Improving Privacy Preservation Approach for Healthcare Data using Frequency Distribution of Delicate Information

Ganesh Dagadu Puri, D. Haritha
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Vaddeswaram, AP, India

***Abstract*—In the modern world, everyone wishes that their personal information wouldn't be made public in any manner. In order to keep personal information hidden from prying eyes, privacy protection is essential. The data may be in the form of big data and minimization of risk and protection of sensitive data is important. In this research, a revolutionary customized privacy-preserving method is implemented that addresses the drawbacks of earlier personalized privacy as well as other anonymization methods. There are two main components that make up the proposed method's core. Delicate Information and Delicate Weight are two additional attributes which are used in the record table, are covered in the first section. The record holder's Delicate Information (DI) decides whether or not secrecy should be kept or if it should be shared. How delicate an attribute value is compared to the rest is indicated by its Delicate weight (DW). The second part covers a new representation used for anonymization termed the Frequency Distribution Block (FDB) and Quasi-Identifier Distribution Block (QIDB). According to experimental findings, the proposed system executes more quickly and with less data loss than current approaches.**

Keywords*—Privacy preservation approach; quasi identifier distribution block; frequency distribution block; big data; anonymization

I. Introduction

Electronic Medical Records (EMRs) are currently widely used in healthcare networks. It makes it possible for people to easily and adaptably exchange their medical data. For instance, instead of needing to search through multiple physical records, a patient or his/her physician merely needs to access the data from a database to locate their diagnostic report. Advanced electronic medical record systems face a significant issue when it comes to securely storing and accessing electronic medical records because healthcare information is so sensitive [1]. Hadoop and big data analytics play a significant part in analyzing and processing the patient information in many forms to provide potential uses [2]. Investigation can leverage private data from several organizations to identify patterns. For instance, if a patient's private data is available across various hospitals, researchers can utilize it to better understand the patterns associated with a given disease and, as a result, make a more accurate diagnosis. The unprocessed information found in hospitals includes specific information on the patient, such as identity, address, date of birth, zip code, symptoms and illness[3]. Before being delivered to the data receiver, the name and residential address information that are deemed private are stripped from the raw data which is also known as micro data. Furthermore, this micro data includes information like postal code and date of birth that can be connected to other external, publicly accessible data bases to re-identify sensitive value[4]. Linking attack refers to the process of re-identifying a record by connecting published data to publicly available data. Let us consider the patient records released by the hospital in Table I, for instance, which excludes data like name, residential data, and other private details. By joining the information from Table I with the publicly accessible external data base given in Table II, the intruder can disclose personal information. The query may appear like,

Select name, disorder from external_table as A, patient_table as B where A.postal=B.postal and A.age=B.age;

Since people are reluctant to volunteer their private data, it is extremely concerning that the answer to this query provides complete data about the illness and the name of the person. The join, which is referred to as Record Level Disclosure, may provide a value for age 36 and postal code 38677. Researchers employ techniques categorized as Privacy Preserving Data Publishing (PPDP) to hide confidential material from recipients. Quasi-Identifier (Q) attributes are characteristics found in Released Patient Data that can be connected to external, publicly accessible data bases, such as Postal Code, Date of Birth, etc. Data is modified in a way that leads to duplicate rows in the resulting table, limiting disclosure. Through the use of generalization, there has to be more than one implicit connection to the external data base. Thus, the k-anonymity algorithm is implemented for measuring this. Each entry in a table is indistinguishable from minimum k-1 other entries with regard to each and every set of quasi-identifier attributes if it fulfills the k-anonymity condition; such a table is known as a k-anonymous table.

With personalized anonymization, a guard node is utilized to determine if the record holder is willing to disclose the level of sensitivity upon which the anonymization will be carried out. As the record owner sensitivity is a generic one, the majority of the sensitive values that are included in the secret data base do not necessitate privacy protection. Therefore, just a small portion of the distribution's records need to be private. For instance, a record holder with malaria will not really mind

sharing his identity, in contrast to a record holder with HIV. The fact that some HIV-positive record owners are willing to expose their identities justifies this proposed privacy preservation strategy.

In other words, every group of quasi-identifier values needs a minimum k-1 records, and they can be tricked by connecting a record from the disclosed data to a database with many entries that is publicly accessible. A two-anonymous generalization for Table I is shown in Table III. Assuming that the intruder uses a publicly accessible database and discovers that Ramesh is 36 years old with a postal code of 38677 and that he has a disorder, the intruder looks at anonymized Table III and learns that 38677 and 36 have been generalized to 386** and [30-40] which can be associated with two entries of releases table and that the disorder cannot be derived from this information. Lungs disease has been hidden and is not intended for publishing in this table (<386**,[50-60],Lung Disease>). Similar findings occur if the intruder attempts to infer Sitaram's illness, which belongs to category 3, but since every member of the category possesses the same sensitive property, the attacker deduces that Sitaram has fever.

Attribute level disclosure results from this release of confidential information. This occurs when a set of disorders are indeed symptoms of the same condition. To tackle this issue l-diversity was introduced. If the sensitive characteristic has at minimum l "well-represented" values, then an equivalence class has l-diversity. If each equivalence class in a table possesses l-diversity, the table has l-diversity. Additionally, skewness and similarity attacks are a drawback of l-diversity. Proximity was viewed as a method of overcoming this. The distribution of sensitive attributes in this strategy must match the anonymized chunk. Thus, there is a data loss.

In this paper, the research work is arranged into five sections. In Section 2, related work of various researchers and research limitations are described in detail. In Section 3, our proposed model is discussed. Experimental findings and discussion of each test is described in Section 4. Thus, in Section 5, research work is concluded and future scope of work is discussed.

TABLE I. PATIENT RELEASED DATA

Postal Code	***Age***	***Disorder***
38677	36	Mouth ulcer
38602	38	Brain cancer
38678	42	Fever
38685	46	Fever
38905	52	Fever
38906	56	Fever
38909	53	Fever
38673	58	Lungs Disease
38607	65	Lungs Disease
38655	68	Brain cancer

TABLE II. EXTERNAL DATABASE

Name	***Postal Code***	***Age***
Ramesh	38677	36
Laxmi	38677	45
Suresh	38602	38
Nageshrao	38602	32
Anupama	38678	42
Sitaram	38905	52
Kishor	38909	53
Vijay	38906	56
--	--	--

TABLE III. ANONYMOUS DATA

Postal Code	***Age***	***Disease***
386**	[30-40]	Mouth ulcer
386**	[30-40]	Brain cancer
386**	[40-50]	Fever
386**	[40-50]	Fever
389**	[50-60]	Fever
389**	[50-60]	Fever
389**	[50-60]	Fever
386**	[50-60]	Lungs Disease
386**	[60-70]	Lungs Disease
386**	[60-70]	Brain cancer

II. LITERATURE SURVEY

Two anonymous techniques were presented by Xingguang Zhou et al. [5] that not only ensure data secrecy but also realize anonymity for patient. When attackers select attack destinations before gathering data from the electronic health record, the first strategy obtains modest security. The second strategy ensures total security by having attackers select attack targets in an adaptive manner upon contact with the electronic medical record system. It also suggested a method for EMR holders to use an anonymous search engine to find their electronic health records. As per Safa Bahri et al. [6] enormous amount of information, especially clinical data, has recently been amassed as a result of the intensification of emerging innovations that the large majority of people in the globe have accepted. Medical associations have acquired and analysed this clinical information and gain information and ideas that may be used to a variety of clinical judgments, including recommendations for medications and improved diagnoses. This paper mentions the significant effects that Big Data has on healthcare stakeholders, including patients, doctors, pharmaceutical and medical technicians, and medical insurance companies. It also examines the various difficulties that must be overcome in order to maximize the advantages of all the Big Data and the software that are presently accessible. Such large data can be stored on the devices customized to application processing [7].

A Secured as well as Anonymous Biometric Based User Authentication technique is introduced by B D Deebak et al. in 2017 [8] to guarantee secure data transmission in medical applications. This study demonstrates that a hostile cannot pretend to be a registered user in order to get unauthorized entry to or revoke an intelligent mobile card. For the purpose of demonstrating security and energy efficiency in healthcare application systems, a formal study relied on the random-oracle approach and resource evaluation is presented. The suggested method also incorporates some efficiency study to demonstrate that it offers high-security characteristics for developing intelligent medical application systems in the IoM. In 2019, Jorge Bernal Bernabe et al. [9] conducted a thorough evaluation of the State-of-Art (SoA) for privacy-preserving research approaches and methods in blockchain, and also the primary connected privacy issues in this exhilarating and disruptive technology. The survey includes privacy strategies in permissioned and privatized blockchains along with privacy-preserving research report and methods in accessible and private blockchain, such as Bitcoin and Ethereum. The analysis of various blockchain use cases includes looking at areas including Electronic-Government, Electronic-Health, crypto currency, developed cities, and cooperative ITS.

A Privacy-Preserving-Reinforcement-Learning (PPRL) architecture for the cloud computing system is proposed by Jaehyoung Park et al. in 2020 [10]. The proposed methodology makes use of learning with errors based cryptosystem for completely homomorphic encryption. Various cloud computing dependent intelligent service contexts are used to carry out effective analysis and assessment for the developed PPRL architecture. A stateless cloud monitoring approach for non-manager adaptive group data with preserving the privacy is proposed by Xiaodong Yang et al. [11]. With the random masking approach, the proposed methodology not only achieves individual identity privacy preservation but also data confidentiality preservation. Marwa Keshk et al. [12] present a thorough analysis of the most recent privacy-preserving methods for defending Cyber Physical System (CPS) technologies and their data against online threats in 2021. The ideas of privacy preservation and CPSs are examined, with an emphasis on the parts of cyber physical systems and how these systems might be hacked physically or digitally. Abdullah Al Omar et al. [13] presented an approach for the healthcare system which ensures data security and transparency. Additionally, the Ethereum platform is used to integrate insurance policies into the suggested system's blockchain, and cryptographic techniques are used to protect private information.

A mathematical formulation for an identity-based encryption strategy for the protection of patient confidentiality during the gathering of clinical records for evaluation is presented by Kissi Mireku Kingsford et al. in 2017[14]. The submission of medical data for analysis is becoming an essential element of daily life. To protect the confidentiality of patient, the model dissociates the identity of the patient from the investigated data upon data submission. A thorough analysis of privacy protection in big data from the communication point of view is presented by Tao Wang et al. in 2018[15]. It focuses on privacy-preserving methods, especially differential privacy, and the basic privacy-preserving paradigm. Additionally, it examines the difficulties with differential privacy as well as its variations and modifications for various novel apps. Muneeb Ul Hassan et al. [16] have performed a detailed analysis of differential privacy approaches for CPSs as presented in 2019. Specifically, it looks at how differential privacy is used and implemented in four key CPS uses: energy, medical, transportation, healthcare & industrial Internet of things. It also outlines unresolved problems, difficulties, and prospective research directions for CPS differential privacy approaches. This investigation can be used as the foundation for the creation of cutting-edge differential privacy methods to handle numerous issues and CPSs' data privacy contexts.

The privacy of Kim's approach was assessed by Kefei Mao et al. [17], who show that the plan is actually vulnerable to the stolen smart - card threat. The plan also has some impassable stages, and the privacy assumption is excessively rigid. In addition, a novel technique built on Kim's as well as the quadratic residue hypothesis is investigated. In contrast to the current plans, the latest proposal does not call for the electric medical record database to personally communicate different secure values with patients and physicians. As a result, it is more useful and practical. It demonstrates that the suggested approach can offer greater protection than Kim's earlier plan. A unique architecture to enable privacy-preserving Machine learning (ML) was proposed by Kaihe Xu et al.[18], where the training data are spread and every shared data chunk is of enormous volume. To accomplish privacy preservation, it actually makes use of the Apache Hadoop platform's data locality attribute and just a few cryptographic functions at the Reduce functions. The comprehensive simulations used to show the presented strategy's robustness and consistency demonstrates that it is safe in the semi-honest framework.

In 2017, Tanashri Karle et al. [19] focused on protecting privacy by utilizing an anonymization methodology and a thorough investigation of two anonymization techniques are discussed namely - Datafly Technique and the Mondrian Algorithm. While Mondrian method is more suited for real datasets, Datafly technique is better suited for synthetic datasets. By using privacy preservation on a medical dataset in 2017, Balaji K. Bodkhe et al. [20] preserve a person's identity and any associated disorders (sensitive feature). The techniques including slicing, generalization, suppression and bucketization are utilized. These techniques guarantee privacy preservation while maintaining the usefulness of the data. The goal of S.Sathya et al. [21] is to take advantage of the new privacy difficulties posed by big data and focus on effective, privacy-preserving computation in the big data era. In order to address the effectiveness and privacy needs of Data Mining (DM) in the big data era, it first formalizes the overall framework of big data analytics, identifies the related privacy requirements, and introduces an effective and Privacy-Triple-DES as an instance.

To minimize and protect the data from unwanted parties, S. Shimona et al. [22] offer the PPDM strategies in a concise manner together with other privacy preservation measures in 2020. In 2020, Suneetha V et al. [23] introduced a unique concept called spark that uses Apache Spark to manage big data in the health care industry quickly and effectively while

using K-anonymization as well as L-diversity to disguise private data. The suggested method ensures that shared information will not reveal the actual information and that sensitive data is separated before being sent to Hadoop distributed file system. In 2019, Hui Jiang et al. [24] highlighted the fundamental steps of Hadoop-based big data analysis and included technical recommendations for common actual and off-line application scenarios. These recommendations were based on a review of the ecological structure of Hadoop. In order to have some reference value for the development of a big data platform and for the analysis and processing of huge data, Hadoop was utilized to construct the application context and the WordCount scenario was merged to assess the MapReduce calculation procedure.

A cooperation privacy preservation strategy for wearable technology was developed in 2018 by Hong Liu et al. [25] with id validation and data access control concerns in the space and time-aware settings. To obtain a secure healthcare pathway query under e-medical cloud servers without disclosing the secret data of patients like name, sex, age, location and also the information of hospitals like diagnosis, medication, and cost. Mingwu Zhang et al. [26] suggested a Privacy Preserving Enhancement of medical pathway query method. To maintain confidentiality in the e-Healthcare system, the suggested methodology first develops a number of privacy-preserving protocols like privacy-preserving medical comparison, privacy-preserving phase selection, and privacy-preserving phase update. It then implements the greedy approach in a secure way to carry out the query as well as the Min-Heap innovation to make it more efficient. This approach is feasible and effective with regard to computational time and cost, according to test findings. In 2018, Abdulatif Alabdulatif et al. [27] set out to propose a cloud-based solution for real-time patient monitoring that protects user privacy by spotting changes in a variety of important health indicators of participants of smart communities. IoT-enabled wearable devices' produced vital sign information is analysed in real-time on the cloud. The construction of a predictive method for the smart community while taking into account the sensitivity of information processing in a third-party context is the main topic of this paper (e.g., cloud computing). For enabling data prediction with patterns, it designed a crucial sign change detection method employing Holt's linear trend approach, where completely homomorphic encryption technique is applied to carry out calculations on an encrypted area that may protect data privacy. Additionally, a parallel strategy for encrypted operations using the MapReduce method of Apache Hadoop was proposed in order to minimize the burden of the completely homomorphic encryption technique across massive healthcare data.

The difficulties and needs of creating frameworks and procedures for globally distributed data processing are investigated and discussed by Shlomi Dolev et al. in 2017 [28]. It categorizes and studies the overhead problems associated with batch, stream and SQL-style processing using geo-distributed architectures, methods, and techniques. Using differential privacy, Miao Du et al. [29] present and put into practice a ML technique for smart edges in 2018. In a wireless big data situation, anonymization in training datasets is the main priority. Additionally, it designs two distinct techniques, Output and Objective Perturbation which fulfill differential privacy, and guarantees privacy and security by including Laplace techniques. Additionally, for correlated datasets, differential privacy preservation algorithms are offered, providing privacy through theoretical inference. Ultimately, tests were conducted using TensorFlow and the effectiveness of the technique was assessed using the four datasets STL-10, SVHN, MNIST and CIFAR-10. The suggested approach effectively ensures accuracy upon benchmark datasets while safeguarding the confidentiality of training datasets.

A scalable approach to the local-recoding issue for big data anonymization over proximity privacy violations was investigated by Xuyun Zhang et al. [30]. The study proposes a proximity privacy framework that provides the semantic proximity of sensitive values including numerous sensitive attributes. It also models the local recoding issue as a proximity-aware clustering issue. It presents a scalable two-phase clustering method that combines the proximity-aware agglomerative clustering technique and the t-ancestors clustering technique. The methods were created using MapReduce to provide good scalability using cloud-based data-parallel processing. Numerous tests using real data sets show that the method greatly outperforms existing methods in terms of scalability, time efficiency, and capacity to fight against proximity information leakage.

As per Haiping Huang et al. [31], Electronic-healthcare has substantially benefited from the industrialization of cloud computing, Internet of things and Wireless-body-Area-Networks (WBANs). Furthermore, there are still several obstacles standing in the way of e-Healthcare's growth, especially issues with data security and privacy protection. Healthcare system architecture is formulated to overcome these issues. It gathers health information from WBANs, transfers it across a substantial wireless sensor network, and then releases it into Wireless-Personal-Area-Networks (WPANs) through a gateway. Additionally, healthcare system uses the Homomorphic Encryption Dependent on Matrix scheme to assure confidentiality, the Groups of Send-Receive Model strategy to accomplish key distribution, and an intelligent system capable of autonomously analyzing the encrypted health data and reporting the findings. The confidentiality, privacy, and improved efficiency of healthcare system are evaluated theoretically and experimentally in comparison to existing systems or techniques. Lastly, the practicality of the healthcare system prototype implementation is examined. A privacy-preserving approach is put forth by Marwa Keshk et al. in 2019 [32] in order to obtain both safety and confidentiality in intelligent power networks. A two-level privacy component and an anomaly detection component are the framework's two core components. Using open datasets, the outlier detection module trains and validates the outcomes of the two-level privacy component using a Long-Short-Term-Memory DL approach. In contrast to various cutting-edge methodologies, the experiments demonstrated that the proposed architecture can effectively secure data of intelligent power networks and identify anomalous behaviors. The term "optimal distributed estimate" refers to a conceptual framework created by Jianping He et al. in 2018 [33] to examine how to maximize the

assessment of a neighbor's original data using the collected local data. The disclosure probability is then looked into as part of the best estimation for the data privacy evaluation. The privacy-preserving average consensus method's data privacy has been further examined using the established framework, and the best noises for the technique are identified.

In 2018, Weichao Gao et al. [34] used the idea of homomorphic encryption as well as secured network protocol development to tackle the issues of privacy preservation for information auction in CPS. A general Privacy-Preserving Auction Strategy is put forth, in which an unreliable third-party trade platform is made up of the two distinct entities of the auctioneer and interim platform. A winner in the auction procedure is defined and all bidder data is hidden by using homomorphic encryption as well as a one-time pad. However, it also suggests an Enhanced Privacy Preserving Auction Method that makes use of an extra signature verification technique in order to increase the overall security of the privacy preserving auction. Each strategy's viability is confirmed through in-depth theoretical analysis and thorough performance tests, which also include an examination of attack tolerance. A unique privacy-preserving anomaly - based detection methodology, known as PPAD-CPS, is suggested by Marwa Keshk et al. in 2018 [35] for safeguarding private data and identifying hostile findings in power technology and associated network traffic. There are two primary components in the architecture. In order to meet the goal of privacy preservation, a data pre-processing component is first proposed for filtering and changing original information into a new format. Secondly, an anomaly-based detection component utilizing a Kalman Filter as well as Gaussian Mixture Model for accurately predicting the posterior probabilities of normal and malicious events is proposed. Two open datasets, the Power System as well as UNSW-NB15 dataset, are used to test the efficiency of the architecture.

III. Proposed System

The privacy-preserving method we propose overcomes the drawbacks of existing techniques and other anonymization methods. There are two main parts that make up the proposed method's core. The first part of the equation concerns with additional attributes utilized in the table namely Delicate Information and Delicate weight. The DI indicates whether the privacy of the record owner's private data should be protected or released. DW determines the sensitivity of the attribute. DW is necessary for DI.

When the person provides their data, DI can be accessed easily from them. DW could be based on previously acquired sensitive attribute information. The same level of protection is provided for every sensitive attributes by conventional privacy approaches, which has been addressed in this approach by the implementation of DI and DW. The flag DI=0 indicates that the entry holder is not willing to share his confidential attribute, while DI=1 indicates that he has no problem doing so. The publisher has highlighted DW for any sensitive attributes where confidentiality is crucial. For instance, a record holder with the fever or gastroenteritis is less reluctant to expose his identify than a label owner with cancer. Whenever the sensitive attribute is a very common disorder like the fever or mouth ulcer, DW=0 is being used; for a sensitive attribute like brain cancer, which is uncommon, DW=1 is utilized. For DW=0, DI has a default value of 1, and for DW=1, the record holder's DI values are accepted.

TABLE IV. DW for Disorders

Disease	***DW***
Mouth ulcer	0
Brain cancer	1
Fever	0
Lungs Disease	1

The second section discusses a novel approach for evaluating the distribution known as the FDB and QIDB. Each disorder's spread in the FDB is based on original, personal data. QIDB is formed for each entry with DW=1 and DI=0. Several QIDB chunks will exist. These chunks are needed to make sure that each particular QIDB and distribution of FDB is synchronized.

TABLE V. Patient Released Data with DW and DI

Postal Code	***Age***	***Disorder***	***DW***	***DI***
38677	36	Mouth ulcer	0	1
38602	38	Brain cancer	1	0
38678	42	Fever	0	1
38685	46	Fever	0	1
38905	52	Fever	0	1
38906	56	Fever	0	1
38909	53	Fever	0	1
38673	58	Lungs Disease	1	1
38607	65	Lungs Disease	1	0
38655	68	Brain cancer	1	1

TABLE VI. Frequency Distribution Block

Disease	***Probability***
Mouth ulcer	0.1
Brain cancer	0.2
Fever	0.5
Lungs Disease	0.2

A. Model and Terminology for Proposed Personalized Privacy

Let R be a connection providing personal information about a set of people. There are four groups of attributes in R.

- Unique Identifiers Uj - It can be used to identify individuals who are eliminated from R.
- Quasi identifiers QIj - its value can be combined with publicly available information to determine a person's identity.

- Delicate attributes Dj – It is secretive or delicate to the record holder.
- Non quasi identifiers NQIj – It doesn't fall into any of the three categories.

The goal of proposed method is to obtain a generalized table R* such that distribution of every QIDB is comparable to the diversity of the entire distribution as seen in FDB. For ease of use, the full set of quasi identifiers is denoted by QI, and its values by q. In a similar manner, there is a single delicate attribute Di and its value d. Relation R comprises of m number of tuples R={r1,r2,…,rn}. Record holder data can be obtained by referring as rj.d to represent delicate value and rj.q for quasi identifier value $1 \leqslant j \leqslant m$.

1) Delicate Weight- for every tuple $r \in R$, its delicate weight is added. This value is derived from Relation W(ds,dw) where ds indicates disorder and dw indicates delicate weight. W contains p records

rj.dw={ wj.dw if wj.ds=ri.d $1 \leqslant i \leqslant p$ } for every $1 \leqslant j \leqslant m$

Table IV provides the dw value for every disorder. Table I is used to create this distribution.

2) Delicate Information - for every tuple $r \in R$, its Delicate Information is indicated as r.di.

rj .di= {1 if ri .dw=0 ud rj.dw=1 } for every $1 \leqslant j \leqslant m$

The value of user defined (ud), is either 0 or 1. If the value of ri.di is zero, the user is not prepared to share his information, and if it is one, the user agrees.

Table V shows the values of dw and di assuming that the record holder will approve di value for DW=1. Additionally, it can be seen that if dw=0, the corresponding di is set to 1, showing that the entries' sensitivity is not really important.

3) Thresholds - To improve and enhance effectiveness of disclosure, generalization, and suppression, values of threshold are established for a number of personalized privacy aspects.

- Tn - It indicates minimum number of entries in R.
- Titr - It indicates maximum number of required iterations.
- Tsup - It indicates minimum number of delicate values for suppression.
- Tdis – It indicates minimum number of delicate values for disclosure
- Tacc – It indicates minimum number of thresholds for addition or subtraction.

Several threshold values are suggested because the dispersion aspect is being taken into account. The first value, which was never specified in the earlier representations, denotes the bare minimum number of item sets that must be provided in order to execute anonymization. Titr is calculated using information of the Value domain hierarchy's height. The generalization is greater and information loss is correspondingly greater when the value of Titr is high. Tsup denotes the absolute minimal amount of sensitive distribution that could exist in QIDB for that block's deletion following Titr. The threshold value Tdis represents the amount that can be added or removed from every frequency distribution for every disorder in order to make it equal to the FDB distribution. The frequency of QIDB and FDB will not be completely the same, thus while examining the distribution of every disorder is examined if the frequency in that qidb.v.d± Tacc always Tdis > Tacc.

4) Frequency Distribution Block - Distribution of every wj.ds in regards to the original distribution ri.d is stored in relation FDB(ds,p) where d represents disorder and p represents probability distribution of it Every p for ds is computed by mapping every ds in R (values of ri.d=fdbv.ds) to the total no. of tuples in R, for every $1 \leqslant v \leqslant k$. Considering there are m entries in the relation.

5) Quasi ‑ Identifier Distribution Block- for every rj.d where rj.dw=1 & r.di=0 a new QIDB is generated comprising ri.s for every $1 \leqslant j \leqslant m$. The relation QIDB.V(q,d) where qidb.vl.q=rj.q & qidb.vl.d=rj.d. Considering there are m QIDB chunks.

TABLE VII. QIDB.1 DATA

Postal Code	***Age***	***Disorder***
38602	38	Brain cancer

TABLE VIII. QIDB.2 DATA

Postal Code	***Age***	***Disorder***
38607	65	Lungs Disease

Table VI illustrates the frequency distribution of every disorder. This distribution demonstrates that the fever is a widespread disorder with a higher frequency—roughly 50 percent in the reported data. Every QIDB maintains the exact similar distribution. Due to the fact that the quasi values <38602, 28> and <38607, 55> have the DW and DI values of 1 and 0 respectively, in the first cycle 2 blocks of QIDB will be produced for these values as shown in Table VII and Table VII. Table VII shows Brain cancer disease probability is 0.2 in distribution block. In the same way Table VIII shows Probability of Lungs disease is 0.2 in the frequency distribution block. It is calculated from delicate weight of delicate information.

6) Generalization - A generalization function provides the general domain of an attribute R.Q. Function will return a generalized value in the domain provided a value r.q in the original domain.

7) Check Frequency- for every QIDB, examine CFq(QIDB.V) with QIDB.V FD which is equal to the FD in FDB. It is performed as follows

Let c be the total number of entries in QIDB.V for every UNIQ(qidb.vl.d) obtain total number of mappings which match qidb.vl.d to the total number of entries that is x in QIDB.V, thus CFq will return true if

For every $1 \leqslant v \leqslant m$ such that fdbv.ds=qidb.vl .d

fdbv.p = (unique (qidb.vl .d) / x) ± Tacc

This is examined in each cycle if a QIDB satisfies the FD then this chunk won't be taken into account for the next iteration.

8) Suppression- After Titr iterations, SUP(QIDB.v) remove the chunk if it meets the following criteria

For every $1 \leqslant u \leqslant m$ such that for every fdbu.ds=qidb.vl.d ^ fdbu.ds=wj.ds ^ wj.dw=1 for every j $1 \leqslant j \leqslant k$

Count (qidb.vl.d) $\leqslant$ Tsup

9) Disclosure - After Titr iterations, DIS(QIDB.v) adds extra records if it meets the following criteria for every $1 \leqslant l \leqslant x$ such that for every fdbu.ds=qids.vl.d ^ fdbu.ds=wi.d ^ wi.dw=1 for some i $1 \leqslant i \leqslant k$

(unique (qidb.vl .d) / x) = Tdis ± fdbv.p

B. Personalized Privacy Breach

Assume an attacker who tries to estimate important information from a record holder h. In the worst situation, the attacker only pays attention to the tuples tuples $r^* \in R^*$ whose Q value rj*.q covers x.q for all j such that $1 \leqslant j \leqslant n$ since it is assumed that the adversary knows Q of H. Q-group is formed by these tuples. That is, if rj* and rjp* are two such tuples then rj*.q=rjp*.q for all j such that $1 \leqslant j \leqslant n$. The adversary cannot deduce a sensitive attribute of h if this group is not established.

1) Required Q-Group/ Actual (h) - Given an individual h, the Required Q-group ReqG(H) is the only Q-group in r* covers h.q. Considering Actual (X) represents those records which are generalized to RG(H).

The attacker has no knowledge about Actual (H). To acquire Actual (H), the adversary must locate some external data base External (H) that should be covered in ReqG(H).

2) External DataBase Ext (x)- External (H) is a collection of individuals whose value is covered by ReqG(H).

Actual (H) ⊆ External (H)

The adversary uses a combinational strategy to deduce sensitive attribute of h. let us consider that h.s is present in one of ri* and h is not repeated. The possible reconstruction of the ReqG(X) contains h different record holders h1, h2, h3,…,hr who belong to External(H) but there can be only y in ReqG(H). This can be seen by the probabilistic nature and can be represented as perm(x,y).

perm(x,y) is Possible Reconstruction that can be created by with h holders and y mappings. Breach Probability represents the probability of inferred information. Let us consider Actual N represents actual number of entries with sensitive attribute from which h can be deduced.

Breach probability = Actual N/perm(x,y)

Breach probability will decide the privacy factors, If it is 100 percent then h can be deduced; if it is poor then the inference will be tough for the adversary.

C. Quasi-Identifier Distribution Block - Anonymization Algorithm

Since it is assumed that the sensitivity distribution in every location is typically fairly uniform, this technique processes quasi values sequentially. Consider the following algorithm of QIDB.

Algorithm 1: QIDB-Anonymization

Input: personal data R with DW-DI, threshold values T_n, T_{itr}, T_{sup}, T_{dis}, T_{acc} and initialized FDB(ds,p)

Output: Released table T *
Step 1: if ($n < T_n$) then return value 1
Step 2: for each r_j.s where r_j.dw=1 & r_j.di=0 a new QIDB is generated comprising r_j.d and r_j.q for every $1 \leq j \leq n$.
Step 3: inital_iteration=0,
receive_flag=0
gen=Initial G(R)
Step 4: while (initial_iteration< T_{itr} and receive_flag=0)
QIDB chunks are deleted if CFq() returns true then examines the value of QIDB if it is 0 then receive_flag=1
Iteration = iteration + 1
gen = next G(R)
Step 5: if receive_flag=0 then
execute sup( ) and dis( )
Step 6: Examine value of QIDB if it is 0 then receive_flag=1
Step 7: release R* if receive_flag=1

The resultant anonymization after implementing Personal Anonymization of one of the QIDB with T_{acc} =0.1 chunk is depicted in Table IX.

TABLE IX. RESULTANT DW-DI BASED QIDB ANONYMIZATION WITH T_{ACC} =0.1

ZIP Code	*Age*	*Disorder*
386**	[30-50]	Brain cancer
386**	[30-50]	Mouth ulcer
386**	[30-50]	Lungs Disease
386**	[30-50]	Fever
386**	[30-50]	Fever

IV. EXPERIMENTAL FINDINGS AND DISCUSSION

Effectiveness of proposed method in comparison to k-anonymity as well as l-diversity is obtained. The investigation made use of a common dataset. 400-records of adult dataset are taken into account with the relevant quasi-attributes: age, gender, marital status, and profession. Age is the only attribute that is numerical; all other attributes are categorical. For DW=1, probability is utilized to determine the DI value.

In Fig. 1, it is shown that data loss for proposed method is less than k-anonymity and l-diversity. Number of records can be increased in the proportion to see the information loss in three methods and compare it.

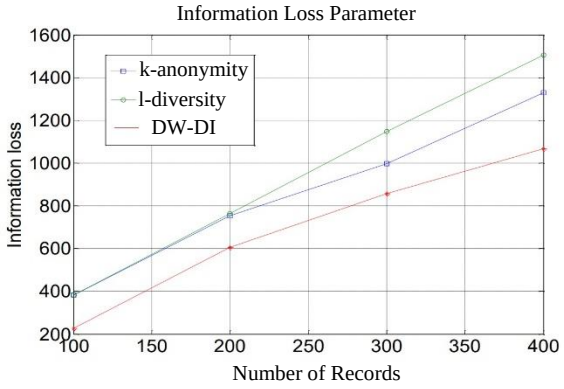

Fig. 1. Information Loss of DW-DI Proposed Personal Anonymization Technique Compared with l-Diversity and k-Anonymity Technique.

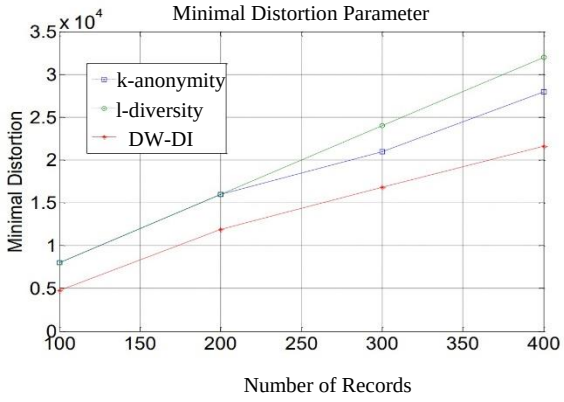

Fig. 2. Minimal Distortion Parameter of DW-DI Personal Anonymization Compared to l-Diversity k-Anonymity Technique.

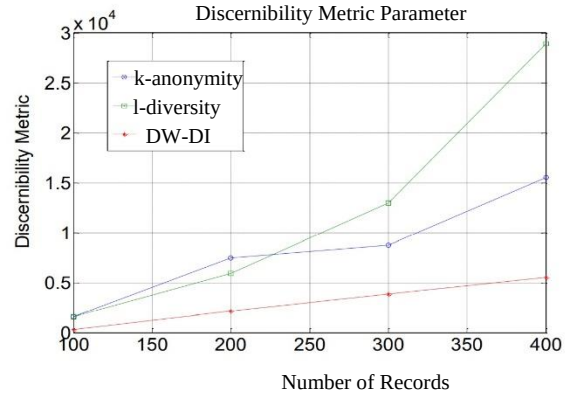

Fig. 3. Discernibility Metric Parameter of DW-DI Personal Anonymization Compare to l-Diversity and k-Anonymity Technique.

For quasi identification, a generalization hierarchy is created and employed and a distance vector is produced and is used in this approach. The generalization hierarchy can go up to a maximum level of 10. In Fig. 1, the information loss factor is displayed. The data quality improves when there is less data loss. The concept of minimal distortion centers on penalizing every value that has been generalized or repressed. When a hierarchy inside the domain generalization hierarchy is extended to the next level, it is given a penalty. In Fig. 2, minimum distortion is displayed. A penalty of 10 is applied in test for each generalization. Fig. 3 illustrates how this Discernibility Metric determines the cost by penalizing every tuple for being unrecognizable from other tuples. In Fig. 4, runtime is displayed. For the test, the threshold values Tn = 400, Titr = 10, Tdis = 0.01, Tsup = 1, Tacc = 01 was used.

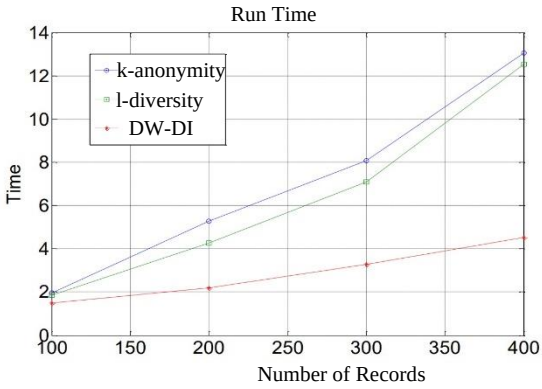

Fig. 4. Run Time of DW-DI Personal Anonymization Compare to l-Diversity and k-Anonymity Technique.

V. Conclusion and Future Work

Since the runtime and quality of the data are better with personalized privacy, it is an essential research direction. Because all entries do not need to be private, using DW not only enhances the signal of sensitivity but also increases the usefulness of the data. Since many of the record holders are willing to expose their identities, DI is an extra flag that increases the quality of the data in the DW record. Therefore, DW-DI is a better solution for personalized privacy than employing a guarding node alone. Using anonymization depending on QIDB, several quasi groups can be separately generalized. This method improves confidentiality by checking each QIDB chunk for a FD of sensitive values that is roughly equivalent to the FD of sensitive values in the original contents. Additionally, it defeats probabilistic assault, attribute connection and record connection. When a specific sensitivity's frequency distribution is localized in a small area of an individual pattern, this method performs effectively.

Future research can go in a number of different ways as it examines QIDB anonymization of DW-DI personal privacy. Firstly, the impact of sequential and multiple distributions of released data have not been taken into account. Research on sensitivity weighting can be taken into consideration. In this method, records are processed sequentially to see if the generalized record fits the QIDB generalized value, and if they do, the record is added to the block. Different techniques can be investigated as an option to sequential processing. Multi-dimensional data and unorganized schema can both be used with this technique.

References

[1] S. Kim, M. K. Sung, and Y. D. Chung, "A framework to preserve the privacy of electronic health data streams," J. Biomed. Inform., vol. 50, pp. 95–106, 2014, doi: 10.1016/j.jbi.2014.03.015.

[2] G. D. Puri and D. Haritha, "Survey big data analytics, applications and privacy concerns," Indian J. Sci. Technol., vol. 9, no. 17, 2016, doi: 10.17485/ijst/2016/v9i17/93028.

[3] G. D. Puri and D. Haritha, "Framework to avoid similarity attack in big streaming data," Int. J. Electr. Comput. Eng., vol. 8, no. 5, 2018, doi: 10.11591/ijece.v8i5.pp.2920-2925.

[4] G. D. Puri and D. Haritha, "A novel method for privacy preservation of health data stream," Int. J. Adv. Trends Comput. Sci. Eng., vol. 9, no. 4, pp. 4959–4963, 2020, doi: 10.30534/ijatcse/2020/110942020.

[5] X. Zhou, J. Liu, Q. Wu, and Z. Zhang, "Privacy Preservation for Outsourced Medical Data with Flexible Access Control," IEEE Access, vol. 6, pp. 14827–14841, 2018, doi: 10.1109/ACCESS.2018.2810243.

[6] S. Bahri, N. Zoghlami, M. Abed, and J. M. R. S. Tavares, "BIG DATA for Healthcare: A Survey," IEEE Access, vol. 7, pp. 7397–7408, 2019, doi: 10.1109/ACCESS.2018.2889180.

[7] P. Ganesh D, P. Dinesh D, and W. Manoj A., "RAID 5 Installation on Linux and Creating File System," Int. J. Comput. Appl., vol. 85, no. 5, pp. 43–46, 2014, doi: 10.5120/14841-3107.

[8] B. D. Deebak, F. Al-Turjman, M. Aloqaily, and O. Alfandi, "An authentic-based privacy preservation protocol for smart e-healthcare systems in iot," IEEE Access, vol. 7, pp. 135632–135649, 2019, doi: 10.1109/ACCESS.2019.2941575.

[9] J. Bernal Bernabe, J. L. Canovas, J. L. Hernandez-Ramos, R. Torres Moreno, and A. Skarmeta, "Privacy-Preserving Solutions for Blockchain: Review and Challenges," IEEE Access, vol. 7, pp. 164908–164940, 2019, doi: 10.1109/ACCESS.2019.2950872.

[10] J. Park, D. S. Kim, and H. Lim, "Privacy-preserving reinforcement learning using homomorphic encryption in cloud computing infrastructures," IEEE Access, vol. 8, 2020, doi: 10.1109/ACCESS.2020.3036899.

[11] X. Yang, M. Wang, X. Wang, G. Chen, and C. Wang, "Stateless Cloud Auditing Scheme for Non-Manager Dynamic Group Data with Privacy Preservation," IEEE Access, vol. 8, pp. 212888–212903, 2020, doi: 10.1109/ACCESS.2020.3039981.

[12] M. Keshk, B. Turnbull, E. Sitnikova, D. Vatsalan, and N. Moustafa, "Privacy-Preserving Schemes for Safeguarding Heterogeneous Data Sources in Cyber-Physical Systems," IEEE Access, vol. 9, pp. 55077–55097, 2021, doi: 10.1109/ACCESS.2021.3069737.

[13] A. Al Omar et al., "A Transparent and Privacy-Preserving Healthcare Platform with Novel Smart Contract for Smart Cities," IEEE Access, vol. 9, pp. 90738–90749, 2021, doi: 10.1109/ACCESS.2021.3089601.

[14] M. D. N. Ayeh. and A. M. Kissi Mireku Kingsford, Fengli Zhang, "A Mathematical Model for a Hybrid System Framework for Privacy Preservation of Patient Health Records," in 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), 2017, pp. 119–124, doi: 10.1109/COMPSAC.2017.21.

[15] S. Yao. and Z. Huo, T. Wang, Z. Zheng, M. H. Rehmani, "Privacy Preservation in Big Data From the Communication Perspective—A Survey," IEEE Commun. Surv. Tutorials, vol. 21, no. 1, pp. 753–778, 2019, doi: 10.1109/COMST.2018.2865107.

[16] M. U. Hassan, M. H. Rehmani, and J. Chen, "Privacy preservation in blockchain based IoT systems: Integration issues, prospects, challenges, and future research directions," Futur. Gener. Comput. Syst., vol. 97, 2019, doi: 10.1016/j.future.2019.02.060.

[17] J. Liu. and M. Wang. Kefei Mao, Jie Chen, "Security enhancement on an authentication scheme for privacy preservation in Ubiquitous Healthcare System," in 2015 4th International Conference on Computer Science and Network Technology (ICCSNT), 2015, pp. 885–892, doi: 10.1109/ICCSNT.2015.7490882.

[18] K. Xu, H. Yue, L. Guo, Y. Guo, and Y. Fang, "Privacy-Preserving Machine Learning Algorithms for Big Data Systems," Proc. - Int. Conf. Distrib. Comput. Syst., vol. 2015-July, pp. 318–327, 2015, doi: 10.1109/ICDCS.2015.40.

[19] T. Karle and D. Vora, "Privacy preservation in big data using anonymization techniques," 2017 Int. Conf. Data Manag. Anal. Innov. ICDMAI 2017, pp. 340–343, 2017, doi: 10.1109/ICDMAI.2017.8073538.

[20] B. K. Bodkhe, "Privacy Preservation for Medical Dataset using Hadoop," in 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), 2017, pp. 3463–3468.

[21] S. Sathya and T. Sethukarasi, "Efficient privacy preservation technique for healthcare records using big data," 2016 Int. Conf. Inf. Commun. Embed. Syst. ICICES 2016, no. Icices, 2016, doi: 10.1109/ICICES.2016.7518878.

[22] S. Shimona, "Survey on Privacy Preservation Technique," Proc. 5th Int. Conf. Inven. Comput. Technol. ICICT 2020, pp. 64–68, 2020, doi: 10.1109/ICICT48043.2020.9112584.

[23] P. S. V, "Privacy Preservation of Healthcare Big Data," no. Icimia, pp. 743–749, 2020.

[24] H. Jiang, "Research and practice of big data analysis process based on Hadoop framework," Proc. 2019 IEEE 3rd Inf. Technol. Networking, Electron. Autom. Control Conf. ITNEC 2019, no. Itnec, pp. 2044–2047, 2019, doi: 10.1109/ITNEC.2019.8729522.

[25] H. Liu, X. Yao, T. Yang, and H. Ning, "Cooperative privacy preservation for wearable devices in hybrid computing-based smart health," IEEE Internet Things J., vol. 6, no. 2, pp. 1352–1362, 2019, doi: 10.1109/JIOT.2018.2843561.

[26] M. Zhang, Y. Chen, and W. Susilo, "PPO-CPQ: A Privacy-Preserving Optimization of Clinical Pathway Query for E-Healthcare Systems," IEEE Internet Things J., vol. 7, no. 10, pp. 10660–10672, 2020, doi: 10.1109/JIOT.2020.3007518.

[27] A. Alabdulatif, I. Khalil, A. R. M. Forkan, and M. Atiquzzaman, "Real-Time Secure Health Surveillance for Smarter Health Communities," IEEE Commun. Mag., vol. 57, no. 1, pp. 122–129, 2019, doi: 10.1109/MCOM.2017.1700547.

[28] S. Dolev, P. Florissi, E. Gudes, S. Sharma, and I. Singer, "A Survey on Geographically Distributed Big-Data Processing Using MapReduce," IEEE Trans. Big Data, vol. 5, no. 1, pp. 60–80, 2017, doi: 10.1109/tbdata.2017.2723473.

[29] M. Du, K. Wang, Z. Xia, and Y. Zhang, "Differential Privacy Preserving of Training Model in Wireless Big Data with Edge Computing," IEEE Trans. Big Data, vol. 6, no. 2, pp. 283–295, 2018, doi: 10.1109/tbdata.2018.2829886.

[30] X. Zhang et al., "Proximity-aware local-recoding anonymization with MapReduce for scalable big data privacy preservation in cloud," IEEE Trans. Comput., vol. 64, no. 8, pp. 2293–2307, 2015, doi: 10.1109/TC.2014.2360516.

[31] H. Huang, T. Gong, N. Ye, R. Wang, and Y. Dou, "Private and Secured Medical Data Transmission and Analysis for Wireless Sensing Healthcare System," IEEE Trans. Ind. Informatics, vol. 13, no. 3, pp. 1227–1237, 2017, doi: 10.1109/TII.2017.2687618.

[32] M. Keshk, B. Turnbull, N. Moustafa, D. Vatsalan, and K. K. R. Choo, "A Privacy-Preserving-Framework-Based Blockchain and Deep Learning for Protecting Smart Power Networks," IEEE Trans. Ind. Informatics, vol. 16, no. 8, pp. 5110–5118, 2020, doi: 10.1109/TII.2019.2957140.

[33] J. He, L. Cai, and X. Guan, "Preserving data-privacy with added noises: Optimal estimation and privacy analysis," IEEE Trans. Inf. Theory, vol. 64, no. 8, pp. 5677–5690, 2018, doi: 10.1109/TIT.2018.2842221.

[34] W. Gao, W. Yu, F. Liang, W. G. Hatcher, and C. Lu, "Privacy-Preserving Auction for Big Data Trading Using Homomorphic Encryption," IEEE Trans. Netw. Sci. Eng., vol. 7, no. 2, pp. 776–791, 2020, doi: 10.1109/TNSE.2018.2846736.

[35] M. Keshk, E. Sitnikova, N. Moustafa, J. Hu, and I. Khalil, "An Integrated Framework for Privacy-Preserving Based Anomaly Detection for Cyber-Physical Systems," IEEE Trans. Sustain. Comput., vol. 6, no. 1, pp. 66–79, 2019, doi: 10.1109/tsusc.2019.2906657.