Journal of Theoretical and Applied Information Technology

15th September 2022. Vol.100. No 17

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A MACHINE LEARNING BASED FRAMEWORK FOR BALANCING USER PRIVACY AND UTILITY WITH MULTIPLE SENSITIVE ATTRIBUTES IN HEALTH CARE DATA PUBLISHING

JAYAPRADHA. J[1], PRAKASH. M[2]

Department of Computing Technologies, SRM Institute of Science and Technology, Kattankulathur, 603203, Tamil Nadu, India,

Department of Data Science and Business Systems, SRM Institute of Science and Technology, Kattankulathur, 603203, Tamil Nadu, India,

Email: [1]jayapraj@srmist.edu.in, [2]prakashm2@srmist.edu.in

ABSTRACT

Due to the increase in the emergence of health care system information, patient records are needed for analyzing certain diseases. It leads to the need for privacy and several challenges in the current health care system. Privacy-preserving data publishing is essential to protect the patient's record from numerous attacks. The main aim of privacy-preserving data publishing is to safeguard an individuals' personal information, though it is made available for various analysis purposes. Initially, the paper has analyzed different services of an electronic health record, need for privacy and has proposed a novel anonymization technique for electronic health records. The proposed approach will overcome the following drawbacks i) generalization of all attributes which significantly reduces the information ii) identity disclosure even with adversaries having intense background knowledge and iii) the trade-off between privacy and utility. Compared to the existing system the proposed approach achieves a better result by using hierarchical id-based generalization approach. Hence, the proposed approach helps significantly in protecting individuals' information even though an intruder has intense background knowledge. Additionally, it focuses on achieving balanced utility in the anonymized data by avoiding over generalization. The proposed approach consists of four phases i) vertical partitioning, ii) Quasi-identifier bucket(QB) anonymization, iii) Sensitive attribute bucket(SAB) anonymization iv) Evaluation of classification accuracy. As per the experimental result, the proposed approach achieves improved data privacy and utility in privacy-preserving data publishing. The experimental results are not evaluated using the utility metrics. However, the results, evidently illustrate that the proposed algorithm achieves improved accuracy than the standard utility aware data anonymization algorithm in health care records. The proposed approach thwarts the identity disclosure and attribute disclosure for the static data.

Keywords: *Anonymization, Health Care, Generalization, Privacy, Utility, Classification Model.*

1. INTRODUCTION

E-Health is an evolving field in recent years and incorporates domains such as medical informatics, health services business and public health information delivered through the internet and associated technologies. Hospitals maintain records of their patients for the purpose of analysis and diagnosis. The patient data are also released to researchers for scientific researches. However, the publishing of data for researches become a major privacy issue due to presence of sensitive attributes that should not be disclosed to a stranger [1].

A Patients' record may include some sensitive attributes like physician details, test results, and diseases details that the patient does not wish to disclose. The publishing of such data to an unknown person for research purpose may lead to a privacy breach. Nowadays, sensors are laid inside the human body for tracking the health condition of a patient. The doctors typically monitor these sensors through external devices. However, these implanted sensors can also cause a privacy breach by various other attacks. It is possible to hack the sensors information connected to the internet for extracting the information of the patients which may lead to the possibility of causing loss in human lives. Such emerging technologies tend to new privacy requirements [2].

E-Health system has numerous advantages like lessening human resources, manual files, and storage rooms compared to the paper-based records. Although, E-Health system gives quick access to the stored files with reduced cost, easy tracking of the patient condition workflow, and strong support in diagnosis for decision making however, there will be a compromise in the privacy, security, and confidentiality of the patient information[3][4]. Different techniques have been proposed in the past few years for privacy-preserving data publishing and among them anonymization plays a significant role in privacy preserving. The ultimate goal of anonymizing the original data is to ensure the privacy of an individual and to prevent more information loss.

Privacy provides us the ability to choose data sharing ownership and protect personal information from being publically disclosed. Numerous anonymization methods have been proposed to overcome the trade-off between privacy and utility as it is the significant problem in privacy-preserving data publishing [5][6][7]. Although numerous anonymization techniques are proposed, the commonly used privacy methods are generalization and suppression. Nonetheless, too much of generalization or suppression on data will spoil the knowledge of the original data. The k-anonymity model plays a significant role in preventing linking of patient's health record with other external sources to get a particular individual's sensitive information.

Likewise, various classification models such as logistic regression, decision tree, random forest, gradient-boosted tree and naïve bayes are built to ensure better utility in particular. The patients do not have any concern in building the classification models with their data provided if the data is secured. However, the owner of the health record should be responsible for data privacy when it is published to a third party. Different methods such as the data distortion [8], uncertainty [9] and information loss [10] have been presented to measure the utility.

In this study, 1:1 microdata with multiple sensitive attributes is used. A model is proposed and (k, l) anonymity and id based approach is implemented for anonymizing the data and various models have been implemented on the anonymized dataset to achieve the accurate classification accuracy.

The paper is organized as follows: Section 2 presents the related work of the paper; Section 3 describes the need for privacy and security in health care data and the various services in e-health; Section 4 defines the problem definition and the explanation of the different terms related to the work and the proposed approach and techniques used in the paper are well represented in section 5; and section 6 illustrates the proposed algorithm. The experimental results are evidently explained with pictorial representation in section 7; finally, section 8 concludes the work with limitations and future directions.

2. RELATED WORK

The digital transformation of health records has increased the demand for data privacy and there have been various researches carried in health care domain. The data published for scientific researches require privacy and so various privacy models have been proposed earlier. The privacy-preserving data publishing models are categorized into five types, namely i) record linkage model, ii) attribute linkage model iii) table linkage model iv) probabilistic model and v) adversaries background knowledge model[11].

1:1 Single Sensitive Attribute

To overcome the record linkage attacks, Sweeney proposed k-anonymity [12]. A table is said to satisfy k-anonymity if an individuals' information in the published data is indistinguishable from at least k-1 individual records present in the released table. As k-anonymity suffers from homogeneity and background knowledge attacks, *l*-diversity and t-closeness models were proposed. The *l*-diversity [13] and t-closeness [14] is an extension of k-anonymity with different requirements to prevent personal information disclosure. δ-Presence, an attack model resists the table linkage problem[15] and ε -Differential Privacy model avoids personal information disclosure through the introduction of the noise in the original data. Several models were proposed assuming that adversaries have less knowledge about the published data, and do not guarantee preventing re-identification. Skyline Privacy was proposed to protect the privacy of the data, even though the adversary had background knowledge [16]. However, in the era of increased information flow, the challenges in processing data also increases.

1:1 Multiple Sensitive Attributes

Different privacy models have been proposed earlier considering the fact that an individuals' record will

have at least one sensitive attribute [17][18][19]. Later, researchers have proposed a wide variety of anonymization methods such as clustering, bucketization, slicing, to handle multiple numerical sensitive attributes [20][21][22]. The key challenge faced in anonymizing multiple numerical attributes is that it anonymizes only numerical data without considering the categorical data. Later, several anonymization methods were proposed to handle only categorical sensitive attributes [23][24]. Few researchers concentrate on heterogeneous multiple sensitive attributes, categorical, numerical and personalized privacy[25].

Customized privacy is not considered in several models, which prevent much information loss. The approach used for multiple heterogeneous sensitive attributes using correlation and personal sensitivity flags, helps in reducing the utility loss that happens due to overall generalization[26]. A technique called "ANGELMS" has been proposed for multiple sensitive attributes by performing vertical partitioning[27] and an approach (p,k)_Angelization has also been proposed to handle multiple sensitive attributes[28] (p,k)_Angelization eradicates the non-membership attack, background joins attack and achieves a seamless result in terms of utility. (p+)-sensitive, t-closeness[29] model, which is a combination of the t-closeness, p-sensitive k-anonymity models, gives a successful result to prevent similarity attack and skewness attack of the anonymized data.

1:M with Single and Multiple Sensitive Attributes

Several researches have been done assuming that an individual has only one micro data record. However, the actual scenario is that there might be multiple records for an individual. Minimal work has been done considering that an individual has multiple data records in original data 1: M[30][31]. *F*-slip model was proposed that resist various attacks such as background knowledge attack, Qausi-identifier correlation attacks, MSA correlation attack, Membership correlation and Non membership correlation attacks. A novel technique *f*-slicing was proposed to anonymize the sensitive attributes. *F*-slip model was proposed for the 1:M dataset. The privacy-preserving work is also conceded in data mining [32].

Classification Models on PPDP

Recently, lot of works has been carried out in building classification models on the anonymised data for measuring accuracy of the data. An algorithm IACk[33] has been proposed to maximize the classification utility by anonymizing the table using generalization and suppression. An algorithm called Infogain Mondrian [34] also works well for large scale data to achieve the classification utility. An enhancement of the Top-down specialization (TDS) method called Top-down refinement (TDR), improves the TDS functionalities to a greater extent. TDR helps to achieve an improved classification utility in anonymized data, and TDR can handle the attributes with or without the hierarchical generalization tree for numerical and categorical data[35]. An approach KACTUS[36], addresses the process of multidimensional generalization for achieving better classification utility and it uses a decision tree C4.5 to suppress the multifaceted regions.

3. NEED FOR SECURITY AND PRIVACY IN E-HEALTH

E-Health is an evolving domain with the union of different e-health services such as telemedicine, e-prescription, e-health applications, e-appointment, e-consultant, remote monitoring, and much other improved communication. The hospitals provide electronic health records of patients for various scientific purposes. The term Electronic Health Record (EHR) is not new; it prevails around, for the past few decades. Recently, the usage and development of electronic devices in connection with the internet produces vast data. The hospitals maintain a separate digital database to store the information of the patients and each record contains patient history and his family details. So, the health information of a patient contains ample sensitive information about him and his family too.

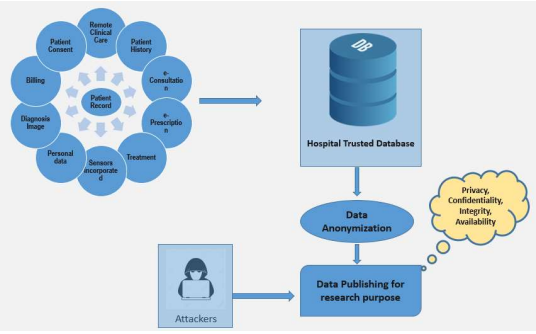


Figure 1 Different Services Of EHR And The Need For Privacy.

Legislative acts have been passed to safeguard the details of the patient so as to prevent unauthorized

usage, misuse of information, modifications, or personal information disclosure. In Figure 1, the different services of EHR and the need for privacy are depicted. Privacy cannot be maintained without appropriate security. The three following terms play a significant role in security i) Confidentiality ii) Integrity iii) and Availability. Maintaining confidentiality is to ensure that disclosure of patient information to unauthorized should be avoided. Trust needs to be developed in the originality of the Occurrence of privacy breach affects not only a particular individual, it also affects the reputation of a trusted institution. Attacks on implanted sensors not only lead to privacy breach, it may also cause loss of human lives. Therefore, privacy and security play a significant role in E-health domain in all aspects. The release of EHR for scientific research purposes, needs to be anonymized before publishing data. The anonymization techniques need to be carefully chosen to uphold privacy as well to avoid information loss.

4. PROBLEM STATEMENT

4.1 Problem Definition

Vertical partitioning method is applied, to fragment the given dataset into quasi-identifier (QID) (def.1) attributes and sensitive attributes (SA). k-anonymity is applied on the QID fragment and (k, l) diversity, a hybrid of k-anonymity and *l*-diversity is applied on the SA fragment to produce an anonymized data set for data publishing and classification models are implemented on the anonymised data to evaluate the accuracy.

4.2 Why Still K-Anonymity?

There is always a thought why k-anonymity is still used as a stronger foundation for all anonymization algorithms? Though several enhanced privacy preservation models such as l-diversity, t-closeness[37], (α,k)anonymity[38], were proposed. k-anonymity (def.3) is a significant concept that prevents the re-identification risk in anonymized data by linking it with external sources. K-anonymity strongly maintains the authenticity of data and anonymizes the quasi-identifier attributes to thwart attaining the sensitive attributes. K-anonymity is a powerful method when it is applied in the right place and it is an NP-Hard problem.

Definition 1 (Quasi-identifier) - A set of attributes that is considered non-sensitive in a dataset and can hypothetically identify an individual. {Gender, Age, Zip code} in Table 1 are quasi-identifiers and it can hypothetically disclose the personal information of an individual. The attributes available to the adversaries constitute quasi-identifier attributes. Given a data of entities E, with entity-specific table RT (A_1.,….,An), p_c: E $\rightarrow$RT and $p_g \rightarrow$E'; where E $\subseteq$E'.A quasi-identifier of RT, written $Q_{RT,}$ is a set of attributes $\{A_i\ldots,..,A_j\} \subseteq \{A_1,\ldots A_n\}$.

Definition 2 (Equivalence Class (EC)) – The table is partitioned into different subclasses S, which consists of "all elements that are equivalent to each other." It is a subset of the form {z∈Z: z RT x} where x is an element of Z, and the term z RT x indicates that there is an equivalence relation between z and x.

Definition 3 (k-anonymity) – A table T satisfies the k-anonymity property if the record t of an individual (t∈T) in the published table cannot be distinguished from "at least k-1 individuals" whose information also appear in the published table [ti1, ti2…..tik-1 ∈T] and the probability of re-identification risk of an individual is not greater than 1/k where k is a threshold being set. Let T(A_1.,…A_n) be a relational table and QI_T be the quasi-identifier related with it. The table T is said to be k-anonymized if each sequence of the attribute value in T QI_T] seems with "at least k occurrences" in T[QI_T].

Definition 4 (*l*-diversity) - A table T satisfies *l*-diversity property if table T contains "*l* well represented" values for sensitive attributes. Likewise, every equivalence class (def.2) in table T should satisfy the *l*-diversity property.

Definition 5 (k, l) anonymity – It is an improved version of k-anonymity. For a given data, generalization approach is used and two thresholds k and l are set. The parameter k designates the "anonymization level of an identifying attribute in a class," and the parameter *l* refers to the "diversity level of a sensitive attribute in a class". Every individual has the privileges to define the threshold k and *l*. (i.e) Each n_{th} record of the original data, there can be "at least k_n record for the same anonymized

identifying attributes" values corresponds to the sensitive attribute equivalence class *ln* where the size is ($l_{n\leq} k_n$).

Definition 6 (Quasi identifier Bucket (QB) - For a given dataset, the identified quasi-identifier attributes form a bucket and the Quasi-identifier bucket should satisfy the k-anonymity property. In Table 1, {Gender, Age, Zip code} ⊆ QB.

Definition 7 (Sensitive Attribute Bucket (SAB)) - For a given dataset, the identified sensitive attributes form a bucket and the Sensitive attribute bucket should satisfy the (k, *l*) anonymity property. In Table 1, {Occupation, Education, Work class, Disease} ⊆ SAB

Definition 8 (Generalization) – It replaces the quasi-identifier values with the range of values. If an attribute age has the value 29, it may be replaced with a value <=29 or may be represented as a range [21-30].

Definition 9(K^{th} record) – The value of k_n on the n^{th} record in a table refers the size of the anonymized quasi-identifier equivalence class mapping to the n^{th} SI record, where SI denotes the sensitive attribute set in a table.

Definition 10(l^{te} record) – The l_n value on the n^{th} record in a table refers to the count of different values in the SI equivalence class mapping to the n^{th} quasi-identifier tuple.

5. PROPOSED APPROACH

The workflow of the proposed approach is depicted in Figure 2 and it involves the following modules i) Pre-processing and vertical partitioning of the original data ii) Classification of attributes iii) QB anonymization iv) SAB anonymization v) Merging of QB and SAB vi) Evaluation of classification utility.

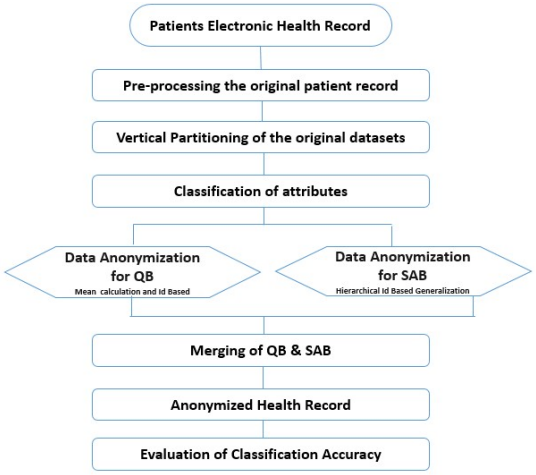


Figure 2: The work flow of the proposed approach.

5.1 Pre-processing and Fragmentation of the `Original Data

The main aim of pre-processing the raw data is to get accurate results. Before progressing to the other models initially, the dataset needs to be checked for out of range values, missing values, unrelated data because it may lead to inaccurate and misleading results if unattended. The analysis process may also increase the complexity, if the dataset is not pre-processed. Generally, the steps included in data pre-processing are data cleaning, data smoothing, data selection, extraction, transformation, etc. The dataset used for the work is already pre-processed so, we have cleaned our dataset just by filling the missing value with the appropriate value.

In general, the vertical partitioning technique segments the original dataset into multiple tables, here it is only two tables with different number of columns containing the same rows. The original data in Table 1 is a patient record. Here, the technique's primary focus is to partition the original data into two disjointed subsets so as to anonymize them effectively, as shown in Table 2a &2b. Quasi-identifier attributes constitute the first fragment, and the sensitive attributes constitute the second fragment. The primary purpose of vertical partitioning is to anonymize the two identifier buckets separately by implementing different anonymization method such as mean and id-based approach for the QB and a "hierarchical id-based generalization" for SAB.

Table 1: Original data

	PID	Age	Sex	Zip code	Occupation	Education	Work class	Disease
rt1	0(Mave)	39	M	225855	Admin-Clerical	Bachelors	State-gov	Flu
rt2	1(John)	50	M	226482	Exec-Managerial	Bachelors	Self-emp	Diabetes
rt3	2(Joey)	38	M	215646	Handlers-Cleaners	HS-grad	Private	Cancer

rt4	3(Bob)	53	M	234721	Handlers-Cleaners	11th	Private	Bronchitis
rt5	4(Rosy)	28	F	338409	Prof-speciality	Bachelors	Private	Ebola
rt6	5(Ham)	37	F	284582	Exec-Managerial	Masters	Private	HIV
rt7	6(Mary)	49	F	160187	Other-services	9th	Private	Flu
rt8	7(Ben)	52	M	209642	Exec-Managerial	HS-grad	Self-emp	Diabetes
rt9	8(Anna)	31	F	457810	Prof-speciality	Masters	Private	Hypertension

Table 2a: Quasi identifier Bucket (QB)

PID	Age	Sex	Zipcode
0(Mave)	39	M	225855
1(John)	50	M	226482
2(Joey)	38	M	215646
3(Bob)	53	M	234721
4(Rosy)	28	F	338409
5(Ham)	37	F	284582
6(Mary)	49	F	160187
7(Ben)	52	M	209642
8(Anna)	31	F	457810

Table2b: Sensitive Attribute Bucket (SAB)

PID	Occupation	Education	Work class	Disease
0(Mave)	Admin-Clerical	Bachelors	State-gov	Flu
1(John)	Exec-Managerial	Bachelors	Self-emp	Diabetes
2(Joey)	Handlers-Cleaners	HS-grad	Private	Cancer
3(Bob)	Handlers-Cleaners	11th	Private	Bronchitis
4(Rosy)	Prof-Speciality	Bachelors	Private	Ebola
5(Ham)	Exec-Managerial	Masters	Private	HIV
6(Mary)	Other-services	9th	Private	Flu
7(Ben)	Exec-Managerial	HS-grad	Self-emp	Diabetes
8(Anna)	Prof-speciality	Masters	Private	Hypertension

When a trusted institution, says a hospital collects the information of the patients, the hospital should take the full responsibility for the data. For a specific purpose of analysis, the trusted institution releases the reports of patients to a third party that conducts research experiments on the released data. The primary outcomes may correlate with age, gender and sometimes the specific diseases are even related to the occupation. Mostly, the e-health record contains age, gender, zip code, disease, contact address, contact number, occupation, and treatment history details. For certain conditions, the details of the patients' family members are also collected and stored. In the proposed approach, a novel technique is adopted to anonymize the original data.

5.2 Classification of Attributes

Once the vertical partition is done, the attributes are categorized categorically or numerically, as shown in Table 3.The classification of attributes is done to adopt different anonymization methods for both categorical and numerical data. During publishing of data for research purposes, Name (ID), the identifying variable will be obliterated. Though quasi-identifier attributes and sensitive attributes are anonymized independently, they are linked through the unique id created in original data for data publishing.

Keeping in mind the trade-off between privacy and utility, mean and id-based anonymization for QB and a novel variant technique hierarchical id-based generalization is adopted for SAB anonymization. As, the E-health record constitutes both categorical and numerical data, the proposed approach follows two different anonymization techniques. In Table 3, we have classified the attributes of the patient record and the primary purpose of attributes classification is to anonymize them according to the proposed approach. In most of the existing model, all the quasi-identifier attributes and sensitive attributes are over-generalized to preserve privacy, which reduces the utility to a greater extent. Here, a hierarchical id-based generalization approach is adopted for categorical sensitive attributes and the ID-Based Technique for quasi-identifier categorical attributes (i.e.) converting the categorical to numerical[39]. The original values of numerical attributes are replaced with the equivalence class mean value to anonymize in both QB and SAB. The quasi identifier and the sensitive attribute identifier anonymization processes are explained below.

Table 3: Attributes Classification

Attributes	Numerical Attribute	Categorical Attribute
Name	No	Yes
Age	Yes	No
Sex	No	Yes
Zip code	Yes	No
Occupation	No	Yes
Education	No	Yes
Work Class	No	Yes
Disease	No	Yes

5.3 Quasi-Identifier Bucket Anonymization

Generalization (def.8) upholds an individual privacy, but it loses its data utility when it is used in a greater manner, so the data cannot be used for research analysis purposes. Generalization does not work well for a few implementation techniques such as correlation analysis and data mining. In Table 4, k=3, so equivalence class with minimum of three records have been formed. For each record in table RT, written quasi-identifier as $RT[QI_{RT}]$ satisfies 3-anonymity, where each value sequence in $RT[QI_{RT}]$, there should be at least 3 occurrences of those record values in $RT[QI_{RT}]$. Therefore, in Table 1, $rt1[QI_{rt}]= rt2[QI_{rt}]= rt3[QI_{rt}]$, $rt4[QI_{rt}]= rt5[QI_{rt}]= rt6[QI_{rt}]$ and $rt7[QI_{rt}]= rt8[QI_{rt}]= rt9[QI_{rt}]$ forms the equivalence classes. If an adversary has intense background knowledge about the individual, then generalization cannot ensure data privacy. If an adversary has the background knowledge that Joey is a 38 old Male from zip code 215646, then the adversary can infer that joey is either suffering from Flu, Diabetes, or Cancer. Furthermore, if the adversary still has a strong knowledge that Joey is an unhealthy person, he can quickly identify joey is suffering from cancer or diabetes.

Table 4: k=3, Generalization

PID	Age	Sex	Zipcode	Occupation	Education	Work class	Disease
0(Mave)	35-50	M	[210000-230000]	Admin-Clerical	Bachelors	State-gov	Flu
1(John)	35-50	M	[210000-230000]	Exec-Managerial	Bachelors	Self-emp	Diabetes
2(Joey)	35-50	M	[210000-230000]	Handlers-Cleaners	HS-grad	Private	Cancer
3(Bob)	25-55	M	[230000-340000]	Handlers-Cleaners	11th	Private	Bronchitis
4(Rosy)	25-55	F	[230000-340000]	Prof-speciality	Bachelors	Private	Ebola
5(Ham)	25-55	F	[230000-340000]	Exec-Managerial	Masters	Private	HIV
6(Mary)	30-60	F	[150000-460000]	Other-services	9th	Private	Flu
7(Ben)	30-60	M	[150000-460000]	Exec-Managerial	HS-grad	Self-emp	Diabetes
8(Anna)	30-60	F	[150000-460000]	Prof-speciality	Masters	Private	Hypertension

In this approach, the first step is to categorize the data according to the attribute type, as shown in Table 3. First, the numerical data is anonymized followed by the categorical data. For numerical attributes, the original values are replaced with the average value of the equivalence class. (i.e) Mean value of the equivalence class. K-anonymization technique is implemented for privacy-preserving in QB. K-anonymity helps to prevent the record linkage attack through quasi-identifier attributes. Though many extended and advanced techniques are proposed, the k-anonymity plays a major role in preventing the interconnection of quasi-identifier attributes to identify a person. The probability of the risk of identifying the individual is max {1/k}. After applying k-anonymity in QB, the numerical attributes are anonymized as follows. In all the equivalence classes, the age values are replaced by the mean of the original data. Equivalence class constitutes a group of k-anonymized tuples that have the same value for all the quasi-identifier attributes. So, the first three records of quasi-identifier buckets form an equivalence class EC1, the next three tuples form EC2 and the last three records form EC3. The PID {Mave, John,Joey} represents the first equivalence class that has same generalized values for the quasi-identifier attributes age, sex and zipcode in Table 4. In the first equivalence class, there are three values 39, 50, and 38. The mean of the first equivalence class is calculated, as shown in (1).

$$\text{Mean} = \frac{A_1 + A_2 + \ldots + A_n}{n} \quad (1)$$

Where A_1 and A_2 are the different values of the numerical attributes in each equivalence class.

$$\text{Mean (Age (EC1))} = \frac{39 + 50 + 38}{3} = 42 \quad (2)$$

Mean (Age (EC1)) is the mean value of age for the first equivalence class; likewise, the same calculation is done for the remaining equivalence classes EC2 and EC3. The same procedure is followed to anonymize the zip code in Table 2a as followed for the attribute age.

$$\text{Mean (Zip code (EC1))} = \frac{225855 + 226482 + 215646}{3} = 222661 \quad (3)$$

Where Mean (Zip code (EC1)) is the mean value of the zip code for the first equivalence class; likewise, the same calculation is done for the remaining

equivalence classes EC2 and EC3. To ggggvanonymize the categorical data in QB, the "ID-Based" approach is followed and each value in an attribute is assigned an ID. For example, the attribute Sex is a categorical attribute, where Male=0 and Female =1, as shown in Figure 3. Table 5 is a 3-anonymous table with our id-based approach for categorical and mean values for the numerical attribute.

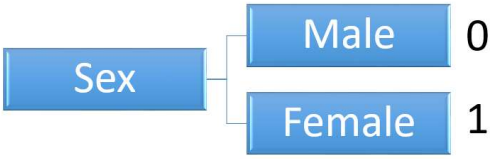


Figure 3: An ID-Based approach for attribute Age.

5.4 Sensitive Attribute Bucket Anonymization

Table 5: Anonymized Quasi identifier Bucket (k=3)

PID	Age	Sex	Zip code
0(Mave)	42	0	222661
1(John)	42	0	222661
2(Joey)	42	0	222661
3(Bob)	39	0	285904
4(Rosy)	39	1	285904
5(Ham)	39	1	285904
6(Mary)	44	1	275880
7(Ben)	44	0	275880
8(Anna)	44	1	275880

Since sensitive attributes play an essential role in a patients' record, the proposed approach focuses more on SAB anonymization. In QB, the mean value calculation and id-based approach is adopted for anonymizing. In SAB, "hierarchical id-based generalization approach" is implemented for anonymizing the SA.

In quasi-identifier bucket, the attribute sex is assigned with ID 0 and 1 for male and female. To prioritize the sensitive attributes, a "hierarchical id-based generalization approach" is implemented for the SAB. Hierarchical id-based generalization approach is having few similarities with the smoothing of data. Attribute disease is a 2-level hierarchy taxonomy and it is assigned with hierarchy id. In the patient record, three diseases are considered i) respiratory -0, ii) epidemic -1, and iii) dangerous virus -2. The respiratory department diseases are categorized as flu-0, bronchitis-1; the epidemic department diseases are diabetes-0, cancer-1, hypertension-2, and the hazardous virus disease are HIV-0 and Ebola-1. The hierarchy tree of disease with the id assigned, is shown in Figure 4. The parent category are the respiratory problem, epidemic disease, a dangerous virus with assigned ids 0, 1, 2. The diseases that come under respiratory problems are flu and bronchitis with assigned ids 0, 1. So id of the disease flu for Mave is 00(which is a combination of flu-id and respiratory problem id). A top-down approach is implemented for assigning id and (k, *l*)-anonymity model is used for sensitive attribute bucket after hierarchy id-based generalization approach. (k, *l*)-anonymity model is an improved version of k-anonymity.

Likewise, the hierarchy tree is constructed for other sensitive attributes also. The attribute work class is a 3-level hierarchy, as shown in Figure 6; education and occupation is a 2-level hierarchy, as shown in Figure 7&5. To prevent sensitive attribute bucket from re-identification attack, we have implemented (k,*l*)-anonymity (def.5)[40], a hybrid of k-anonymity and *l*-diversity(def.4) where the two-parameters k and *l* are set. The k indicates the identifying anonymization level, and *l* indicates the diversity level of a sensitive attributes equivalence class. Both parameters k and *l* can be changed according to the user. The parameter k designates the "anonymization level of an identifying attribute in a class," and the parameter *l* refers to the "diversity level of a sensitive attribute in an equivalence class".

The model is much flexible and can work as either k-anonymity or *l*-diversity. It can only use as a k-anonymity model when the parameter l_n is set to 1 and all k_n to a constant k. It can also work as a standalone *l*-diversity model if all l_n and k_n to a constant value *l*. K-anonymity alone cannot perform well in preventing the linking attack for sensitive attributes, so *l*-diversity together formed a (k,*l*)-anonymity. Table 6 is the outcome of the (k,l)-anonymity model with k=3, *l*=3, along with hierarchical id-based generalization approach. When adversary tries to identify an individual named John from Table 7, with the background knowledge age=50, sex=male and zip code =226482, he cannot be able to identify John and his sensitive attributes. Thus identity disclosure is not possible even with intense background knowledge. The proposed approach enhance the privacy, reduces information loss and prevents the identity disclosure and the attribute disclosure.

Table 6: (k,l)anonymity k=3, l=3

Journal of Theoretical and Applied Information Technology
15th September 2022. Vol.100. No 17

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

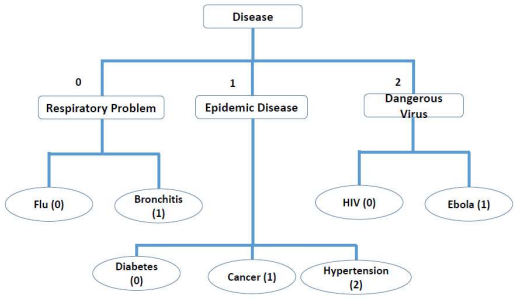


Figure 4: 2-Level Hierarchical Taxonomy For The Attribute Disease.

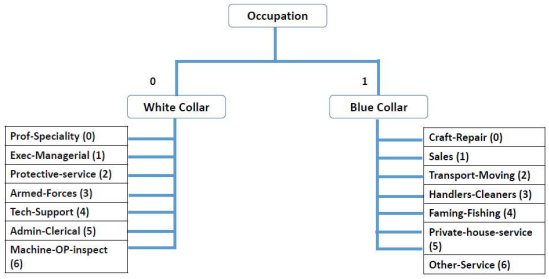


Figure 5: 2-Level Hierarchical Taxonomy For Attribute Occupation

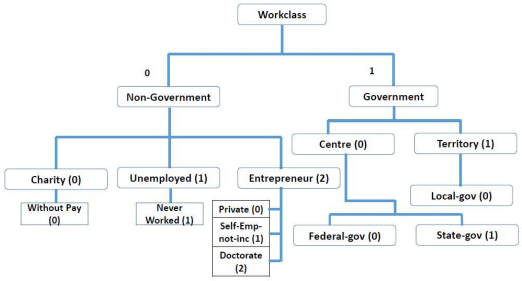


Figure 6: 3-Level Hierarchy Taxonomy For Attribute Work Class.

PID	Occupation	Education	Work class	Disease
0(Mave)	05	10	101	00
1(John)	01	10	021	10
2(Joey)	13	07	020	11
3(Bob)	13	05	020	01
4(Rosy)	00	10	020	21
5(Ham)	01	21	020	20
6(Mary)	16	03	020	00
7(Ben)	01	07	021	10
8(Anna)	00	21	020	12

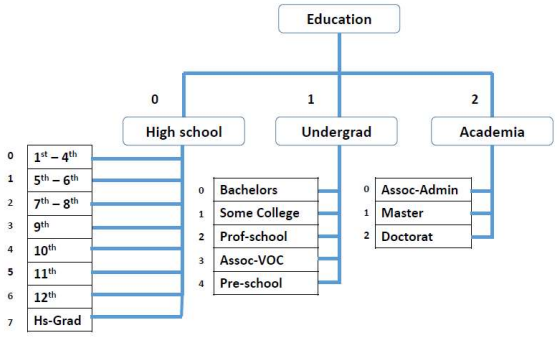


Figure. 7: 2-Level Hierarchical Taxonomy For Attribute Education

5.5 Linking of Anonymized Buckets and Data Publishing

The anonymization process in the proposed approach is carried out after the vertical partitioning of the original table. A unique id is created for original patient table, which will be used as an index for both the QB and SAB to link them after anonymization process. The resulted anonymized data is shown in Table 7. However, this unique id (PID) will not be disclosed to a third party; it is just used for linking the buckets after anonymization.

Table 7: Anonymized Data

PID	Age	Sex	Zip code	Occupation	Education	Work class	Disease
0(Mave)	42	0	222661	05	10	101	00
1(John)	42	0	222661	01	10	021	10
2(Joey)	42	0	222661	13	07	020	11
3(Bob)	39	0	285904	13	05	020	01
4(Rosy)	39	1	285904	00	10	020	21
5(Ham)	39	1	285904	01	21	020	20
6(Mary)	44	1	275880	16	03	020	00
7(Ben)	44	0	275880	01	07	021	10
8(Anna)	44	1	275880	00	21	020	12

Since two different methods are applied for a quasi-identifier bucket and sensitive attributes bucket, there might be shuffling of records, leading to the data mismatch. Creating the unique id helps linking the correct records of both the buckets before releasing the anonymized record. Finally, the anonymized record will be published to the third party with greater privacy and utility. The evaluation of the accuracy using classification models are explained in proof of experiment and results section.

6. PROPOSED ALGORITHMS

The proposed algorithm deals with the initial three phases of the work i) vertical partitioning, ii) QB anonymization iii) SAB anonymization. Algorithm 1 outlines the vertical partitioning, lines 1-3, assigns a unique id for the original data, and then fragments the table into two buckets quasi-identifier and sensitive attribute. In algorithm 2, line 1 passes the k parameter as an input and the user can fix the k parameter. Line 2-5, the attributes are classified, and the list of categorical and numerical attributes are identified. The k-anonymity is applied for categorical attributes from lines 6-10 and ids are assigned to the categorical attributes from lines 11-17. For numerical attributes, the mean values are calculated from lines 18-22. The values in each EC are replaced with the calculated mean values from lines 23-30.

In algorithm 3, three parameters SAB, k, *l* are passed as arguments to SAB_anonymity for anonymizing the sensitive attributes. As already mentioned in definition 5, the parameter k designates the "anonymization level of an identifying attribute in a class," and the parameter *l* refers to the "diversity level of a sensitive attribute in a class." The type of the attributes are classified, and a list of categorical and numerical attributes are identified from lines 2-5, and k-anonymity is performed. The outcome of performing k-anonymity is passed to *l*-diversity to form the unique elements in the equivalence class. L-diversity is applied for the categorical attributes from lines 8-20, and for the numerical attribute, the *l*-diversity is performed from 21-35. Line 36 represents the hierarchy encoding for the categorical sensitive attributes in the SAB. Then the mean calculation and replacement of mean values are implemented for numerical attributes. Finally, in line 38, the merge function combines the QB and SAB with the help of the unique id that finally results in the anonymized table T*.

Algorithm 1 Vertical partitioning

```
Input: Table T
Output: QB, SAB of Table T.
// Generation of unique ID for all the records to ensure that it will be used for further grouping.
[1] T =Function id (RT)
// Vertical Partition the Table T into Quasi-identifier Bucket and Sensitive Attribute Bucket
[2] QB, SAB = vertical_function Split (T)
[3] return T;
```

Algorithm 2 QB Anonymization

```
Input: Partition QB
Output: Anonymized Table QB*
// Applying k-anonymity on Quasi-identifier bucket.
[1] QB_anonymity = Function k_division(QB,k)
// Classifying the categorical and numerical
// Identify the list of the categorical attribute.
[2]QBcat = list( function cat(attribute))
[3]end
// Identify the list of numerical attribute.
[4]QBnum =list(function num(attribute))
[5]end
// k-anonymity for categorical attribute in Quasi-identifier bucket.
[6]for each element in QBcat
[7]for each attribute in element
[8]   k_cat = list(unique(element))
[9]     end for
[10]end for
// Assigning id for categorical attribute values.
[11]for each element in QBcat
[12]   for each attribute in QBcat
[13]     for each value in attribute
[14]value = k_cat[index(attribute)]
[15] end for
[16]   end for
[17]     end for
// k-anonymity for numerical attribute in Quasi-identifier bucket.
[18]for each element in QBnum
[19]   for each attribute in QBnum
[20]k_num = list(mean(attribute))
[21]end for
[22]    end for
[23]for each element in QBnum
[24]   for each attribute in QBnum
[25]     for each value in attribute
[26]value = k_num[index(attribute)]
[27]end for
[28]    end for
 [30]      end for
```

Algorithm 3 SAB Anonymization

```
Input: Partition SAB
Output: Anonymized Table SAB*, Anonymized Table T*.
// Applying (k, l)-anonymity on the Sensitive attribute bucket.
[1]  SAB_anonymity = Function (k,l)division(SAB,k,l)
[2]  SAB_cat=function_cat(SAB)
[3]  SAB_num=function_num(SAB)
[4]  SAB_all_cat = list()
[5]  SAB_all_num= list()
[6]  for each attribute in SABcat
[7]   SAB_cat_sub = list()
// Line 6-10 in algorithm 2 will perform the k-anonymity for SAB.
[8]   If length(unique(SAB_cat [attribute])) > l
[9]    Function append(SAB_all_cat,SAB_cat)
[10]  else
[11]   for each item in SABcat
[12]    if item[attribute] not in SAB_cat
[13]      Function append(SAB_cat_,item[attribute])
[14]      break
[15]    else
[16]      continue
[17]    end if
[18]   end for
[19]  end if
[20] end for
[21] for each attribute in SABnum
[22]   SAB_num = list()
[23]  if length(unique(SAB_num[attribute])) > l
[24]    Function append(SAB_all_cat,SAB_num)
[25]  else
[26]    for each item in SABnum
[27]      if item[attribute] not in SAB_num
[28]        Function append(SAB_num, item[attribute])
[29]        break
[30]      else
[31]        continue
[32]      end if
[33]    end for
[34]  end if
[35] end for
// Performing Hierarchical encoding for categorical variable.
[36]  Hierarchy_Encoding=Function hierarchy( SAB_cat)
// Line 18-30 in Algorithm 2 will perform the assigning mean values for the numerical attributes in each EC.
[38] Anonymized table(T*)=function merge( SAB_num, SAB_cat, QBcat, QBnum)
    return T*
```

7. PROOF OF EXPERIMENT AND RESULTS

The main objective of the proposed approach is privacy-preserving and evaluating the classification accuracy of the anonymized dataset. The anonymized dataset utility is compared with the standard "utility-aware" algorithms such as Info Gain Mondrain and IACK algorithms. Researchers mostly have used adult data set from UCI machine repository for implementing k-anonymity. In the proposed work, the adult dataset from UCI Machine repository and the cardiovascular study dataset from the Kaggle repository are used. After pre-processing the data by filling out the missing and unknown values, the resulting dataset in adult is 45,223 tuples and the tuples in cardiovascular dataset is 70,000. The adult dataset have 14 attributes. The five attributes: age, sex, marital_status, relationship, the native country are used as the quasi-identifier attributes, and four attributes occupation, education, education_num, and work class as the sensitive attributes described in Table 8. The cardio vascular dataset have 13 attributes, where age, gender, height, weight are used as the quasi-identifiers and attributes cholesterol, gluc, smoke as sensitive attributes. As the cardiovascular data is already pre-processed, we have taken all 70,000 tuples for the experimental work.

Table 8: Classification Of Attributes In Adult Dataset

Attribute	Quasi-identifier	Sensitive	Classification
Age	✓		Numerical
Occupation		✓	Categorical
Sex	✓		Categorical
Marital Status	✓		Categorical
Education		✓	Categorical
Relationship	✓		Categorical
Native country	✓		Categorical
Education num		✓	Numerical
Work class		✓	Categorical

Table 9: Classification Of Attributes In The Cardiovascular Dataset

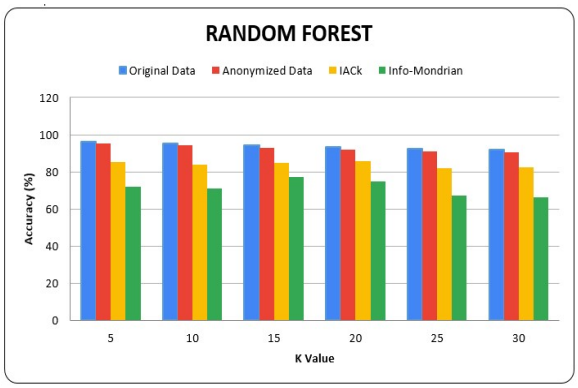
Attribute	Quasi-identifier	Sensitive	Classification
Age	✓		Numerical
Gender	✓		Numerical
Height	✓		Numerical
Weight	✓		Numerical
Cholesterol		✓	Numerical
Gluc		✓	Numerical
Smoke		✓	Numerical

The attributes of cardiovascular dataset are classified in Table 9. The proposed approach achieves improved privacy and classification utility and has comprehensive advantages compared to the available generalization methods. To implement the classification models, the adult dataset is divided into two-third portion 30,178, as the training data and the remaining 15,045 as the test data. In the cardiovascular study dataset, 46,000 tuples are considered as the training data and the remaining 24,000 as the test data. The proposed approach achieves a balanced classification utility and privacy when applied to both the datasets independently. The approach is compared with already existing standard methods namely IACK and Info Gain Mondrian.

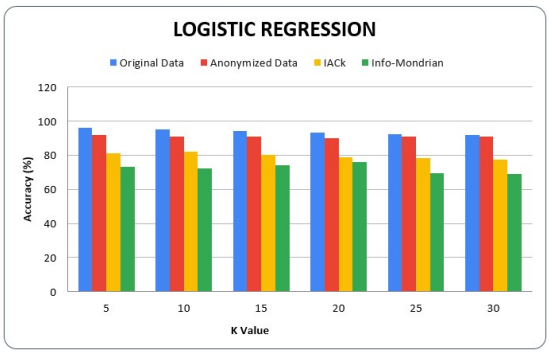
The proposed approach attains a consistent increase in accuracy compared to IACK and a significant increase than InfoGain Mondrian. The accuracy is compared with other models such as Random Forest, Logistic Regression, Adaboost, and Support Vector

Classifier for different k-values. The different k values scaling from 5-30 are used to show the classification accuracy in Figures 8 and 9. Compared to IACK and Info-Mondrian, the proposed approach has lower utility loss and attains greater privacy. Figure 8 shows the accuracy evaluations using Random Forest, Logistic regression, AdaBoost, and Support Vector Classifier compared with the original data, anonymized data, IACk and Info-Mondrain for the adult dataset.

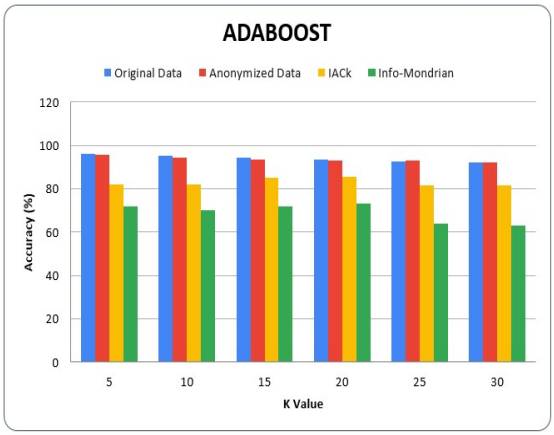
The random forest, Figure 8a, and adaboost, Figure 8c performs well in the proposed approach than the logistic regression Figure 8b and support vector classifier Figure 8d. Figure 9 shows the accuracy evaluations using Random Forest, Logistic Regression, Adaboost, and Support Vector Classifier compared with original data, anonymized data, IACk and Info-Mondrain for cardiovascular dataset.



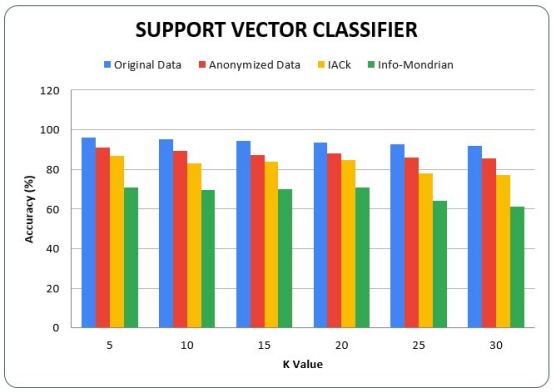
(a) Accuracy evaluations using RF



(b) Accuracy evaluations using LR



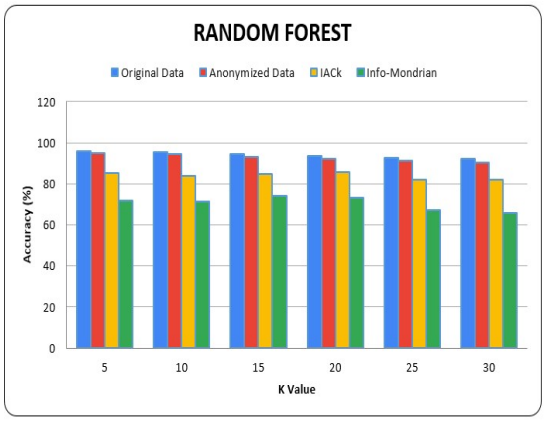
(c) Accuracy evaluations using AdaBoost



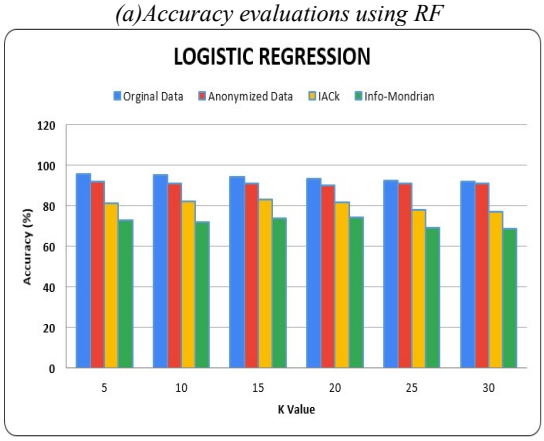
(d) Accuracy evaluations using SVC

Figure 8 Accuracies Evaluation: QIB-MHSAB, IACK, Info-Mondrian For The Adult Dataset.

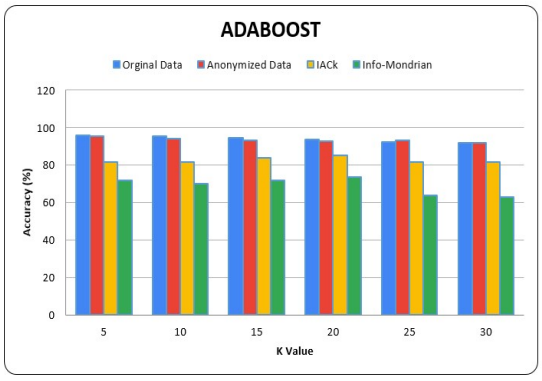
In the cardiovascular dataset, the random forest, Figure 9a and adaboost Figure 9c perform well compared to the logistic regression Figure 9b and support vector classifier Figure 9d. The proposed approach proved to protect the multiple sensitive attributes. The proposed hierarchical id-based generalization approach helps in protecting the individual's privacy with less information loss and for the better understanding of the disease data trends with reduced cost.



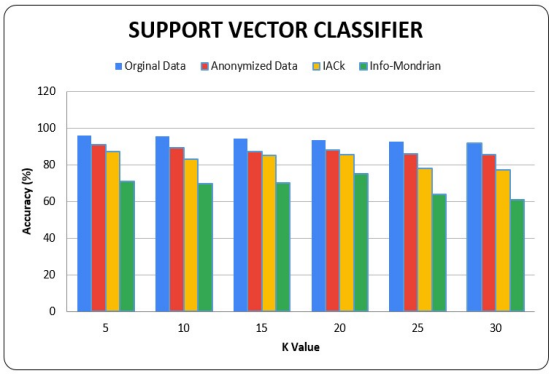
(a)Accuracy evaluations using RF



(b) Accuracy evaluations using LR



(c) Accuracy evaluations using AdaBoost.



(d) Accuracy evaluations using SVC

***Figure** 9: Accuracies Evaluation: QIB-MHSAB, IACK, And Info-Mondrian For Cardiovascular Dataset.*

8. CONCLUSION AND FUTURE DIRECTIONS

The privacy of health care data is a candid problem that requires special attention in the research community. Various types of researches have been done by implementing several anonymization method for protecting privacy with enhanced classification utility. However, most of the models could not achieve it. The paper has discussed several existing privacy models and its limitations to achieve a trade-off between privacy and utility. The need for security and privacy in the health care domain and the purpose of the user's privacy are also analyzed. Vertical partitioning is performed with a unique id in the original data which is used in linking the quasi-identifier attributes and sensitive attributes. The paper implemented two anonymization approaches: i) mean value replacement for numerical and id-based for the categorical quasi-identifier attributes and ii) hierarchical id-based generalization for the categorical sensitive attributes. The proposed hierarchical id-based generalization plays a significant role. The primary notion of the proposed approach is to classify the health care record attributes to apply the anonymization method appropriately. The proposed approach follows the concept of converting the categorical attribute into a numerical attribute. The proposed approach can be used in associated systems to protect users' privacy with improved classification utility. The experimented is conducted on anonymized data by building four existing classification models i) Random Forest ii) Logistic Regression iii) AdaBoost iv) Support Vector Classifier to evaluate the accuracy. As per the experimental results, the Random Forest, and the AdaBoost performs well

than the other two models. The work is just a beginning, and many things that require improvisation are considered as future directions. The limitations of the work are, it cannot be applied on streaming data and unstructured data. The proposed approach needs to be implemented in big dataset in order to check its better accuracy. Also, the proposed approach can be implemented across various applications apart from health care to achieve more significant result. The same anonymization model needs to be implemented on various datasets in future to check the proposed model is stable.

REFERENCES

[1] Ismail Keshta, Ammar Odeh, (2021) Security and privacy of electronic health records: Concerns and challenges, Egyptian Informatics Journal 22(2): 177-183.

[2] Mohanad Dawoud, D.Turgay Altilar, (2017) Cloud-Based E-Health Systems: Security and Privacy Challenges and Solutions, International Conference on Computer Science and Engineering, IEEE Explore 2017: 861-865. 10.1109/UBMK.2017.8093549

[3] Clemens Scott Kruse, Micheal Mileski, Alekhya Ganta Vijaykumar, Sneha Vishnampet Viswanathan, Ujwala Suskandla, and Yazhini Chidambaram, (2017) Impact of electronic health records on long-term care facilities: Systematic review, JMIR Medical Informatics; 5(3). 10.2196/medinform.7958

[4] Nureni Ayofe Azeez , Charles Van der Vyver, (2019) Security and privacy issues in e-health cloud-based system: A comprehensive content analysis, Egyptian Informatics Journal;20(2): 97-108. https://doi.org/10.1016/j.eij.2018.12.001

[5] ThomasAsikis, Evangelos Pournaras, (2020) Optimization of privacy-utility trade-offs under informational self-determination, Future Generation Computer Systems; 109: 488-499. https://doi.org/10.1016/j.future.2018.07.018

[6] Chaobin Li,u Shixi Chen, Shuigeng Zhou, Jihong Guan, YaoMa , (2019) A novel privacy preserving method for data publication, Information Sciences; 501: 421-435.

[7] Zhitao Guan, Zefang Lv, Xiaojiang Du, Longfei Wu, Mohsen Guizani, (2019) Achieving data utility-privacy trade-off on Internet of Medical Things : A machine learning approach, Future Generation Computer Systems;98:60-68. https://doi.org/10.1016/j.future.2019.01.058

[8] Jiuyong Li, Raymond Chi-Wng Wong, Ada Wai-Chee Fu, Jian Pei, (2008) Anonymization by local recoding in data with attribute hierarchical taxonomies, IEEE Transactions on Knowledge and Data Engineering;20: 1181–1194. 10.1109/TKDE.2008.52

[9] Jian Xu, Wei Wang, Jian Pei, Xiaoyuan Wang, Baile Shi, Ada Wai-Chee Fu, (2006) Utility-based anonymization using local recoding, ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: 785–790. https://doi.org/10.1145/1150402.1150504

[10] Hui Wang, Ruilin Liu, (2011) Privacy-preserving publishing microdata with full functional dependencies, Data & Knowledge Engineering; 70: 249–268. https://doi.org/10.1016/j.datak.2010.11.002

[11] Benjamin C. M. Fung, Ke Wang, Ada Wai-Chee Fu, and Philip S Yu, (2011) Introduction to Privacy Preserving Data Publishing Concepts and Techniques,.

[12] Latanya Sweeney, (2002) K -anonymity: A model for protecting privacy. International Journal Uncertainty Fuzziness Knowledge Based System; 10 (5): 557–570.

[13] Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramaniam, M, (2006) *l* -diversity: Privacy beyond k -anonymity. 22nd International Conference on Data Engineering: 1-24

[14] Ninghui Li, Tiancheng Li, Suresh Venkatasubramanian, (2010) Closeness: a new privacy measure for data publishing. IEEE Transactions on Knowledge and Data Engineering; 22 (7): 943–956. 10.1109/TKDE.2009.139

[15] Mehmet Ercan Nergiz, Maurizio Atzori, Chris Clifton, (2007) Hiding the presence of individuals from shared databases, ACM SIGMOD International Conference on Management of Data: 665–676. https://doi.org/10.1145/1247480.1247554

[16] Bee-Chung Chen, Kristen LeFevre Raghu Ramakrishnan, (2007) Privacy skyline: Privacy with multidimensional Adversarial knowledge, 33rd International Conference on Very Large Databases: 770-781.

[17] Chaobin Liu, Shixi Chen, Shuigeng Zhou, Jihong Guan, Yao Ma, (2021) A general framework for privacy-preserving of data publication based on randomized response techniques, Information Systems; 96: 1-11. 10.1016/j.is.2020.101648

[18] Yufei Tao, Hekang Chen, Xiaokui Xiao, Shuigeng Zhou, (2009) ANGEL: Enhancing the Utility of Generalization for Privacy Preserving Publication, IEEE Transactions on Knowledge and Data Engineering; 21(7), 1073-1087. 10.1109/TKDE.2009.65

[19] Xiaokui Xiao Yufei Tao, (2006) Anatomy: Simple and Effective Privacy Preservation, Proceedings of the 32 International conference on Very Large Databases: 139-150.

[20] Qinghai Liu , Hong Shen , Yingpeng Sang, (2015) Privacy-preserving data publishing for multiple numerical sensitive attributes, Tsinghua Science and Technology; 20 (3): 246–254 . 10.1109/TST.2015.7128936

[21] Qinghai Liu, Hong Shen, Yingpeng Sang, (2014) A Privacy-preserving Data Publishing Method for Multiple Numerical Sensitive Attributes via Clustering and Multi-Sensitive Bucketization, , 6th International Symposium on Parallel Architectures, Algorithms and Programming: 220-223. 10.1109/PAAP.2014.56

[22] Dharavathu Radha and Prof. Valli Kumari Vatsavayi, (2017) Bucketize: Protecting Privacy on Multiple Numerical Sensitive Attribute, Advances in Computational Sciences and Technology;10(5); 991-1008.

[23] Yuelei Xiao and Haiqi Li, (2020) Privacy Preserving Data Publishing for Multiple Sensitive Attributes Based on Security Level, MDPI Information; 11: 1-27. https://doi.org/10.3390/info11030166

[24] Zakariae El Ouazzani and Hanan El Bakkali, (2018) Proximity Test for Sensitive Categorical Attributes in Big Data, 4th International Conference on Cloud Computing Technologies and Applications (Cloudtech):1-7. 10.1109/CloudTech.2018.8713359

[25] Jayapradha. J*, Prakash. M, Yenumula Harshavardhan Reddy, (2020) Privacy Preserving Data Publishing for Heterogeneous Multiple Sensitive Attributes with Personalized Privacy and Enhanced Utility, Systematic Review Pharm ; 11(9):1055-1066. 10.31838/srp.2020.9.151

[26] Ashoka K, Dr. Poornima B, (2017) Enhanced Utility in Preserving Privacy for Multiple Heterogeneous Sensitive Attributes using Correlation and Personal Sensitivity flags, International Conference on Advances in Computing, Communications and Informatics: 970-976. 10.1109/ICACCI.2017.8125967

[27] Fangwei Luo, Jianmin Han, Jianfeng Lu and Hao Peng, (2013) ANGELMS: a privacy preserving data publishing framework for microdata with multiple sensitive attributes, Third international conference on information science and technology: 393–398. https://doi.org/10.1109/ICIST.2013.6747576

[28] Adeel Anjum, Naveed Ahmad, Saif U. R. Malik, Samiya Zubair, Basit Shahzad, (2018) An efficient approach for publishing microdata for multiple sensitive attributes, The Journal of Supercomputing; 74:5127-5155. https://doi.org/10.1007/s11227-018-2390-x

[29] Sowmyarani, C.N.; Srinivasan, G.N. (2015) A robust privacy preserving model for data publishing, International Conference on Computer Communication and Informatics: 1-6. 10.1109/ICCCI.2015.7218095

[30] Jayapradha. J*, Prakash. M, (2021) An efficient privacy-preserving data publishing in Health care records with multiple sensitive attributes, 6th International Conference on Inventive Computation Technologies: 623-629. 10.1109/ICICT50816.2021.9358639

[31] Xinning Li, Zhiping Zhou, (2020) A generalization model for multi-record privacy preservation, Journal of Ambient Intelligence and Humanized Computing; 11:2899–2912. https://link.springer.com/article/10.1007/s12652-019-01430-y

[32] M. Prakash* and G. Singaravel, (2018) Haphazard, enhanced haphazard and personalised anonymisation for privacy preserving data mining on sensitive data sources, International Journal of Business Intelligence and Data Mining; 13(4): 456-474. https://doi.org/10.1504/IJBIDM.2018.094983

[33] Jiuyong Li , Jixue Liu, Muzammil Baig and Raymond Chi-Wing Wong, (2011) Information based data anonymization for classification utility, Data & Knowledge Engineering; 70: 1030–1045. 10.1016/j.datak.2011.07.001

[34] Bee-Chung Chen, Kristen LeFevre Raghu Ramakrishnan, (2007) Privacy skyline: Privacy with multidimensional Adversarial knowledge, 33rd International Conference on Very Large Databases: 770-781.

[35] B.C.M. Fung, K.Wang, P.S.Yu. (2007) Anonymizing Classification data for privacy preservation, IEEE Transactions on Knowledge and Data Engineering; 19 (5): 711–725. 10.1109/TKDE.2007.1015

[36] Slava Kisilevich, Lior Rokach, Yuval Elovici, Bracha Shapira, (2010) Efficient multidimensional suppression for k-anonymity, IEEE Transaction on Knowledge and Data Engineering; 22 (3) : 334–347. 10.1109/TKDE.2009.91

[37] Ninghui Li, Tiancheng Li, Suresh Venkatasubramanian (2007) t-closeness: privacy beyond k-anonymity and l-diversity, International Conference on Data Engineering, IEEE Computer Society, 106–115.

[38] Raymond Chi-Wing Wong, Jiuyong Li, Ada Wai.-Chee Fu, Ke Wang, (2006) (α, k)-anonymity: An enhanced k-anonymity model for privacy preserving data publishing, ACM International Conference on Knowledge Discovery and Data Mining ; 754–759. https://doi.org/10.1145/1150402.1150499

[39] Abdul Majeed, (2019) Attribute-centric anonymization scheme for improving user privacy and utility of publishing e-health data, Journal of King Saud University – Computer and Information Sciences; 31(4): 426–435, https://doi.org/10.1016/j.jksuci.2018.03.014

[40] Zude Li, Guoqiang Zhan, and Xiaojun Ye, (2006) Towards an Anti-inference (K,l)-Anonymity Model with Value Association Rules, International Conference on Database and Expert Systems Applications: 883-893. https://doi.org/10.1007/11827405_86