Evaluation of Re-identification Risk using Anonymization and Differential Privacy in Healthcare

Ritu Ratra, Preeti Gulia, Nasib Singh Gill
Department of Computer Science and Applications
Maharshi Dayanand University
Rohtak, Haryana, India

***Abstract*—In the present scenario, due to regulations of data privacy, sharing of data with other organization for research or any medical purpose becomes a big hindrance for different healthcare organizations. To preserve the privacy of patients seems like a crucial challenge for Healthcare Centre. Numerous techniques are used to preserve the privacy such as perturbation, anonymization, cryptography, etc. Anonymization is well known practical solution of this problem. A number of anonymization methods have been proposed by researchers. In this paper, an improved approach is proposed which is based on k-anonymity and differential privacy approaches. The purpose of proposed approach is to prevent the dataset from re-identification risk more effectively from linking attacks using generalization and suppression techniques.**

Keywords*—Data privacy; anonymization; differential privacy; re-identification risk analysis; privacy preserving data publishing

I. INTRODUCTION

Due to the advancements in the areas of business intelligence, generally organizations for instance banks, healthcare, health insurance are converted into "data-driven" organizations. These organizations used to apply new mechanisms to analyze a high volume of data. It is the responsibility of the data controller to ensure the user about their privacy and it should be done before publishing the data to a third party. There is no protection of privacy in the original dataset. PPDP (Privacy-Preserving Data Publishing) offered numerous tools and mechanisms to preserve privacy. [1][2][3][4]. Anonymization must be done on the datasets before publishing to various organizations because they may contain personal information. It is well known that personal information can be gathered from these types of records and there are many people who assess the re-identification risk. European Medicines Agency (EMA) recommends an anonymization approach for risk analysis based on qualitative technique and quantitative technique [5].

PPDP process consists of different phases i.e. collection of data; providing storage for collected data; perform anonymization; data publishing after modification and perform data mining process as shown in the conceptual scenario of PPDP described in Fig. 1. There are some persons such as record owner, data holder; data publisher; data recipient, and adversary are involved in this process. The record owner is the entity of record, data holder can be person or organization that holds the data; data publisher is responsible for the publishing of anonymous data; data recipient is any entity that has access to published data and adversary is the entity whose objective is to gather user's information. At the time of the data publishing process, sensitive records may be leaked out. To overcome this problem one possible solution is to modify the dataset. There are many methods for modification of datasets in PPDP [6]. Data anonymizaton is most commonly used to achieve privacy protection in data publishing. Several methods have been proposed to handle the security issues related to datasets. In particular, anonymisation and differential privacy are two techniques that have been used for implementation practically. The k-anonymity used to perturb datasets by generalization and suppression. K-anonymity algorithm is used to preserve user's identity through linking attacks [7]. Differential privacy is also used to prevent privacy by furnishing individuals' personal information ability. However, instead of using k-anonymity's deterministic approach to in distinguishability, differential privacy invokes stochastic in-distinguishability by adding noise or perturbing values. Both k-anonymity and Ɛ-differential privacy suffer from a number of drawbacks. In particular, the curse of dimensionality of adding extra quasi identifiers to the k-anonymity framework results in greater information loss [8]. On the other hand, differential privacy has long been criticized for the large information loss imposed on records. The proposed technique in this paper shows how to overcome these drawbacks by combining k-anonymity and Ɛ-differential privacy, while simultaneously benefitting from their advantages. This paper presents the k-anonymity and differential privacy technique. Both techniques have their own limitations. This can be improved upon in their combination. To implement such a concern is focus of their paper is on re-identification risk analysis.

The rest of the organization of the paper is as follows: Section II provides the literature survey related to anonymization and differential privacy. Section III elaborates the materials and methods used in the paper. Section IV describes the proposed work. Section V presents the experimental details of proposed technique and corresponding results. Section VI concludes the paper.

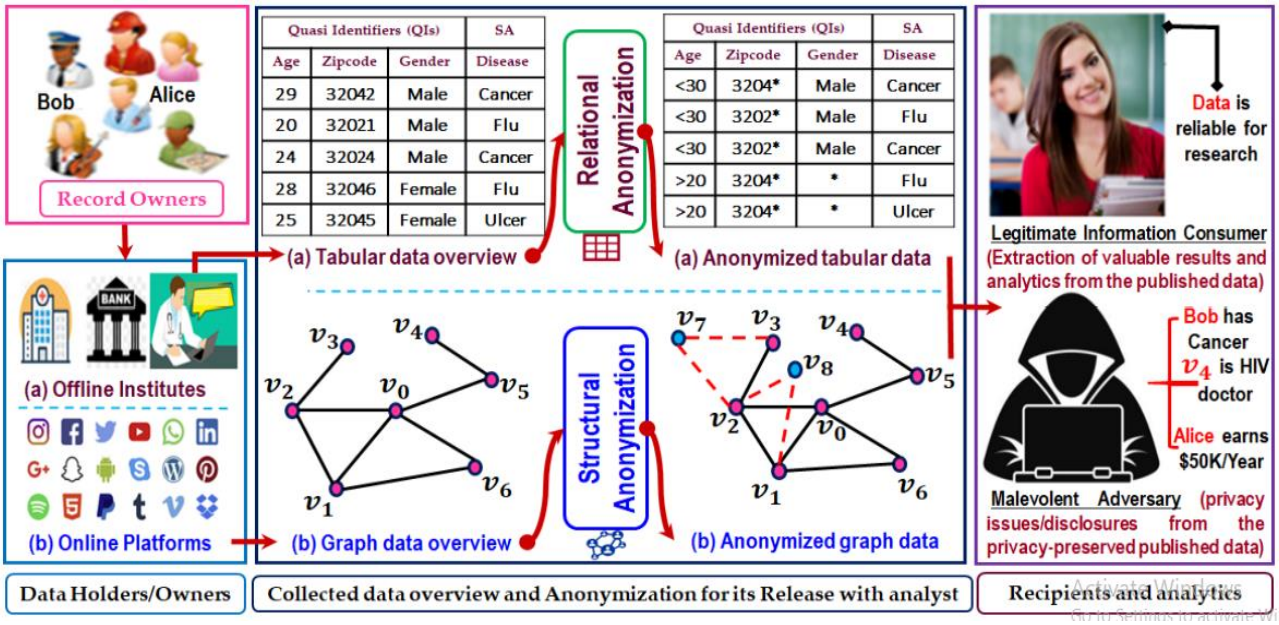


Fig. 1. Privacy Preserving Data Publishing (PPDP) Process [6].

II. Related Work

Protection of sensitive data and extraction of useful information from distributed data is also a challenging task. It is need to preserve the privacy of the before publishing. More than sufficient work has been proposed and implemented in the field of privacy-preserving data publishing. There are several methods used to protect sensitive data. There are various privacy-enhanced mechanisms that are related to the preservation of privacy [9].

Luc Rocher et.al [10] proposed an approach based on the generative copula method. This approach estimated more accurately the probability of anyone to be rightly re-identified.

Boris Lubarsky [11] described a method that proved to be successful even in the heavily incomplete dataset shared. Re-identification can occur due to insufficient anonymization of datasets or combining the datasets. Pseudonym reversal may also be one of the causes of re-identification risk.

Branson, et al [12] have presented a study of testing the re-identification problem. They presented a study through the testing of how a prescribed drug can be subjected to cause of re-identification.

Suman et al [13], introduced a novel technique based on anonymization. The proposed algorithm's performance was measured by using information loss and accuracy. In various experiments, proposed approach provided minimum information loss and maximum accuracy.

Sumana and Hareesh [14] described various anonymization methods in PPDM which are used to provide privacy of the data. Anonymization's main goal is to secure access to personal information and is also used to provide accumulated information.

Vibhor Sharma et.al [15] presented a new Evolutionary privacy-preserving technique in data mining. Whenever data mining is applied to large datasets a number of threats are automatically introduced to privacy. To provide protection to the sensitive data of individuals, data should be masked before it is revealed for data mining.

Marques et.al [16] discussed a complete analysis study on anonymization. A number of techniques of anonymization can be applied to datasets to prevent re-identification risk. They discussed different tools such as ARX, µ-Argus, SDC Micro, and Privacy Analytics Eclipse.

Manoj Kumar Gupta et.al [17] determined various approaches like a generalization, k-anonymity, l-diversity, suppression, shuffling, noise addition, etc. l-diversity is based on the inside group diversity of sensitive attributes. According to the definition of l-diversity, there must be minimum value for each private attribute when each group contains one sharing combination of key attributes. Only then the dataset will be considered as satisfied l-diverse.

P Ram Mohan Rao et.al [18] introduced a novel approach named "Synthesize Quasi Identifiers and apply Differential Privacy" (SQIDP) for privacy-preserving in data mining. This approach was applicable to text data set with 100% data utility.

III. Methods and Techniques Existing

This section highlights the existing techniques and algorithms that are used in proposed technique i.e. Anonymization and differential privacy. These techniques are used to preserve the privacy before publishing.

A. Anonymization

Anonymization is a type of modification technique used to preserve privacy [19]. In data anonymization, sensitive information is either encrypted or removed from the datasets in order to preserve the privacy. There are two methods of anonymization i.e. generalization and suppression [20]. In

Generalization, individual attributes are substituted with an extensive category. Generalization is also a method used for changing categorical attributes and continuous numeric attributes, while suppression means just removing the values of attributes. In this, certain values of the attributes are converted into an asterisk '*'. Various types of attributes are as [21].

Although these types of information may seem very harmless and individually may not present any harm but by linking them from each other, the attackers can misuse can also change the information. In order to hide these original data, there is need to hide and secure these data which may, in turn, present us with another challenge, information loss.

Nowadays, it is common that some of the datasets are openly available for research purpose. To preserve the privacy of shared data, the owner of data can apply different types of anonymization on the datasets. Generalization, suppression, permutation, and perturbation are some examples of anonymization. Furthermore, more than one approach can be applied to the dataset. It proved more beneficial to protect the privacy of data [22]. Therefore, it is necessary to consider the concept of de-identification and re-identification of data. For this purpose, a medical data set has been used that contains the information of some patients. It is depicted in Table I. Here the name attribute is the personal identification attribute; a sensitive attribute is a disease.

TABLE I. DESCRIPTION OF USER'S ATTRIBUTES AND SOLUTION IN ANONYMIZATION

Attribute type	Meaning of Attribute	Solution in Anonymization
Identifying/ Direct	Some attributes like name, mail identity, or aadhar number come under this category. These attributes can certainly recognize the person' personal information	These attributes are removed in anonymization process.
Quasi identify	When one attribute linked with some other attribute caused the disclosure of privacy then those are called quasi identify attributes. For example, age and sex when linked to some other database can easily disclose the person's identity	These attributes are suppressed or generalized in order to preserve the privacy of an individual.
Sensitive	These attributes are crucial and should not be shared. For example. Disease information, salary information should not be shared against any organization.	Mostly do not change for data analyses.
Non-Sensitive	These are the attributes that are publishable publicly because these do not create any problem related to privacy. For example weight, hair color, height, etc.	These are not collected in most cases. If collected, shared as it is.

TABLE II. MICRO TABLE OF HEALTHCARE RECORDS (ORIGINAL)

Name	Zip	Age	Gender	Disease
Wilson	56478	25	M	Heart Disease
Marin	56399	27	F	Blood Cancer
Bob	56789	43	M	Flu Holdon
Emela	56866	34	F	Heart Disease
Peter	56300	24	M	Heart Disease
John	56708	46	M	Prostate Cancer
Boby	56427	33	M	Prostate Cancer

Table II is an example of de-identification. De-identification is the process of altering the dataset to create an alternate use of the dataset so that it is impossible to recognize the identity. De-identification of Table II is shown in Table III, where the field name "Name" is deleted. To provide privacy if the name attribute is removed, then to provide the privacy data can be altered and the altered data is displayed in Table III.

Now the names of patients are not shown in Table III. However, if anyone has access to Aadhar Card Data (as shown in Table IV), it is very easy to discover the information regarding all records. It can be done by joining the two different tables on the common attributes.

TABLE III. HEALTHCARE RECORDS AFTER DELETING THE NAME FIELD (DE-IDENTIFICATION)

Zip code	Age	Gender	Disease
56478	25	M	Heart Disease
56399	27	F	Blood Cancer
56789	43	M	Flu Holdon
56866	34	F	Heart Disease
56300	24	M	Heart Disease
56708	46	M	Prostate Cancer
56427	33	M	Prostate Cancer

These common attributes are known as quasi-identifier. By using the data of Table III and Table IV, an attacker can easily get the information that it is Bob is suffering from a disease of Flu Holdon. So removing the personal information will not be helpful for complete privacy to the data. The method of reversing the de-identification by connecting the identity of the data subject is referred to as Re-identification.

TABLE IV. AADHAR CARD DATASET MICRO DATASET

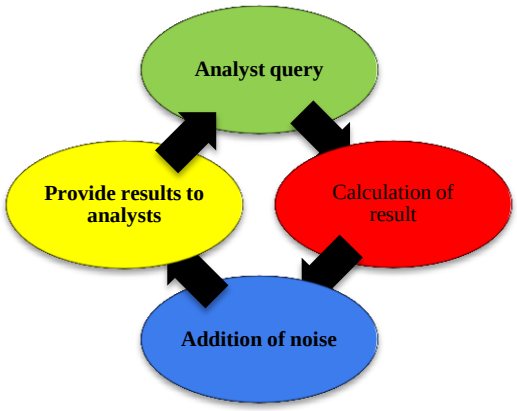
Proposed Name	Zip	Age	Gender
Wilson	56478	25	M
Marin	56399	27	F
Bob	56789	43	M
Emela	56866	34	F
Peter	56300	24	M
John	56708	46	M
Boby	56427	33	M

So in short it can be said that deletion of the personal identification data from relation will not much helpful to protect privacy [23]. To protect privacy first of all personal identification data must be removed and anonymization of the quasi-identifiers is also required.

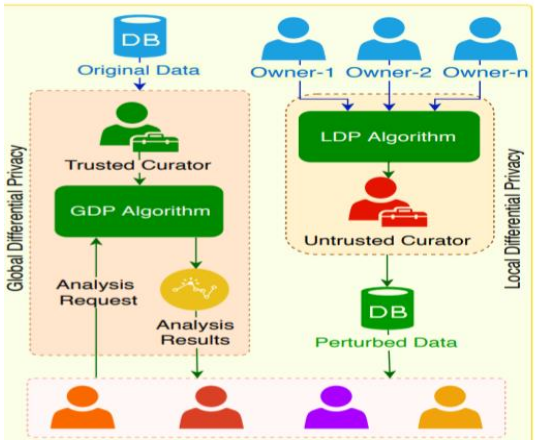
B. Differential Privacy

Differential privacy is also a widely used privacy preservation method. This approach permits the analysts to explore necessary answers from the data repositories that contain sensitive information [24]. In this method, analysts are able to get answers from data stores having sensitive data with secure protection of privacy [25]. In differential privacy, a randomized function R provides Ɛ-differential privacy protection for all data sets named DS1 and DS2. These datasets are differing on at most one data element [26]. This randomized function is such that:

$$Pr [R (DS1) \in S] \leq \exp(Ɛ) \times Pr[\kappa(DS2) \in S]$$



2(a)



2(b)

Fig. 2. (a) Process of Differential Privacy, (b) Global and Local Differential Privacy [16].

Ɛ is the statistical distance, it is use to define the strength of privacy. A lower value of Ɛ means stronger privacy [27]. Different steps of the differential privacy approach are shown in Fig. 2(a). Fig. 2(b) describes the GDP (Global Differential Privacy) and LDP (Local Differential Privacy). A trusted curator recruited in GDP. He can apply gauged noise in order to produce DP (Differential Privacy). The curator should make some practical algorithms or mechanisms that are inappropriate for deep learning. Here the algorithm resides on the server and the original data set has to be uploaded onto the server for training. But in the case of LDP, owners of data modify the data before publishing. There is no need for a trusted curator or any third party to preserve privacy. LDP guaranteed better privacy as compared to GDP. It should be noted that data values are not changed in DP. Here, Users cannot access the database directly. These inaccurate data are sufficient to protect privacy but so small that helpful for the analysts and researchers. Privacy and Utility are not mutually exclusive [28].

IV. Proposed Technique

This paper presents an enhanced privacy –preserving approach based on anonymization and differential techniques. It helps to hide information without abruptly changing the records. The records are k-anonymized as there are k data sets with the same value in each quasi field. To provide anonymization to the original dataset generalization is used. This method is always applied to the quasi attributes [29]. Suppression and generalization techniques are used to provide anonymization. The suppression method is used on quasi attributes in the format of same size intervals. It is done for uniformity in the data set. The proposed enhanced approach tends to solve the privacy issue related to various attacks Generalization is the process through which data can be presented in the form of clustering. The elementary objective of this technique used to collect the links into the cluster and then make a super vertex. Every vertex provides the merged information of the super network. Using this approach, identifying the local data or information is very difficult. To provide protection from re-identification risk, different PPDM (Privacy Preserving Data Mining) techniques [30] are used but the method of anonymity is widely used. This paper proposed the technique of k-anonymity and Ɛ-differential privacy. The proposed method anonymized the data set using a k-anonymity algorithm with k=2 and k=5. The very step first step is to classify the features into sensitive, quasi, and identifiers features. After this, the quasi-identifiers are partitioned into k- quasi on which k-anonymity is applied, and on k- quasi, Ɛ-differential privacy is applied. After this, k-quasi attributes are processed to provide the k-anonymity. After this in the next step differential privacy is applied to the k-quasi attributes. The inspiration to take differential privacy is its stochastic in-distinguishability. Now k-anonymity has applied, an attacker can uniquely recognize the equivalence class. In which any individual's record belongs to that k-quasi. With the help of Ɛ-quasi, it is ensured that the re-identification of records cannot occur.

The proposed method is shown with the help of a flowchart in Fig. 3. It preserves from re-identification risk between equivalence classes. In differential privacy, every equivalent class is considered as a single independent class of an individual's record. In this concept, it is more important to know that differential privacy equivalence class is not the set of attributes. To prevent from re-identification risk records are shuffled.

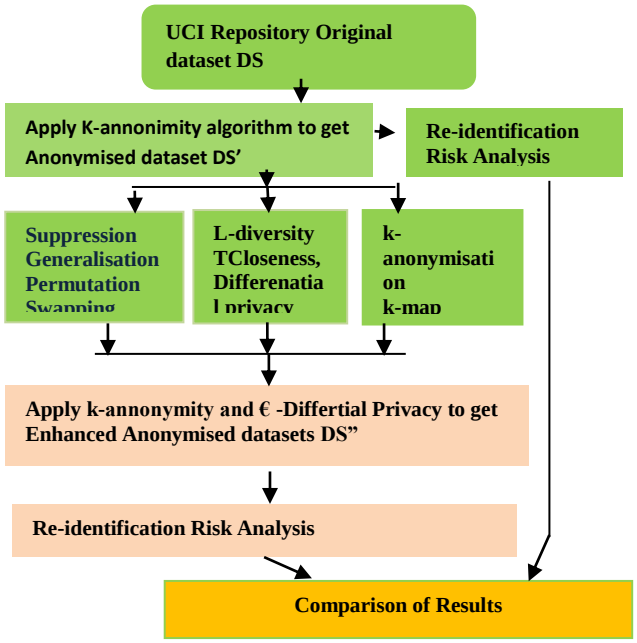


Fig. 3. Flow Chart of Proposed Research Method.

The proposed work is described in Algorithm 1.

Algorithm1: k-ADP (k- Annonymity Differential Privacy)

Input: Original Data set DS

Output: Anonymized data set using k-ADP

Step1. Classify the features (attributes) into quasi, identifiers and sensitive

Step2. Set k-quasi attributes → k-quasi

Step3. Set Ɛ –quasi attribute → Ɛ- quasi

Step4. Apply k-anonymization on k-quasi attributes.

Step5. Apply k-ADP (k- Anonymity Differential Privacy) technique to each equivalence class of k- annonymised dataset

Step6. Now merge k-anonymised records and Ɛ- Differential Privacy records.

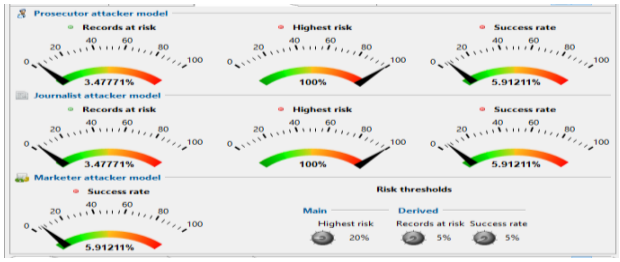
V. Experiments and Results

There are numerous tools and mechanisms for privacy-preservation of datasets. In this paper, anonymization and differential privacy methods are used to provide protection from re-identification risk. From the UCI machine learning repository, Heart dataset is selected for analysis purposes. There are 14 attributes in heart dataset and 2602 records. Out of all attributes, only quasi attributes and sensitive attributes are considered. Here two attributes names as 'age' and 'sex' are considered as quasi attributes and class names as 'result' is considered as a sensitive attribute. Users can directly apply the anonymization method to datasets by using the ARX tool. This tool accepts the files of .csv, .xls, and .xlsv format. Here, k-anonymity with k=2, k=10, and generalization method is selected to perform anonymization on the dataset. Differential privacy is applied to the anonymized dataset. The proposed technique is used to evaate the risk factor of the re-identification. For this purpose, the relationship between k and Ɛ is evaluated. As increases the value of k, the risk is decreased and the risk is decreased with decreasing the value of Ɛ. Now, re-identification risk analysis is done on three datasets i.e. original dataset, anonymised dataset, and enhances anonymised dataset. Experimental results are shown using a tabular and graphical format.

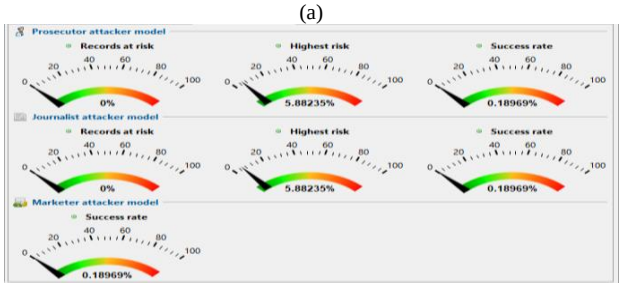
A. Effect on Re-identification Risk

Risk related to privacy can be analyzed using ARX tool [31]. These risks are related to re-identification risk for the prosecutor, journalists and markets attacker. The risk that can be derived from population uniqueness is also included. The impact of data anonymization on the re-identification risk profile for the Heart disease dataset is shown in Fig. 4 and Fig. 5.

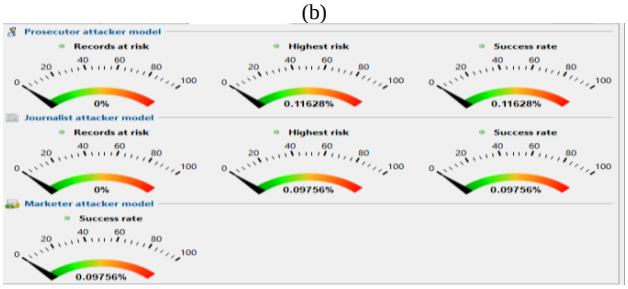
Fig. 4(a) highlights risk of re-identification risk of original dataset at Prosecutor level. Here approximately 3.47% of the total number of records is at risk. The higher risk calculated here is 100%. It means at most all records are at risk in the original dataset. The Success rate is 5.912% in the case of original dataset. At the journalist level, higher risk calculated here is 100%. It means at most all records are at risk in the original dataset. The Success rate is 5.912% in the case of original dataset. It is the same as in the case of the Prosecutor scenario. Fig. 4(b) shows the risk of re-identification of annonymised datasets at the prosecutor level. The highest risk, in this case is 5. 08%. And the effect of the proposed technique is displayed in Fig. 4(c). Here in this case data is purely safe i.e. rate of records at risk is 0% in all scenarios.



(a)



(b)



(c)

Fig. 4. (a) Risk Estimation (Original Dataset), (b) Risk Estimation (Anonymized Dataset), (c) Risk Estimation (Enhanced Anonymized Dataset).

Comparative study of the risk of various attackers of the original dataset, anonymized data set, and enhanced anonymized dataset is given in Table V. Table V lists the risk estimation evaluated at prosecutor level, journalist level, and marketer level. It is depicted that the estimated risk for journalists is higher in the original data set i.e. 33.3% and is lower in enhanced anonymized data set i.e. 0.11%. It can also be noted that estimated Marketer and the Journalist risk are also lowest in enhanced anonymized data set and higher in the original dataset. The detail of various risks is also listed in the Table V. Through the experiments, it is proved that enhanced anonymized data is safer as compared to original data and anonymized data shown in the Fig. 4. The re-identification risk of the original dataset and anonymized dataset is described in Fig. 5 and Fig. 6, respectively.

From Table V, it is stated that the highest Prosecutor risk is higher in the original dataset (100%), and less in enhanced anonymized datasets i.e. 0.11%. Estimated Journalist risk is higher in original dataset (33.30%) and lowers in enhanced anonymized dataset i.e. 0.11%. Estimated marketer risk is higher in original dataset (7.12%), and very less in enhanced anonymised datasets i.e. 0.09%. Re-identification risk estimated in various approaches to the number of records is shown in the following figures.

TABLE V. COMPARISON OF RISK ESTIMATION

Measure	Original dataset	Anonymized dataset	Enhanced Anonymized dataset
Lowest Prosecutor risk	2.12%	0.14%	0.11%
Record at lower risk	4.5%	69.56%	100%
Average Prosecutor risk	7.12%	0.18%	0.11%
Highest Prosecutor risk	100%	.32%	0.11%
Estimated Journalist risk	33.3%	0.32%	0.11%
Estimated Marketer risk	7.12%	0.19%	0.09%.
Estimated Prosecutor risk	33.33%	0.32%	0.11%

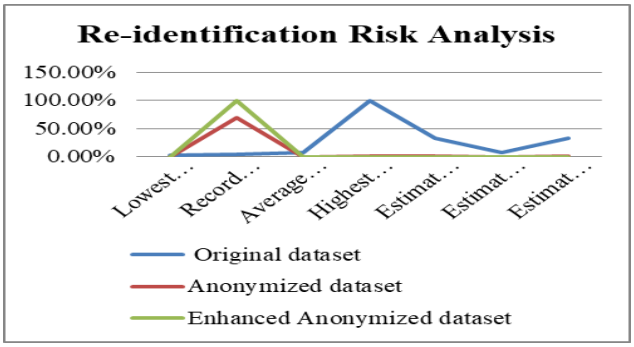
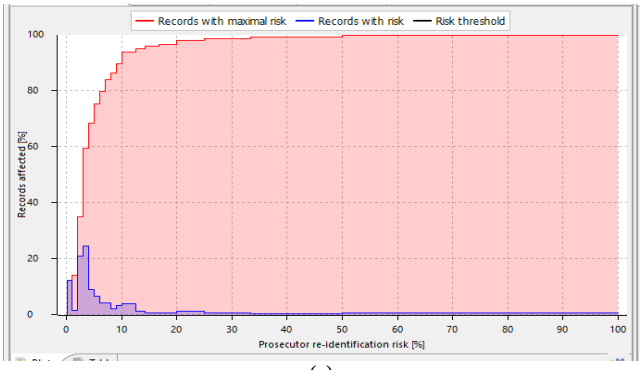
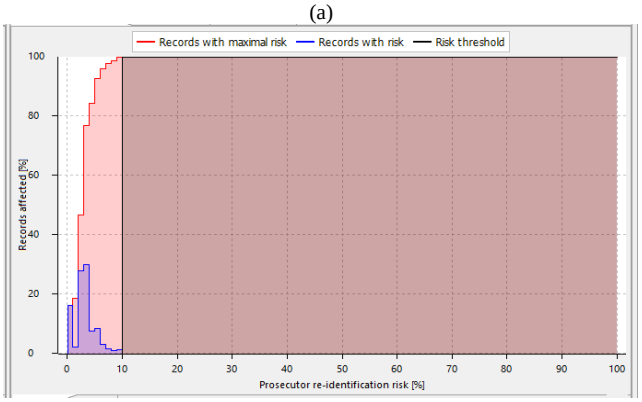


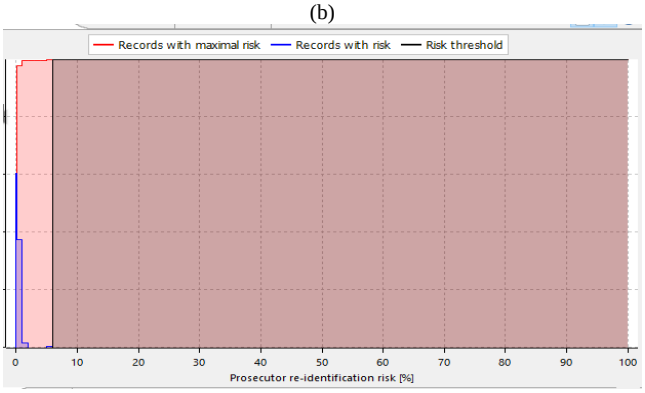
Fig. 5. Re-identification Risk Analysis.



(a)



(b)



(c)

Fig. 6. (a) Re-identification Risk (Original Dataset), (b) Re-identification Risk (Anonymized Dataset), (c) Re-identification Risk (Enhanced Anonymized Dataset).

In the above figures, re-identification risk distribution among the dataset's records is displayed. The calculation of distribution depicted on the input dataset and output dataset. Fig. 6(a) highlights the records with Maximum risk, records of with risk, and risk threshold of the data to prosecutor re-identification risk in percentage. Fig. 6(b) depicted the Maximum risk, Record with risk, and the Risk Threshold of the anonymized dataset at Prosecutor re-identification, and in Fig. 6(c), it is shown that when anonymization with differential privacy is applied on original data set, all three estimations approaches to zero so, the proposed method is much efficient to minimize the re-identification risk.

VI. Conclusion and Future Work

In the era of data sharing, protection of privacy has become an important matter in different organization and in a healthcare industry it is directly concerned with patients. This paper proposed an enhanced anonymized approach to preserve the privacy of patients' data. To preserve the privacy, a proposed technique has been implemented on the dataset related to the heart disease. In this paper, anonymization (K-anonymity) and differential privacy approaches are used to provide privacy to the dataset. Through various experimental results, it is proved that an anonymized dataset achieved more security. The re-identification risk in a modified dataset is very much less as compared to the original dataset. In future, different classification algorithms would be applied to the anonymized dataset to measure the accuracy, execution time, kappa-static, etc.

Acknowledgment

The authors are grateful to the UCI Repository for providing the dataset and also thankful to all members of the Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak, India for their kind support.

References

[1] Deepak Narula, Pardeep Kumar, Shuchita Upadhyaya, "Evaluation of proposed amalgamated anonymization approach", Indonesian Journal of Electrical Engineering and Computer Science, Vol. 16, No. 3, pp 1439-1446, ISSN: 2502-4752, December (2019). http://doi.org/10.11591/ijeecs.v16.i3.

[2] Gregory E. Simon, et.al, "Assessing and Minimizing Re-identification Risk in Research Data Derived from Health Care Records", The Journal for electronic Health Data and Methods, EGEMS (Wash DC). 29;7(1):6, 2019. Doi 10.5334/egems.27.

[3] P Raje ndra Prasada, Tryambak Hirwarkarb, "Efficient Model for Privacy Preserving Classification Of Data Streams". Turkish Journal of Computer and Mathematics Education Vol.12 No.2, pp. 1475 -1481, (2021).

[4] Ritu Ratra, Preeti Gulia, "Privacy Preserving Data Mining: Techniques and Algorithms", International Journal of Engineering Trends and Technology, Volume 68 Issue 11, ISSN: 2231 – 5381, pp. 56-62, (2020). DOI:10.14445/22315381/IJETT-V68I11P207.

[5] Anastasiia Pika, et.al, "Privacy-Preserving Process Mining in Healthcare", International Journal of Environmental Research and Public Health", ISSN: 1660-4601, pp 1-28, (2020); https//:doi:10.3390/ijerph17051612.

[6] Abdul Majeed, Sungchang le et.al., "Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey" , IEEE Access, volume 9, pp 8512- 8545, (2021). https://doi.org/10.1109/ACCESS.2020.3045700.

[7] Can Eyupoglu, Muhammed Ali Aydin, Abdul Halim Zaim and AhmetSertbas "An Efficient Big Data Anonymization Algorithm Based on Chaos and Perturbation Techniques". www.mdpi.com/journal/entropy, pp 1-18, (2018).

[8] Pathum Chamikara Mahawaga Arachchige, Peter Bertok, Ibrahim Khalil, Dongxi Liu, Seyit Camtepe, and Mohammed Atiquzzaman, "Local Differential Privacy for Deep Learning", IEEE Internet of Things Journal, Vol. xx, no. xx, arXiv:1908.02997v3[cs.LG] 9 Nov 2019. https://doi.org/10.1109/JIOT.2019.2952146.

[9] Kunwar Singh kushwah and Abhay Panwar, "A Privacy Preservation Technique Using Machine Learning Technique", International Journal of Engineering and Innovative Technology (IJEIT). pp 3445-3454, (2015).

[10] Luc Rocher, Julien M. Hendrickx and Yves-Alexandre de Montjoye., "Estimating the success of re-identifications in incomplete datasets using generative models". Nature Communications, pp 1-9, (2019). https://doi.org/10.1038/s41467-019-10933-3.

[11] Boris Lubarsky, "Re-identification of "Anonymized Dat Georgetown Law Technology Review, Vol 1:1, pp 202-213. (2018).

[12] Branson et al. , "Evaluating the re-identification risk of a clinical study report anonymized under EMA Policy 0070 and Health Canada Regulations", pp.1-9, (2020). https://doi.org/10.1186/s13063-020-4120-y.

[13] Suman Madan and Puneet Goswami, "Adaptive Privacy Preservation Approach for Big Data Publishing in Cloud using k-anonymization", Recent Advances in Computer Science and Communications. Volume 14, Issue 8, pp 2680-2690, (2021). https://doi.org/10.2174/2666255813999200630114256.

[14] Surname, et al., "Information theoretic-based privacy risk evaluation for data anonymization", Journal of Surveillance, Security and Safety, 2021:2:83-102, (2021). https://doi,org/10.20517/jsss.2020.20.

[15] Vibhor Sharma, Dheresh Soni, Deepak Srivastava and Dr. Pramod Kumar.., "A Novel Hybrid Approach of Suppression and Randomization for Privacy Preserving Data Mining", EEO. 2021; 20(5): pp 2451-2457. (2021). doi:10.17051/ilkonline.2021.05.267. https://doi.org/10.1177/2378023121994014.

[16] Marques, J. and Bernardino, J., "Analysis of Data Anonymization Techniques.", In Proceedings of the 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2020) - Volume 2: KEOD, pages 235-241 ISBN: 978-989-758-474-9, (2020). https://doi.org/10.5220/0010142302350241.

[17] Manoj Kumar Gupta and Abhishek Gupta, "A hybrid-security model for privacy-enhanced distributed data mining", Journal of King Saud University – Computer and Information Sciences xxx (xxxx) xxx, pp 1-13, (2020).

[18] P Ram Mohan Rao, S Murali Krishna and A P Siva Kumar, "Novel algorithm for efficient privacy preservation in data analytics". Indian Journal of Science and Technology, ISSN Print: 0974-6846 Electronic: 0974-5645. pp. 519-526, (2021). https://doi.org/10.17485/IJST/v14i6.1773.

[19] S Kumaraswamy , Manjula S H , K R Venugopal, "Secure Cloud based Privacy Preserving Data Mining Platform", Indonesian Journal of Electrical Engineering and Computer Science Vol. 7, No. 3, pp830,-838, September 2017. https://doi.org/10.11591/ijeecs.v7.i3.

[20] Alpa Shah and Ravi Gulati, "Privacy Preserving Data Mining: Techniques, Classification and Implications - A Survey", International Journal of Computer Applications (0975 – 8887) Volume 137,No.12, pp 40-46, (2016).

[21] D. Kavitha, "A Survey on Privacy Preserving Data Mining Techniques". International Journal of Computer & Mathematical Sciences IJCMS ISSN 2347 – 8527 Vol. 7, Issue 2. pp. 160-169, (2018).

[22] Desmond Ko Khang Siang, et.al, "Comparative Study on Perturbation Techniques in Privacy Preserving Data Mining". International Journal of Innovative Computing 8(1), pp27-32, ISSN 2180-4370, (2019).

[23] Nurislam Tursynbek, Aleksandr Petiushko, Ivan Oseledets , "Robustness Threats of Differential Privacy", arXiv preprint arXiv:2012.07828, 2020 - arxiv.org, Aug (2021).

[24] Xingxing Xiong, Shubo Liu, Dan Li, Zhaohui Cai, and Xiaoguang Niu "A Comprehensive Survey on Local Differential Privacy", Security and Communication Networks, Article ID 8829523, 29 pages, (2020). https://doi.org/10.1155/2020/8829523.

[25] Josep Domingo-Ferrer, David Sánchez, and Alberto Blanco-Justicia, "The Limits of Differential Privacy (and Its Misuse in Data Release and Machine Learning)", Communications of the ACM, Vol. 64 No. 7, Pages 33-35, July 2021, https://doi.org/10.1145/3433638.

[26] Mathew E. Hauer1 and Alexis R. Santos-Lozada, "Differential Privacy in the 2020 Census Will Distort COVID-19 Rates", Socius: Sociological Research for a Dynamic World, Volume 7, pp. 1–6, (2021).

[27] Revathy Swaminathan and T. Arun Kumar, "Survey paper on privacy preserving data mining", International Journal of advanced Research, pp 1120-1127, ISSN: 2320-540, (2016).

[28] Jyothi Mandala, Pragada Akhila, Vulapula Sridhar Reddy, "Integrated Reinforcement DQNN Algorithm to Detect Crime Anomaly Objects in

Smart Cities", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 12, No. 12, (2021).

[29] Fabian Prasser and Florian Kohlmayer, "Putting Statistical Disclosure Control Into Practice: The ARX Data Anonymization Tool", Gkoulalas-Divanis, Aris, Loukides, Grigorios (Eds.): Medical Data Privacy Handbook, Springer, November. ISBN: 978-3- 319-23632-2. (2015).

[30] Preeti Gulia, Hemlata, "Privacy preserving data mining of vertically partitioned data in distributed environment-an experimental analysis", Journal of Theoretical and Applied Information Technology, Vol.96. No 10, ISSN: 1992-8645, pp 2973- 2987, (2018).

[31] Jaap. Wieringa, et.al, ''Data analytics in a privacy-concerned world,'' Journal of Business Research., Volume 122, pp. 915–925, (2021).