the sensitive attribute sensitive level to focus	Need to be implemented on MSA and there was a certain utility loss.
(p+,aisg)-sensitive - anonymity model [66].					on the personalized privacy.	

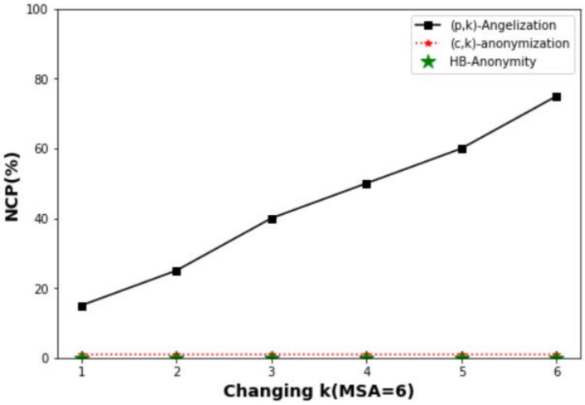


FIGURE 3. Normalized certainty penalty.

and age.

$$Mean\ deviation\ of\ age\ across\ unique\ values(Md_{age}) = \frac{Infoloss_{age}.mean()}{len(unique\ (QI\ [age]))} * 100 \qquad (17)$$

The average information loss using the NCP metric for the patient dataset is 0.083% which is very less as shown in equation18. As per our proposed model, the quasi-identifier is alone generalized and information loss due to the generalization is 0.083%. The sensitive attributes are not generalized so there is no information loss in the sensitive attributes.

$$Average\ info\ loss = \frac{(Md_{ht} + Md_{wt} + Md_{age})}{3} \qquad (18)$$

Figure 3 shows the NCP percentage value by changing the values of k with a fixed number of sensitive attributes (e.g. MSA = 6) for examining HB-anonymity, (c,k)-anonymization, (p,k)- angelization. The NCP% value of (p,k)- angelization increases unceasingly when the value of k increases. Due to this continuous increase in k-value, the utility of the dataset is getting degraded. The bucket formed in the sensitive table may affect the utility in the quasi-identifier table. The HB-anonymity has a utility loss of about 0.083% is almost equal to zero and the loss is consistent though the value of k is increased. (c,k)-anonymization has 0.9 utility loss as per our execution and in the case of NCP % value, HB-anonymity is having negligible utility loss.

D. KL-DIVERGENCE

Kullback–Leibler divergence is a metric to measure the difference in one probability distribution to another probability distribution. KL divergence is implemented considering the relation table as probability distribution d_1. $d_1(a)$ represents the element of records that belongs to A (a∈ A). The anonymized table is denoted as the probability distribution d_2 after applying the HB-anonymity. The Kullback–Leibler divergence for the patient table for the actual probability distribution (d_1) and the estimated distribution (d_2) after applying HB-anonymity is defined as below:

$$KL_d\,(d_1, d_2) = \sum\nolimits_{a \in A} d_1\,(a)\,log\left(\frac{d_1(a)}{d_2(a)}\right) \qquad (19)$$

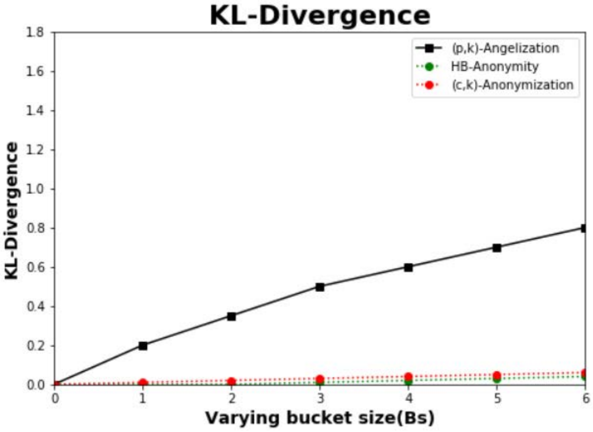


FIGURE 4. KL-divergence.

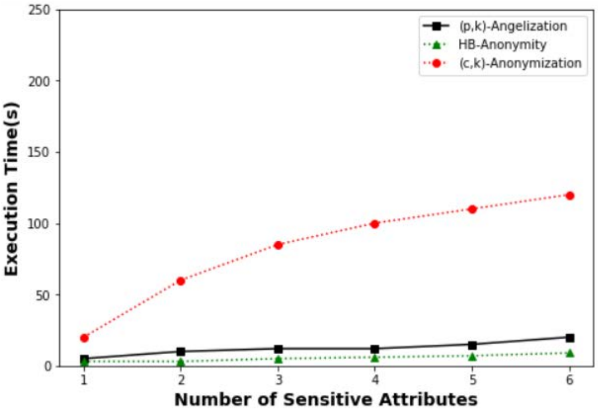


FIGURE 5. Execution time with respect to the number of sensitive attributes.

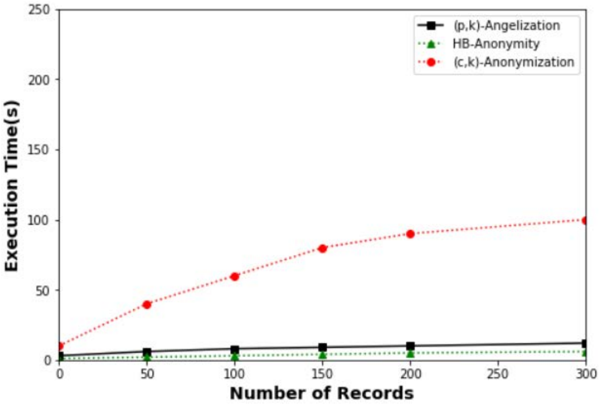
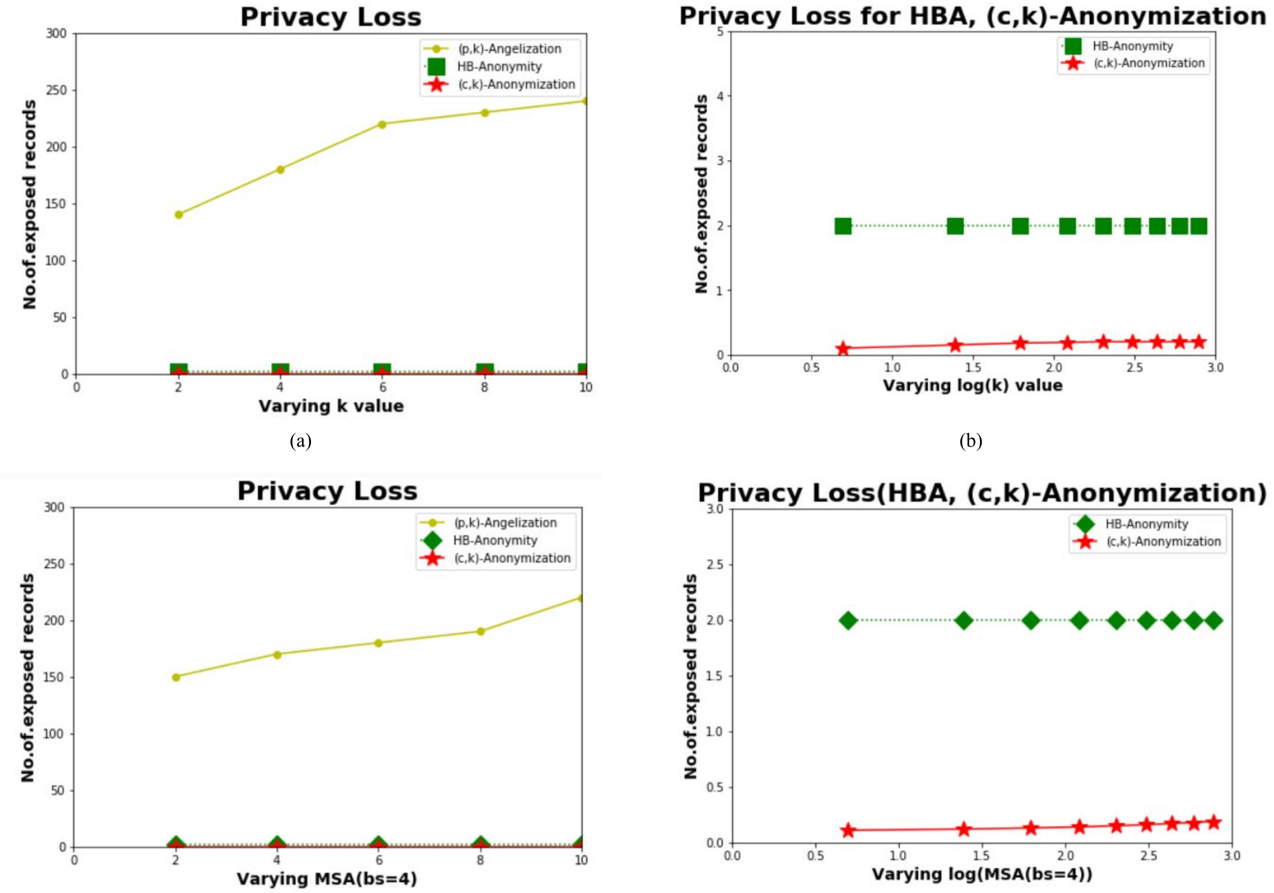


FIGURE 6. Execution time with respect to the number of records.

In HB-Anonymity, the d_1 is the actual distribution of the sensitive attributes in the patient table T_P and the estimated distribution of the sensitive attributes in the patient table after Heap Bucketization is d_2. The KL divergence is performed by changing the group size from 3-15. The (p,k)-angelization has a different score for the probability of estimated distribution for sensitive attribute buckets. In the proposed model, the KL-divergence is calculated for the sensitive attributes part. As the NCP metric has been used to calculate the utility loss in quasi-identifier, to measure the utility loss in sensitive



(a)

(b)

(c)

(d)

FIGURE 7. a. Privacy loss by varying k value. b. Privacy loss for HBA and (c,k)-anonymization by varying log(k) value. c. Privacy loss by varying MSA. d. Privacy loss for HBA and (c,k)-anonymization by varying log(MSA) value.

attributes the metric KL-divergence has been used. There are no generalization or suppression methods in sensitive attributes, so the utility loss is measured through the probability of the distribution of the actual and anonymized records.

In figure 4, the KL-divergence is plotted for the different bucket sizes. By varying the bucket size, the distribution of the data also varies. As the bucket size increases, high privacy is achieved but the utility loss is high. In the (p,k)-angelization, the probability of data distribution increases rapidly with bucket size. Whereas in (c,k)-anonymization there is zero utility loss and there is a slight increase in the utility loss if the bucket size increases. HB-anonymity also results in negligible utility loss for the reasonable bucket size and there is a slight increase in utility loss if bucket size is very high. From figure 5, the conclusion is that the (p,k)-angelization has reasonable utility loss whereas (c,k)-anonymization and HB-anonymity have negligible utility loss and it is very consistent.

E. EXECUTION TIME

When it comes to execution time, the HB-anonymity has a very negligible execution time in connection with the number of sensitive attributes. The execution time of (c,k)-anonymization is greater compared with (p,k)-angelization and HB-anonymity. The HB-anonymity protects the privacy of the data and maintains the consistent execution time. The execution time of HB-anonymity is very small and satisfactory. The HB-anonymity has very little execution time, for the number of records. The execution time of (p,k) angelization is also less and there is only a slight difference in the execution time between the (p,k) angelization and the HB-anonymity. The execution time of (c,k)-anonymization is greater when compared to (p,k) angelization and HB-anonymity.

The main advantage of HB-anonymity is though high privacy is achieved, the execution is also reduced. The proposed model has not incorporated many customized rules to achieve privacy. As the (c,k)-anonymization has imposed many rules to achieve high privacy, the execution time increases as the number of records and number of sensitive attributes increases. Figure 5 depicts the Execution time for the number of sensitive attributes and figure 6 depicts Execution time for the number of records.

F. PRIVACY LOSS

The vulnerable records that can be identified by the intruders can measure the privacy loss in a dataset. Identifying an individual in the released anonymized table is directly proportional to the privacy loss. The higher the records exposed to the intruders, is higher the privacy loss. In (p.k)-angelization and (c,k)-anonymization the privacy loss is measured by varying the values of k and multiple sensitive attributes. Likewise, in HB-anonymity the privacy loss is measured by changing the values of k and multiple sensitive attributes.

Figure 7a represents the privacy loss in (p.k)-angelization, (c,k)-anonymization and HB-anonymity by varying k value. The number of vulnerable records in (p,k)-angelization increases gradually as the k value increases because the number of records with a single sensitive value is high during the intersection of fingerprint buckets. To have a clear insight of the privacy loss for HBA and (c,k)-anonymization, the logarithmic function has been used for the k value as shown in figure 7b. When the sensitive attributes are increased, the vulnerable records also increase in (p,k)-angelization due to the increase in single sensitive value as shown in Figure 7c. The (c,k)-anonymization does not have any privacy loss as there exist no vulnerable records. Furthermore, the HB-anonymity achieves high privacy due to Heap Bucketization. To have a clear insight of the privacy loss for HBA and (c,k)-anonymization, the logarithmic function has been used for the MSA value as shown in figure 7d. The combinations of all the records belonging to one bucket are put together such that the intruder would not be able to identify any individual record. The intruder cannot be able to predict the sensitive attribute values from the intersection of any buckets. Though the (c,k)-anonymization and HB-anonymity have no privacy loss, the (c,k)-anonymization possess a much complex anonymization process, thus the execution time of the (c,k)-anonymization is high.

XI. CONCLUSION AND FUTURE DIRECTION

The paper has presented various related works on privacy-preserving data publishing with MSA. In the paper, an efficient model Heap Bucketization–Anonymity has been proposed to address the challenge of balancing the utility loss and privacy. An HB-anonymity algorithm has been developed based on the anonymization methods adopted for quasi-identifier and sensitive attributes. The HB-anonymity model concentrates on the prevention of breaking of the relationship between the attributes, thus the correlation among the quasi-identifier and sensitive attributes are calculated using Pearson correlation co-efficient to achieve less utility loss. The quasi-identifier has been anonymized by implementing k-anonymity and slicing.

A new approach Heap Bucketization has been implemented to anonymize the sensitive attributes. The proposed model Heap Bucketization makes the re-identification of the individual a challenging task for the intruder in the disclosed dataset. Experimental evaluation has been performed on the unique Ayurveda patient dataset and resulted that the proposed model achieves the balance between utility and privacy with less execution time. Moreover, HB-anonymity eradicates the various attacks such as i) background knowledge attack, ii) quasi-identifier attack iii) membership attack, iv) non-membership attack and v) fingerprint correlation attack. The future direction of the work is to develop models for dynamic data and unstructured data. In addition, we believe that quasi-identifier could be a semi-sensitive attribute and the work can be carried in such a direction. The work could be extended to 1:M microdata which is a challenging research topic.

REFERENCES

[1] Y. Alotaibi, "A new secured e-government efficiency model for sustainable services provision," *J. Inf. Secur. Cybercrimes Res.*, vol. 3, no. 1, pp. 75–96, Dec. 2020, doi: 10.26735/CAAK6285.

[2] S. Rajendran, O. I. Khalaf, Y. Alotaibi, and S. Alghamdi, "MapReduce-based big data classification model using feature subset selection and hyperparameter tuned deep belief network," *Sci. Rep.*, vol. 11, no. 1, pp. 1–10, Dec. 2021, doi: 10.1038/s41598-021-03019-y.

[3] E. Luo, M. Z. A. Bhuiyan, G. Wang, M. A. Rahman, J. Wu, and M. Atiquzzaman, "PrivacyProtector: Privacy-protected patient data collection in IoT-based healthcare systems," *IEEE Commun. Mag.*, vol. 56, no. 2, pp. 163–168, Feb. 2018, doi: 10.1109/MCOM.2018.1700364.

[4] Z. S. H. Abad, D. M. Maslove, and J. Lee, "Predicting discharge destination of critically ill patients using machine learning," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 3, pp. 827–837, Mar. 2021, doi: 10.1109/JBHI.2020.2995836.

[5] S. Banerjee, T. Hemphill, and P. Longstreet, "Wearable devices and healthcare: Data sharing and privacy," *Inf. Soc.*, vol. 34, no. 1, pp. 1–9, Dec. 2017, doi: 10.1080/01972243.2017.1391912.

[6] G. Tsang, S.-M. Zhou, and X. Xie, "Modeling large sparse data for feature selection: Hospital admission predictions of the dementia patients using primary care electronic health records," *IEEE J. Transl. Eng. Health Med.*, vol. 9, pp. 1–13, 2021, doi: 10.1109/JTEHM.2020.3040236.

[7] Z. Zhou, H. Yu, and H. Shi, "Human activity recognition based on improved Bayesian convolution network to analyze health care data using wearable IoT device," *IEEE Access*, vol. 8, pp. 86411–86418, 2020, doi: 10.1109/ACCESS.2020.2992584.

[8] G. Xu, "IoT-assisted ECG monitoring framework with secure data transmission for health care applications," *IEEE Access*, vol. 8, pp. 74586–74594, 2020, doi: 10.1109/ACCESS.2020.2988059.

[9] S. N. Prabhu, C. P. Gooneratne, K.-A. Hoang, and S. C. Mukhopadhyay, "IoT-associated impedimetric biosensing for point-of-care monitoring of kidney health," *IEEE Sensors J.*, vol. 21, no. 13, pp. 14320–14329, Jul. 2021, doi: 10.1109/JSEN.2020.3011848.

[10] A. Alsufyani, Y. Alotaibi, A. O. Almagrabi, S. A. Alghamdi, and N. Alsufyani, "Optimized intelligent data management framework for a cyber-physical system for computational applications," *Complex Intell. Syst.*, pp. 1–13, Aug. 2021, doi: 10.1007/s40747-021-00511-w.

[11] S. S. Vedaei, A. Fotovvat, M. R. Mohebbian, G. M. E. Rahman, K. A. Wahid, P. Babyn, H. R. Marateb, M. Mansourian, and R. Sami, "COVID-SAFE: An IoT-based system for automated health monitoring and surveillance in post-pandemic life," *IEEE Access*, vol. 8, pp. 188538–188551, 2020, doi: 10.1109/ACCESS.2020.3030194.

[12] H. Yu and Z. Zhou, "Optimization of IoT-based artificial intelligence assisted telemedicine health analysis system," *IEEE Access*, vol. 9, pp. 85034–85048, 2021, doi: 10.1109/ACCESS.2021.3088262.

[13] N. Y. Philip, M. Razaak, J. Chang, M. O'Kane, and B. K. Pierscionek, "A data analytics suite for exploratory predictive, and visual analysis of type 2 diabetes," *IEEE Access*, vol. 10, pp. 13460–13471, 2022, doi: 10.1109/ACCESS.2022.3146884.

[14] N. Subramani, P. Mohan, Y. Alotaibi, S. Alghamdi, and O. I. Khalaf, "An efficient Metaheuristic-based clustering with routing protocol for underwater wireless sensor networks," *Sensors*, vol. 22, no. 2, p. 415, Jan. 2022, doi: 10.3390/s22020415.

[15] M. Ali, A. Abbas, M. U. S. Khan, and S. U. Khan, "SeSPHR: A methodology for secure sharing of personal health records in the cloud," *IEEE Trans. Cloud Comput.*, vol. 9, no. 1, pp. 347–359, Jan. 2021, doi: 10.1109/TCC.2018.2854790.

[16] R. Rout, P. Parida, Y. Alotaibi, S. Alghamdi, and O. I. Khalaf, ''Skin lesion extraction using multiscale morphological local variance reconstruction based watershed transform and fast fuzzy C-means clustering,'' *Symmetry*, vol. 13, no. 11, p. 2085, Nov. 2021.

[17] Y. Alotaibi, ''A new database intrusion detection approach based on hybrid meta-heuristics,'' *Comput., Mater. Continua*, vol. 66, no. 2, pp. 1879–1895, 2021, doi: 10.32604/cmc.2020.013739.

[18] A. Zigomitros, F. Casino, A. Solanas, and C. Patsakis, ''A survey on privacy properties for data publishing of relational data,'' *IEEE Access*, vol. 8, pp. 51071–51099, 2020, doi: 10.1109/ACCESS.2020.2980235.

[19] 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.

[20] A. Stadler, ''The health insurance portability and accountability act and its impact on privacy and confidentiality in healthcare, impact of HIPAA on confidentiality,'' M.S. thesis, School Health Sci., Liberty Univ., Lynchburg, VA, USA, 2021.

[21] S. Chenthara, K. Ahmed, H. Wang, and F. Whittaker, ''Security and privacy-preserving challenges of e-health solutions in cloud computing,'' *IEEE Access*, vol. 7, pp. 74361–74382, 2019, doi: 10.1109/ACCESS.2019.2919982.

[22] L. Sweeney, ''Achieving K-anonymity privacy protection using generalization and suppression,'' *Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 571–588, 2002, doi: 10.1142/S021848850200165X.

[23] P. Samarati, ''Protecting respondents identities in microdata release,'' *IEEE Trans. Knowl. Data Eng.*, vol. 13, no. 6, pp. 1010–1027, Nov. 2001, doi: 10.1109/69.971193.

[24] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, ''l-diversity: Privacy beyond K-anonymity,'' *ACM Trans. Knowl. Discovery Data*, vol. 1, no. 1, pp. 1–12. 2007, doi: 10.1109/ICDE.2006.1.

[25] N. Li, T. Li, and S. Venkatasubramanian, ''T-closeness: Privacy beyond k-Anonymity and l-Diversity,'' in *Proc. IEEE 23rd Int. Conf. Data Eng.*, Apr. 2007, pp. 106–115, doi: 10.1109/ICDE.2007.367856.

[26] B. B. Mehta and U. P. Rao, ''Improved l-diversity: Scalable anonymization approach for privacy preserving big data publishing,'' *J. King Saud Univ. Comput. Inf. Sci.*, pp. 1–8, Aug. 2019, doi: 10.1016/j.jksuci.2019.08.006.

[27] X. Jin, M. Zhang, N. Zhang, and G. Das, ''Versatile publishing for privacy preservation,'' in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010, pp. 353–362, doi: 10.1145/1835804.1835851.

[28] J. Wang, K. Du, X. Luo, and X. Li, ''Two privacy-preserving approaches for data publishing with identity reservation,'' *Knowl. Inf. Syst.*, vol. 60, no. 2, pp. 1039–1080, Jun. 2018, doi: 10.1007/s10115-018-1237-3.

[29] L. Yao, Z. Chen, X. Wang, D. Liu, and G. Wu, ''Sensitive label privacy preservation with anatomization for data publishing,'' *IEEE Trans. Dependable Secure Comput.*, vol. 18, no. 2, pp. 904–917, Mar. 2021, doi: 10.1109/TDSC.2019.2919833.

[30] L. Zhang, J. Xuan, R. Si, and R. Wang, ''An improved algorithm of individuation K-Anonymity for multiple sensitive attributes,'' *Wireless Pers. Commun.*, vol. 95, no. 3, pp. 2003–2020, Aug. 2017, doi: 10.1007/s11277-016-3922-4.

[31] H. Zhu, S. Tian, M. Xie, and M. Yang, ''Preserving privacy for sensitive values of individuals in data publishing based on a new additive noise approach,'' in *Proc. 23rd Int. Conf. Comput. Commun. Netw. (ICCCN)*), Shanghai, China, Aug. 2014, pp. 1–6, doi: 10.1109/ICCCN.2014.6911855.

[32] Q. Liu, H. Shen, and Y. Sang, ''Privacy-preserving data publishing for multiple numerical sensitive attributes,'' *Tsinghua Sci. Technol.*, vol. 20, no. 3, pp. 246–254, Jun. 2015, doi: 10.1109/TST.2015.7128936.

[33] W. Widodo and W. C. Wibowo, ''A distributional model of sensitive values on p-sensitive in multiple sensitive attributes,'' in *Proc. 2nd Int. Conf. Informat. Comput. Sci. (ICICoS)*, Semarang, Indonesia, Oct. 2018, pp. 1–5, doi: 10.1109/ICICOS.2018.8621698.

[34] J. C. W. Lin, P. Fournier-Viger, Q. Liu, Y. Djenouri, and J. Zhang, ''Anonymization of multiple and personalized sensitive attributes,'' in *Proc. 20th Int. Conf. Big Data Analytics Knowl. Discovery*, Regensburg, Germany, Sep. 2018, pp. 204–215, doi: 10.1007/978-3-319-98539-8_16.

[35] A. Anjum, N. Ahmad, S. U. R. Malik, S. Zubair, and B. Shahzad, ''An efficient approach for publishing microdata for multiple sensitive attributes,'' *J. Supercomput.*, vol. 74, no. 10, pp. 5127–5155, Apr. 2018, doi: 10.1007/s11227-018-2390-x.

[36] J. Jayapradha, M. Prakash, and Y. H. Reddy, ''Privacy preserving data publishing for heterogeneous multiple sensitive attributes with personalized privacy and enhanced utility,'' *Systematic Rev. Pharmacy*, vol. 11, no. 9, pp. 1055–1066, Sep. 2020, doi: 10.31838/srp.2020.9.151.

[37] Y. Sei, H. Okumura, T. Takenouchi, and A. Ohsuga, ''Anonymization of sensitive quasi-identifiers for l-diversity and t-closeness,'' *IEEE Trans. Dependable Secure Comput.*, vol. 16, no. 4, pp. 580–592, Jul./Aug. 2019, doi: 10.1109/TDSC.2017.2698472.

[38] S. A. Onashoga, B. A. Bamiro, A. T. Akinwale, and J. A. Oguntuase, ''KC-Slice: A dynamic privacy-preserving data publishing technique for multisensitive attributes,'' *Inf. Secur. J., Global Perspective*, vol. 26, no. 3, pp. 121–135, May 2017, doi: 10.1080/19393555.2017.1319522.

[39] N. V. S. L. Raju, M. N. Seetaramanath, and P. S. Rao, ''A novel dynamic KCi–Slice publishing prototype for retaining privacy and utility of multiple sensitive attributes,'' *Int. J. Inf. Technol. Comput. Sci.*, vol. 11, no. 4, pp. 18–32, Apr. 2019, doi: 10.5815/ijitcs.2019.04.03.

[40] S. R. P. Reddy, K. V. Raju, and V. V. Kumari, ''A novel approach for personalized privacy preserving data publishing with multiple sensitive attributes,'' *Int. J. Eng. Technol.*, vol. 7, pp. 197–206, 2018, doi: 10.14419/ijet.v7i2.20.13296.

[41] T. Kanwal, S. A. A. Shaukat, A. Anjum, S. U. R. Malik, K.-K.-R. Choo, A. Khan, N. Ahmad, M. Ahmad, and S. U. Khan, ''Privacy-preserving model and generalization correlation attacks for 1:M data with multiple sensitive attributes,'' *Inf. Sci.*, vol. 488, pp. 238–256, Jul. 2019, doi: 10.1016/j.ins.2019.03.004.

[42] K. Albulayhi, T. P. Tosic, and T. F. Sheldon, ''G-Model: A novel approach to privacy-preserving 1:M micro data publication,'' in *Proc. IEEE Int. Conf. Cyber Secur. Cloud Comput.*, pp. 88–99, 2020, doi: 10.1109/CSCloud-EdgeCom49738.2020.00024.

[43] J. Jayapradha and M. Prakash, ''F-Slip: An efficient privacy-preserving data publishing framework for 1:M microdata with multiple sensitive attributes,'' *Soft Comput.*, pp. 1–18, Oct. 2021, doi: 10.1007/s00500-021-06275-2.

[44] P. Mohan, N. Subramani, Y. Alotaibi, S. Alghamdi, O. I. Khalaf, and S. Ulaganathan, ''Improved metaheuristics-based clustering with multihop routing protocol for underwater wireless sensor networks,'' *Sensors*, vol. 22, no. 4, p. 1618, Feb. 2022, doi: 10.3390/s22041618.

[45] X. Li and Z. Zhou, ''A generalization model for multi-record privacy preservation,'' *J. Ambient Intell. Humanized Comput.*, vol. 11, no. 7, pp. 2899–2912, Jul. 2020, doi: 10.1007/s12652-019-01430-y.

[46] B. Li, K. He, and G. Sun, ''Local generalization and bucketization technique for personalized privacy preservation,'' *Semantic Scholar*, pp. 1–12, Aug. 2020.

[47] E. Gachanga, M. Kimwele, and L. Nderu, ''Feature based data anonymization with slicing method for data publishing,'' in *Proc. 11th Int. Conf. Mach. Learn. Comput.*, 2019, pp. 274–279, doi: 10.1145/3318299.3318389.

[48] Y. Tao, H. Chen, X. Xiao, S. Zhou, and D. Zhang, ''ANGEL: Enhancing the utility of generalization for privacy preserving publication,'' *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 7, pp. 1073–1087, Jul. 2009, doi: 10.1109/TKDE.2009.65.

[49] X. Xiao and Y. Tao, ''Anatomy: Simple and effective privacy preservation,'' in *Proc. 32nd Int. Conf. Very Large Data Bases VLDB Endowment*, Seoul, South Korea, Sep. 2006, pp. 139–150.

[50] R. Khan, X. Tao, A. Anjum, H. Sajjad, S. U. R. Malik, A. Khan, and F. Amiri, ''Privacy preserving for multiple sensitive attributes against fingerprint correlation attack satisfying C-diversity,'' *Wireless Commun. Mobile Comput.*, vol. 2020, pp. 1–18, Jan. 2020, doi: 10.1155/2020/8416823.

[51] A. Majeed, ''Attribute-centric anonymization scheme for improving user privacy and utility of publishing e-health data,'' *J. King Saud Univ.*, vol. 31, pp. 426–435, Mar. 2018, doi: 10.1016/j.jksuci.2018.03.014.

[52] J. W. K. K. Edemacu and B. Jang, ''MPPDS: Multilevel privacy-preserving data sharing in a collaborative eHealth system,'' *IEEE Access*, vol. 7, pp. 109910–109923, 2019, doi: 10.1109/ACCESS.2019.2933542.

[53] V. Arul, C. Vairavel, M. Prakash, and N. V. Kousik, ''Privacy preservation of micro data publishing using fragmentation,'' *J. Soft Comput.*, vol. 9, no. 3, 1945-1949, Apr. 2019, doi: 10.21917/ijsc.2019.0271.

[54] M. H. Afifi, K. Zhou, and J. Ren, ''Privacy characterization and quantification in data publishing,'' *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 9, pp. 1756–1769, Sep. 2018, doi: 10.1109/TKDE.2018.2797092.

[55] H. Zhang, Z. Zhou, L. Ye, and X. Du, ''Towards privacy preserving publishing of set-valued data on hybrid cloud,'' *IEEE Trans. Cloud Comput.*, vol. 6, no. 2, pp. 316–329, Apr. 2018, doi: 10.1109/TCC.2015.2430316.

[56] Q. Wang, Y. Zhang, X. Lu, Z. Wang, Z. Qin, and K. Ren, ''Real-time and spatio-temporal crowd-sourced social network data publishing with differential privacy,'' *IEEE Trans. Dependable Secure Comput.*, vol. 15, no. 4, pp. 591–606, Jul. 2018, doi: 10.1109/TDSC.2016.2599873.

[57] R. Wang, Y. Zhu, C.-C. Chang, and Q. Peng, "Privacy-preserving high-dimensional data publishing for classification," *Comput. Secur.*, vol. 93, no. 1, pp. 1–10, Mar. 2020, doi: 10.1016/j.cose.2020.101785.

[58] J. Zhao, J. Mei, S. Matwin, Y. Su, and Y. Yang, "Risk-aware individual trajectory data publishing with differential privacy," *IEEE Access*, vol. 9, pp. 7421–7438, 2021, doi: 10.1109/ACCESS.2020.3048394.

[59] X. Ding, C. Wang, K. Kwang R. Choo, and H. Jin, "A novel privacy preserving framework for large scale graph data publishing," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 2, pp. 331–343, Feb. 2021, doi: 10.1109/TKDE.2019.2931903.

[60] M. Arif, J. Chen, G. Wang, O. Geman, and V. E. Balas, "Privacy preserving and data publication for vehicular trajectories with differential privacy," *Measurements*, vol. 173, pp. 1–15, Mar. 2021, doi: 10.1016/j.measurement.2020.108675.

[61] T. Kanwal, A. Anjum, S. U. R. Malik, H. Sajjad, A. Khan, U. Manzoor, and A. Asheralieva, "A robust privacy preserving approach for electronic health records using multiple dataset with multiple sensitive attributes," *Comput. Secur.*, vol. 105, pp. 1–21, Jun. 2021, doi: 10.1016/j.cose.2021.102224.

[62] C.-Y. Lin, "Suppression techniques for privacy-preserving trajectory data publishing," *Knowl.-Based Syst.*, vol. 206, pp. 1–15, Oct. 2020, doi: 10.1016/j.knosys.2020.106354.

[63] E. K. Budiardjo and W. C. Wibowo, "Privacy preserving data publishing with multiple sensitive attributes based on overlapped slicing," *Information*, vol. 10, no. 12, pp. 1–18, Nov. 2019, doi: 10.3390/info10120362.

[64] H. Attaullah, A. Anjum, T. Kanwal, S. U. R. Malik, A. Asheralieva, H. Malik, A. Zoha, K. Arshad, and M. A. Imran, "F-classify: Fuzzy rule based classification method for privacy preservation of multiple sensitive attributes," *Sensors*, vol. 21, no. 14, p. 4933, Jul. 2021, doi: 10.3390/s21144933.

[65] Y. Xiao and H. Li, "Privacy preserving data publishing for multiple sensitive attributes based on security level," *Information*, vol. 11, no. 3, pp. 1–27, Mar. 2020, doi: 10.3390/info11030166.

[66] H. Song, N. Wang, J. Sun, T. Luo, and J. Li, "Enhanced anonymous models for microdata release based on sensitive levels partition," *Comput. Commun.*, vol. 155, pp. 9–23, Apr. 2020, doi: 10.1016/j.comcom.2020.02.083.

J. JAYAPRADHA (Student Member, IEEE) received the Bachelor of Technology degree in electrical and electronics engineering from SASTRA University, Thanjavur, in 2008, and the Master of Technology degree in computer science and engineering from the SRMIST, Chennai, Tamil Nadu, India, in 2011, where she is currently pursuing the Ph.D. degree in trust computing. She also works as an Assistant Professor with the Department of Computer Science and Engineering, SRMIST. She has three years of industrial experience in the business intelligence domain. She has published many research articles on machine learning, data mining, and privacy. Her research interests include machine learning, databases, and privacy. She is a Professional Member of CSI, ISCA, and IAENG. She has also awarded the University 3rd Rank Holder for her master's degree.

M. PRAKASH received the Doctor of Philosophy and Master of Technology degrees from Anna University, Chennai, India. He is currently working as an Associate Professor with the Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India. He has 13 years of experience in teaching and learning. His research interests include big data analytics, databases, and security. He is a Professional Member of IEEE, ISTE, IE (I), ISCA, IAENG, CSTA, IACSIT, and UACEE.

YOUSEEF ALOTAIBI received the master's degree in information technology (computer network) from La Trobe University, Melbourne, VIC, Australia, in 2009, and the Ph.D. degree from the Department of Computer Science and Computer Engineering, La Trobe University, in 2014. He is currently an Associate Professor with the Department of Computer Science, College of Computer and Information Systems, Umm Al-Qura University, Saudi Arabia. He has published several international journals and conference papers. His research interests include business process modeling, business process reengineering, information system, security, business and IT alignment, software engineering, system analysis and design, sustainability, and smart cities development.

OSAMAH IBRAHIM KHALAF received the B.Sc. degree in software engineering field from Al_Rafidain University College, Iraq, the M.Sc. degree in computer engineering field from Belarussian National Technical University, and the Ph.D. degree in the field of computer networks from the Faculty of Computer Systems & Software Engineering, University Malaysia Pahang. He is currently a Senior Engineering and Telecommunications Lecturer with Al-Nahrain University. He has hold 17 years of university-level teaching experience in computer science and network technology and has a strong CV about research activities in computer science and information technology projects. He has had many published articles indexed in (ISI/Thomson Reuters) and has also participated and presented at numerous international conferences. He has a patent and has received several medals and awards due to his innovative work and research activities. He has good skills in software engineering, including experience with .Net, SQL development, database management, mobile applications design, mobile techniques, Java development, android development, and IOS mobile development, cloud system and computations, website design. He is the Editor-in-Chief and a main guest editor in many Scopus and SCI index journals. His brilliant personal strengths are in highly self-motivated team player who can work independently with minimum supervision, strong leadership skills, and outgoing personality. He has overseas work experiences in university, such as Binary University, Malaysia, and University Malaysia Pahang.

SALEH AHMED ALGHAMDI received the Bachelor of Education degree (Hons.) from the Department of Computer Science, Teachers College, Riyadh, Saudi Arabia, in 2004, the Master of Information Technology degree from La Trobe University, Melbourne, VIC, Australia, in 2010, and the Doctor of Philosophy degree in computer science from the Royal Melbourne Institute of Technology (RMIT) University, Melbourne, in 2014, thesis title A Context-Aware Navigational Autonomy Aid for the Blind. He is currently an Associate Professor with the Department of Information Technology, College of Computers and Information Technology, Taif University, Taif, Saudi Arabia. His research interests include context awareness, positioning and navigation, and visually impaired assistance.