DOI: http://dx.doi.org/10.26483/ijarcs.v16i2.7203

Volume 16, No. 2, March-April 2025

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

SECURING HEALTHCARE DATA FROM RE-IDENTIFICATION ATTACK USING A HEURISTIC DATA ANONYMIZATION MODEL WITH PRIVACY AND ACCURACY

Dr. D. Anuradha
Assistant Professor, Vellore Institute of Technology,
Vellore, Tamil Nadu, India

***Abstract*:** New technologies in healthcare industry provide improved quality of treatment with reduced cost. These technologies deals with the valuable sensitive data and they should be kept safe while we have to maximize their usage. In this paper a novel and efficient data pre-processing method is used which improves the completeness, accuracy and appropriateness of the chosen dataset. The proposed data anonymization method computes different data intervals and replace original sensitive data with computed values. Tests are conducted with a lung cancer dataset, ML algorithms and an existing data anonymization model. The test results show the effectiveness of the model against re-identification attack and the improved accuracy in predicting the lung cancer possibilities.

Keywords: Data pre-processing, Data cleansing, Data anonymization, Healthcare data, Fixed intervals of attribute values, Masking and encoding attribute values.

1. INTRODUCTION

New innovations and inferences from the healthcare's data is very much vital for inventing advanced medicines and to diagnose the disease accurately. To achieve these results Patient Health Records (PHR) are to be shared among researchers and pharmaceutical companies [1]. These researches reduce the cost of treatment and availability issues as well. Telemedicine and remote healthcare facilities made the treatment very easy and advanced. It has been observed that researches on medical data is very much essential for the development of e-healthcare sector and to introduce innovative solutions of treatments [2]. Some of the prominent applications of e-healthcare services are depicted in figure1.

Even though patient data analysis yields many advantages, data providers are very cautious about sharing their data due to privacy issues. Privacy preserving data publishing [3] enables data sharing is completely protected with the help of effective tools and methods. Sharing data among different users involve better understanding of data and new innovative ideas [4, 5, 6]. Only trust-worthy data service providers can help in arresting unethical data leakage and it has become an unavoidable requirement in e-healthcare industry.

According to a data privacy study [4], a dataset containing anonymized personal information was shared for research purposes. Since it was publicly made available an adversary with some background information re-identified the individuals with their sensitive information. Various data privacy attacks and the possible protection techniques are discussed in [7]. Actually, the attacks on health care data privacy are evolving day by day. Hence, the data privacy solutions for health care sector should also require intensive research to address these data privacy attacks. This study insists that the need of a novel data privacy technique is inevitable.

Managing data is a convoluted process which requires skilful efforts at different stages of data life cycle as shown in figure2. Attack on data can happen at all stages in the data lifecycle. The malicious data breach can be better understood with the diagram given in figure3. It is possible that any adversary can acquire the data from any data source and can use classy and refined methods of data re-identification to link different datasets [8]. Hence, an upgradation of the specified method is obligatory to preserve the data privacy. Long-time research on data privacy have resulted in various solutions like, k-anonymity [4]; l-diversity [9], (a,k)-anonymity [10] and t closeness [11] etc. but, adversaries can still trace down the required information using effective malwares [12].

A detailed study on data generalization is presented in here. This novel work on data privacy using generalization method is twofold. Firstly, the numerical data such as age and zip code in the PHR are anonymized using fixed intervals and data masking techniques. Then the text data such as gender and disease are anonymized using substitution method. This data anonymization scheme will provide the data security professionals and researchers a new approach in data science and help them to device further new methods of data anonymization. This research paper is organized as few sections of research. Section 2 elaborates how data privacy is essential in electronic health care records. In section 3, related works on data generalization are discussed. The proposed research idea is presented in section 4 and the section 5 deals with the novelty and advantages of the proposed scheme. The final conclusion and further research ideas and extensions are discussed in the last section.

2. THE NEED OF DATA PRIVACY IN E-HEALTH CARE SYSTEMS

The evolution of health care industry resulted in understanding of the importance of availability of heath care services at any time and any place. This reassures that having e-health care data in the cloud environment to confirm availability of heath care services [13]. Health care services in the cloud environment are more vulnerable to data privacy attacks. Confidentiality and integrity of data are the main concerns in the cloud computing for assuring data privacy [14]. These data security and privacy requirements ensures the expected way of data publication. And, we have to be very clear in understanding these security and data privacy requirements.

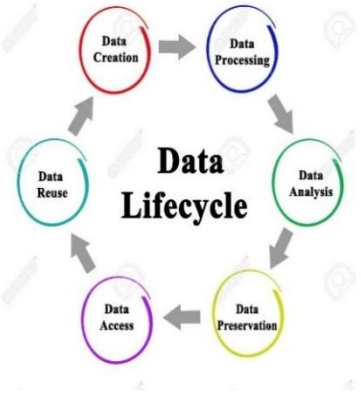

Figure1–Applications of e-healthcare services

Figure2 - Data lifecycle

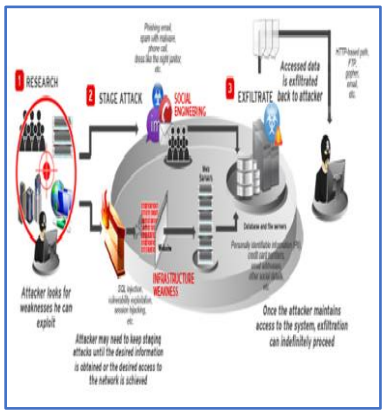

Figure3 - Malicious data breach

Table1: Comparative study of data privacy techniques

	Data privacy technique	Methodology	Uses
1	Generalization	It is a prominent method to preserve data privacy using classical categories or values.	Hiding personal identity and sensitive data.
2	Suppression	It can be treated as a variation of data generalization techniques in which the original values are substituted by a masking value.	Helps in arresting spoofing attack, assuring patient consent.
3	Pseudonymization	Pseudonyms are used for replacing the sensitive data	Data consistency and anonymity are preserved
4	Bucktization	In this method, user data are categorized and placed in different data buckets. Separate security schemes are implemented for buckets.	Protecting data from accidental disclosure and data relevance are assured.
5	Slicing	The data elements are sliced into many slices and kept in relevant memory locations.	Helpful in data security audits, hiding the identity of the person whose information is available for data analysis.
6	Randomization	User data is added with a specified amount of noise data so that the resultant data become more random. This will make the attacker to put more effort to retrieve user's sensitive data.	Data becomes more anonymous; attacker job becomes more tedious and complex
7	Cryptographic methods	More prominently used to secure cloud data.	Assurance, authenticity, audit, authorization confidentiality, integrity, nonrepudiation

2.1 Data anonymization

Since sharing data has become inevitable for every business and organization, user data privacy has also gained its importance in data sharing. Data privacy involves data modification such that linking back the information becomes impossible [15 - 18]. Even if the attacker gathers information from different information sources, it should be made impossible for the attacker to link the details about an entity in the database.

2.2. Access control

Access control policies restrict unauthorized access to the information. Different categories of users are created as user profiles. Each user profile has different priorities of data access and data modifications. Each legitimate user is

assigned with a user profile and the operations which are eligible in the profile alone will be permitted for the user. Thus, the access control guarantees proper usage of data [19].

2.3. Non-repudiation

In all database systems each activity done by all the users are stored as activity log. So, no user can claim that the activity done by that user is not done by the same user. In the e-health care systems, this activity log information is used for analysis to infer facts about patients, doctors, pharma companies, insurance companies and other stake holders [20].

2.4. Trust management

For any two entities to share their information a very strong trust agreement has to be established between them. Both the entities should agree to share only reliable and accurate information among them. We can design our own trust agreement policies and their implementation. When we want to publish or access e-health records we can get the help of cloud service providers. In this scenario, the trust agreement among data publisher, data user and the cloud service provider have to be made in a fair and appropriate way [21].

2.6. Security audit

The typical security audit examines the activity log for any abnormal activities and security attacks happened. The analysis of such activities provides a better understanding of security threats which will help in preventing such kind of attacks in future [22]. The general guidelines provided in most of the security audit reports are proper implementation of user authentication, data access authorization, multi-factor authentication, usage of bio-metric credentials, etc.

2.7. Data consistency

Data security breaches cause most of the data consistency issues. In the context of e-health records, it is important to analyse latest and reliable information to arrive at accurate predictions. So, data consistency of e-health records has to be ensured by implementing proper security measures [23].

2.8. Patient's consent

The data owner has full right over his/her own data and he/she can decide whether to broadcast or not his/her data to researchers [24]. So, obtaining the patient's consent to disseminate his/her information for analysis purpose is must for the cloud service providers before publishing the data.

2.9. Data backup

As information of each patient in a health care environment grows exponentially it is not advisable to archive entire e-health records in any medical institution [25]. Instead, we can get the help of reliable cloud service providers. The e-health records can be compressed and stored in cloud data centres till the lifetime of the patient.

2.10. Data security attacks

Security attacks can do passive and active attacks on legitimate user communication and data to get the sensitive information such as username, password, OTP, pin, security questions, captcha, etc. Once the attacker gets these credentials, he/she can do any malicious attack on valuable data. The most dangerous attack is insider attack in which the person who is authorized to access entire data happens to be an attacker. In health care environment, the doctors and hospital authorities can become inside attackers which is very dangerous [26].

3. RELATED WORKS

Since e-health care systems are distributed, they ought to face certain challenges in preserving data integrity, confidentiality, and data availability [12]. Not only these non-cryptographic techniques but also some cryptographic techniques such as user authentication and authorization, user access controls, managing user credentials, usage of biometric information, etc. are also used [13] in e-health care systems. Even though these techniques are working efficiently, they require more efforts in monitoring and maintenance of the system. A comparative study of different types of privacy preserving techniques [2] are shown in the table1. Since the data analytics involve huge amount of data, researchers prefer data to be available in a cloud environment. In this scenario, cryptographic protection of data is well suited when compare to other non-cryptographic solutions cited above [27 - 30].

An informative survey on storing and retrieving of e-health records over cloud environment is done by [31]. They concluded that it is efficient to share the e-health records over clouds rather than investing on dedicated data centres. A comprehensive literature survey done by [32] helps us to understand the challenges in providing security and privacy, collaboration and secure sharing of e-health record systems. Data generalization is a data privacy technique in which quasi-identifiers in data records are replaced with less sensitive data, while preserving data consistency. The data privacy models proposed by [5, 6, 9, 33] anonymize the dataset using k-anonymity techniques. Still, they lack the ability to protect the dataset from attribute disclosure. K-anonymity is a suitable technique for protecting against the linking attacks. However, the data owners or data providers have to consider that data utility while strengthening the data privacy to make sure that the data published for research is useful.

4. PROPOSED HEURISTIC DATA ANONYMIZATION MODEL

In this section, the proposed model is discussed in detail. The dataset which is used for implementing this system contains the details of the out patients of a hospital. The proposed anonymization system has two segments, namely,

- Data pre-processing
- Data anonymization scheme

4.1. Data pre-processing

It is always preferable to choose a dataset which is almost complete, i.e. the missing data values are very less. Also, the existing values in the dataset are expected to be appropriate and accurate. If the chosen dataset is lacking these properties, then appropriate data pre-processing is needed, so as to make the accuracy and quality of the dataset are improved. In the proposed model, there are two segments for data pre-processing, which are discussed here.

4.1.1. Data cleansing

In this step of data pre-processing, the empty data cells in the dataset are removed so as to improve the accuracy. Since, filling up the empty data cells with some relevant data may reduce the overall accuracy of the dataset, it would be better to remove those empty data cells. But it may reduce the number of rows and attributes of the resulting dataset to a significant amount. So, in this proposed scheme, utmost care is taken in each step for the better result.

The row or the column in the dataset, which has more empty data cells, i.e. more than 2/3 of the total data values are missing then that row or the column is removed from the dataset. To do this we have two approaches.

- First, the rows in which the number of attribute values available is less than 1/3 of the total attribute values (or number of empty values is more than 2/3 of the number of attributes) are considered. Such rows are identified and removed. Then consider the columns with empty data cells more then 2/3 of the total number of rows in the current dataset. Such columns are identified and removed from the dataset.
- In the second approach, we first consider the columns. The columns having empty data cells more than 2/3 of the number of rows are identified and removed from the dataset. Then, consider rows with empty data cells more than 2/3 of the number of attributes in the current dataset.

4.1.2 Data fill up

It is obvious that the proposed data cleansing scheme will not remove all the empty data cells. Along with empty data cells there are some more data cells which require more appropriate data values to be replaced. These data cells are termed as defected data cells. Hence, the cleansed dataset ought to be processed further for treating the defected data cells. The following steps are carried out to fill up more accurate values in the place of defected data cells.

a) Identifying defected data cells

The defected data cells are classified into one of the three types listed below.

- Appropriate value to be placed in the existing empty cells
- Spurious/wrong data values have to be modified with appropriate values
- The data values which are not in the required format need to be replaced with appropriate values.

Once the defected data cells are identified, the appropriate data values are computed by considering most similar rows as that with the row in which the defected data cell is found.

b) Identifying most similar rows

To identify the most similar rows, the quasi-identifiers of the dataset are used. The quasi-identifiers are the attributes that directly identify an individual row in the dataset, i.e. personal identity number, phone number, age, zip code. This similarity is computed using cosine similarity as follows:

$$\text{Similarity}(A,B) = (\sum_{i=1}^{n} A_i.B_i)/(\sqrt{\sum_{i=1}^{n} A_i^2}).(\sqrt{\sum_{i=1}^{n} B_i^2})$$

where A and B are the two rows. Both the rows are having ‘n’ number of attributes. A matrix is formed with all possible combinations of A and B. Thus, the most similar rows can be identified.

c) Filling up appropriate data values

The dataset that is chosen is having only characters and digits. The value to be placed in a defected data cell is computed by considering similar rows and the corresponding attribute values. The mean of all the numerical values of the attribute of all the similar rows is the value to be set to the defected data cell. If the defected data cell requires a character string value to be replaced, then the most frequent data value of the corresponding attribute of the similar rows is chosen. By finishing this step, a more accurate and complete dataset is obtained which is ready for applying proposed data anonymization scheme.

4.2. Data anonymization scheme

In the proposed model two steps are considered in applying data anonymization.

- Removal of quasi-identifiers and sensitive data
- Applying anonymization scheme

4.2.1 Removal of quasi-identifiers and sensitive data

The quasi-identifiers are the attributes that are used mainly for identifying any individual. For example, personal identity number (SSN, Aadhar number), name, phone number, address are some of the examples of quasi-identifiers. In order to preserve the data privacy of the users whose data is present in the dataset; we can simply remove the quasi-identifiers. But this kind of valuable data absence in the data analytics will reduce the usability of the dataset used. So, we have to very carefully design an anonymization scheme which will completely hide the individual identities from outside world, while improving the usability of the dataset for analysis purpose. On the other hand, if the data owner does not want certain data values not to be disclosed or published then those data values must be removed from the dataset before publishing the dataset publicly.

4.2.2. Applying anonymization schemes

Instead of removing the sensitive information, which would be very useful in data analysis, from the dataset we can anonymize them so as to preserve the data privacy. In the proposed model, four anonymization methods are implemented.

- Masking attribute values
- Generalization of attribute values
- Encoding attribute values

4.2.2.1. Masking attribute values

This method will hide the data values by any chosen character. The patient identity number and phone number of the individuals are chosen for data masking. The first half of the patient id is replaced with a letter ‘p’ and the second half ph_no attribute is replaced with the character ‘*’ as shown below in the table2.

Table2 – Masking the attribute values

	Original value	Masked value
p_id	*621343*	*p343*
ph_no	*9991234567*	*99912******

With the first half of the data no one can identify the individual, hence the privacy of the people is preserved. Still, this anonymized information is very much useful in analysing and relating the information.

4.2.2.2 Generalization of attribute values

In this step, the numerical data values are anonymized. First, the range of the values is identified and the number of intervals is decided with reference to the level of privacy disclosure. The interval is computed as follows:

Interval I = (h – l)/n,

where, h = highest value of the attribute, l = lowest value of the attribute, and n = number of intervals.

From the chosen dataset age and zip code attributes are anonymized in this step and first we discuss about how age attribute is anonymized. In the dataset segment taken for study, it is observed that highest age value is 78 and lowest age value is 12. The required number of intervals is decided to be 3. Hence, the interval is computed as

Interval I = (78-12)/3 = 22

After computing the interval, the interval buckets are finalized as follows:

i) 12 – 34
ii) 35 – 57
iii) 58 – 78

4.2.2.3 Encoding attribute values

In the chosen dataset the gender attribute values will be either ‘Male’ or ‘Female’. To anonymize this attribute encoding method is used. The ‘Male’ and ‘Female’ values are replaced with 0 and 1 respectively.

5. EXPERIMENTS AND RESULTS

In this section the results obtained from testing the proposed model of data anonymization is discussed in detail. In the proposed model the predefined number of intervals influences the constant intervals to be used to generalize the sensitive information, which produces better outputs when compared to replacing approximate values. The proposed model can be applied in many data science applications and guarantees both ultimate data privacy and utmost data usability. The proposed model shows many advantages over other existing system for providing data anonymization and the proposed model is rigid against the security threat identity and membership disclosure, which most of the existing data privacy system. This system is very much suitable for the applications in which the end user interaction and query processing are huge and very frequent. Also, the proposed model makes the data hacker’s analysis of the system to be more complex and time consuming. Thus, the proposed data anonymization system provides a better promising environment for sharing the valuable information for efficient data analysis, while preserving the data privacy of the users. The results obtained in implementing the proposed model is evaluated in two aspects, i.e. data privacy and data usability.

Once the intervals are identified then the actual attribute values are to be replaced with a new mean value, which is computed as follows:

Mean value V = (b_1 + b_2 + (v_1 + v_2 + … +v_n)/(n +2)

where b_1 and b_2 = Boundaries of the interval

v_1, v_2, v_n = Values in the interval

n = No. of values in the interval

The calculated mean values for the three intervals are given below:

Mean value for first interval I_1 = (12 + 34 + 12 + 22 + 32 + 15 + 28) / 7 = 22

Mean value for second interval I_2 = (35 + 57 + 51 + 45) / 4 = 47

Mean value for third interval I_3 = (58 + 80 + 67 + 78 + 75 + 59 + 63) / 7 = 69

The same process is followed for anonymize the zip code attribute values. In this case also, the number of intervals is assumed to be 3. Then the intervals are computed as cited above.

Interval IV = (588344 – 561289) / 3 = 9018

The intervals are then computed as

i) 561289 - 570307
ii) 570308 – 579326
iii) 579327 - 588345

The new mean value for replacing the actual values is,

Mean value for first interval IV_1 = 565992
Mean value for second interval IV_2 = 574277
Mean value for third interval IV_3 = 583607

5.1. Results in providing data privacy

The proposed model shows strong resist against the identity disclosure vulnerability. The prominent data privacy attacks such as background knowledge attacks, identity disclosure, and membership disclosure are handled efficiently in the proposed model of data privacy. If attackers have already known some information about any individual, whose information is published for data analysis, they can easily infer other sensitive information about that individual. Even if the data is anonymized with classical generalization method, sensitive information can be derived from the published anonymized data. But in the proposed model these attack spaces are greatly reduced.

Banet et al [34] mention that the European GDPR states that to assess the residual risk of anonymized data, one should account the reasonable means that can be employed to achieve re-identification. Banet et al. categorizes the reasonable means of re-identification attacks into three categories of attack resources as follows.

1. Background knowledge breadth: This is the number and type (size of the group) of entities to which the target or victim belong to. In the proposed model one cannot find any group. Because in the dataset, instead of generalized data, the data interval is replaced with a new mean value, which will never give any clue for the adversaries to infer any sensitive information about any individual. Still, the more appropriate attribute value will be more useful in getting an inference from the data analysis. For example, the age and zip code attributes are categorized into six categories and the gender is categorized into two categories. Hence, the

1000 tuples are categorized into six unequal categories and only mean values are stored instead of the generalized data. So, the re-identification of a person is not possible logically, unless the adversary has full personal information about the person.

2. Background knowledge depth: This corresponds to the quantity and quality of the background information of an entity i.e. the number of quasi-identifiers in the dataset used to identify any individual. In our model all the quasi-identifiers are removed and hence the re-identification attack is not possible.
3. Computational capabilities: The computational capabilities of the adversaries influence more on the attack. But, in the anonymized dataset generated by our anonymization model does not have any clue to identify any individual that irrespective of the computational capabilities of the adversary, he/she cannot identify any individual.

5.2. Results in providing data utility

During the data anonymization process, it is obvious that the accuracy of the dataset would be decreased. Still, with the proposed model we can expect a required data accuracy which is better than the well-known and efficient data anonymization method named, IACK[35]. A lung cancer dataset with 1000 tuples is used for testing the accuracy of the proposed data anonymization model. This dataset consists of five non-numerical and 20 numerical attributes. Among the 25 attributes, four attributes are identified as quasi-identifiers. The IACK method and the proposed model are applied over the lung dataset separately. The outputs of these two data anonymizing methods along with the original, un-anonymized original dataset (3 datasets) are trained using two machine learning algorithms, namely, Random Forest and Regression trees. Once the data training is completed there were 3 data models trained by Random Forest and 3 data models trained by Regression trees. So, finally 6 trained data models are ready for testing the accuracy of the prediction. For the testing purpose 6 separate datasets are chosen as shown below.

Test dataset1 – Have 5 tuples
Test dataset2 – Have 10 tuples
Test dataset3 – Have 50 tuples
Test dataset4 – Have 100 tuples
Test dataset5 – Have 150 tuples
Test dataset6 – Have 200 tuples

With these 6 datasets and 6 trained data models I have conducted 36 tests and the results are recorded. The accuracies of all the results are shown in the figure 4 and 5. When we analyse the test results it is clearly proved that the proposed data anonymization system provides a better promising environment for sharing the valuable information for efficient data analysis, while preserving the data privacy of the users.

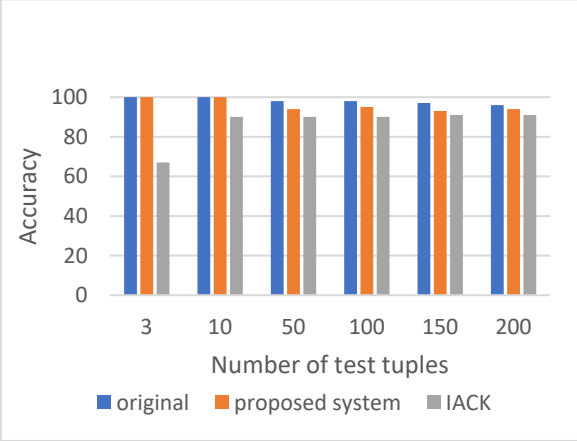

Figure4 – Comparative results using Random Forest

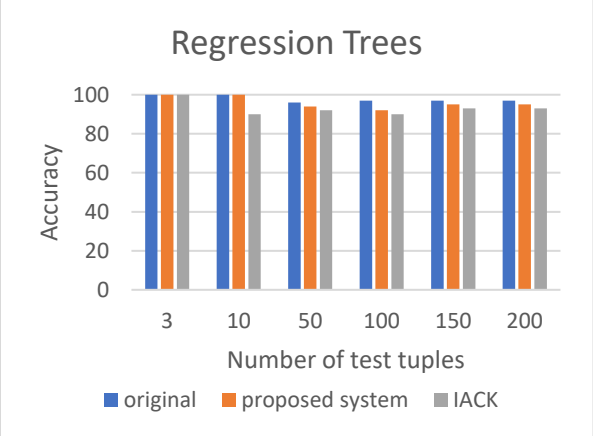

Figure5 – Comparative results using Regression trees

6. CONCLUSIONS AND FUTURE WORK

There are many researches are being carried out in the area of privacy of e-healthcare data. This paper provides a basic understanding of various data privacy techniques which were producing good result. By examining these existing systems, a new idea in data privacy was devised and implemented in the proposed model. In the data pre-processing step of the proposed model the chosen dataset is made more complete, accurate and appropriate for implementing anonymization. In the second step of this proposed model, instead of removing or hiding more valuable and sensitive information they are anonymized in an efficient way. Thereby the quality of the dataset was improved in terms of data privacy and data usability.

The proposed model can further be improved in the following areas.

i. The proposed model does not include any authentication procedure to verify and/or validate the users of the anonymized dataset. This authentication becomes important because the entire dataset is made available for all the users.
ii. The access control policies can also be considered for restricting the unauthorized access by anyone.
iii. In the proposed model it was assumed that the anonymized dataset is kept secure. Hence the encryption of the dataset is not considered. The time taken for encrypting and decrypting the dataset have significant effect on the overall performance of the system. So, an appropriate and efficient encryption algorithm have to be chosen.
iv. The proposed model can be extended to have data validation in which the attribute values which are not in required format can be handled in a better way so as to improve the accuracy of the dataset.

REFERENCES

1. A. Gkoulalas-Divanis, G. Loukides and J. Sun, "Publishing data from electronic health records while preserving privacy: A survey of algorithms," Journal of Biomedical Informatics, pp. 4-19, 2014
2. S. Murthy, A. Abu Bakar, F. Abdul Rahim, R. Ramli, A Comparative Study of Data Anonymization Techniques, 2019 IEEE 5th Intl Conference on Big Data Security on Cloud (Bigdata Security), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS), 27-29 May 2019, DOI: 10.1109/BigDataSecurity-HPSC-IDS.2019.00063
3. A. Majeed and S. Lee, "Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey," IEEE Access, vol. 9, pp. 8512-8545, 2021.
4. J.-H. Weng and P.-W. Chi, "Multi-Level Privacy Preserving K-Anonymity," in 16th Asia Joint Conference on Information Security (AsiaJCIS), Seoul, Korea, 19-20 August 2021.
5. Y. Xu, T. Ma, M. Tang and W. Tian, "A Survey of Privacy Preserving Data Publishing using Generalization and Suppression," Applied Mathematics & Information Sciences, vol. 8, no. 3, pp. 1103-1116, 2014.
6. J. Domingo-Ferrer and V. Torra, "A critique of k-anonymity and some of its enhancements," in Proc. 3rd Int. Conf. Availability Rel. Secure., 2008, pp. 990–993.
7. Olatunji, Iyiola E., et al. "A review of anonymization for healthcare data." Big data (2022).
8. A. Farzanehfar and F. H. Y.-A. d. Montjoye, "The risk of re-identification remains high even in country-scale location datasets," Patterns, vol. 2, 2021.
9. A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "L-diversity: Privacy beyond k-anonymity," ACM Trans. Knowl. Discov. Data, vol. 1, no. 1, Mar. 2007. [Online]. Available: http://doi.acm.org/10.1145/1217299.1217302
10. Raymond Chi-Wing Wong, Jiuyong Li, Ada Wai-Chee Fu, Ke Wang, (a, k)-anonymity: An enhanced k-anonymity model for privacy preserving data publishing, In Proceedings of the 12th ACM SIGKDD, 754-759 (2006)
11. N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity," in 2007 IEEE 23rd International Conference on Data Engineering, Apr. 2007, pp. 106–115
12. MARIA RIGAKI and SEBASTIAN GARCIA, Czech Technical University in Prague, A Survey of Privacy Attacks in Machine Learning, ACM Computing Surveys, Vol. 56, No. 4, Article 101. Publication date: November 2023.
13. F. Alhaddadin, J. Gutierrez and W. Liu, "Privacy-aware cloud-based architecture for sharing healthcare information," Auckland University of Technology, Auckland, New Zealand, 2020.
14. Durga Venkata Sowmya Kaja, Yasmin Fatima and Akalanka B. Mailewa, Data Integrity Attacks in Cloud Computing: A Review of Identifying and Protecting Techniques, International Journal of Research Publication and Reviews, Vol 3, no 2, pp 713-720, February 2022.
15. A. Gkoulalas-Divanis, G. Loukides and J. Sun, "Publishing data from electronic health records while preserving privacy: A survey of algorithms," Journal of Biomedical Informatics, pp. 4-19, 2014.
16. S. Mahloujifar, E. Ghosh, and M. Chase. 2022. Property inference from poisoning. In Proceedings of the IEEE Symposium on Security and Privacy (SP'22). IEEE Computer Society, 1569–1569. https://doi.org/10.1109/SP46214.2022.00140
17. Stacey Truex, Ling Liu, Mehmet Emre Gursoy, Lei Yu, and Wenqi Wei. 2021. Demystifying membership Inference attacks in machine learning as a service. IEEE Transactions on Services Computing 14, 6 (2021), 2073–2089. DOI:10.1109/TSC.2019.2897554
18. Zhikun Zhang, Min Chen, Michael Backes, Yun Shen, and Yang Zhang. 2022. Inference attacks against graph neural networks. In Proceedings of the 31st USENIX Security Symposium (USENIX Security'22). USENIX Association.
19. Lumin Shana, Huan Zhoua, DaochengHonga,b, Qiwen Donga, YeWanga, Shubing Song, Application of access control model for confidential data, Procedia Computer Science 192 (2021) 3865–3874
20. Mohamed Sharaf, (2022) non-repudiation and privacy-preserving sharing of electronic health records, Cogent Engineering, 9:1, 2034374, DOI: 10.1080/23311916.2022.2034374
21. Shehab Thabit, Yan Lian Shan, Yao Tao, AL-badwi Abdullah, Trust management and data protection for online social networks, IET Communications published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology, 2022;16:1355–1368
22. Muhammad Farooq, Mohd Rushdi Idrus, Adi Affandi Ahmad, Ahmad Hanis Mohd Shabli, Osman Ghazali, Security and Privacy of Cloud Data Auditing Protocols: A Review, State-of-the-art, Open Issues, and Future Research Directions, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 14, No. 12, 2023
23. Janet Ngesa, Tackling security and privacy challenges in the realm of big data analytics, World Journal of Advanced Research and Reviews, 2024, 21(02), 552–576, https://doi.org/10.30574/wjarr.2024.21.2.0429
24. Roy McClelland, Colin M. Harper, Information Privacy in Healthcare — The Vital Role of Informed Consent, European Journal of Health Law 29 (2022) 1–12
25. Parisasadat Shojaei, Elena Vlahu-Gjorgievska, and Yang-Wai Chow, Security and Privacy of Technologies in Health Information Systems: A Systematic Literature Review, Computers 2024, 13, 41. https://doi.org/10.3390/computers13020041
26. B P Patil, K G Kharade, R K Kamat, Investigation on Data Security Threats & Solutions, International Journal of Innovative Science and Research Technology ISSN No:-2456-2165, Volume 5, Issue 1, January – 2020
27. Mr Melvin Victor1, Dr D David Winster Praveenraj, Sasirekha R, Ahmed Alkhayyat, Abdullayeva Shakhzoda, Cryptography: Advances in Secure Communication and Data Protection, E3S Web of Conferences 399(3), ICONNECT-2023, DOI:10.1051/e3sconf/202339907010
28. K. Pavani, Kondepati Rohini, Jangala Rani, Sai Sree, T. P. Kumar, Avuthu Siva, Swaroopa Rani, P. Yellamma, Data Security and Privacy Issues in Cloud Environment, 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT), DOI:10.1109/ICSSIT55814.2023.10060925, Corpus ID: 257537540
29. Lei Zhang, Huaping Xiong, Qiong Huang, Jiguo Li, Kim-Kwang Raymond Choo, Jiangtao Li, Cryptographic Solutions for Cloud Storage: Challenges and Research Opportunities, IEEE Transactions on Services, 2022 Computer Science, Engineering, DOI:10.1109/TSC.2019.2937764, Corpus ID: 202781137

30. Marios Vardalachakis, Haridimos Kondylakis, Manolis Tampouratzis, Nikolaos Papadakis, Nikos Mastorakis, Anonymization, Hashing and Data Encryption Techniques: A Comparative Case Study, International Conference on Applied Mathematics & Computer Science (ICAMCS), 2023
31. A. Sonya, A Data Integrity and Security Approach for Health Care Data in Cloud Environment, Published in Journal of Internet Services, December 2022, Computer Science, Medicine, DOI:10.58346/jisis.2022.i4.018, Corpus ID: 256177410
32. M. Marwan, A. Kartit, H. Ouahmane, A Cloud Based Solution for Collaborative and Secure Sharing of Medical Data, Published in International Journal of Enterprise Information Systems, 2018 Medicine, Computer Science, DOI:10.4018/IJEIS.2018070107, Corpus ID: 49684062
33. Junqi Guo, Minghui Yang, Boxin Wan, A Practical Privacy-Preserving Publishing Mechanism Based on Personalized k-Anonymity and Temporal Differential Privacy for Wearable IoT Applications, Published in Symmetry, 2021 Computer Science, Medicine, DOI:10.3390/sym13061043, Corpus ID: 235915642
34. Benet Manzanares-Salor, David Sánchez, Pierre Lison, Evaluating the disclosure risk of anonymized documents via a machine learning-based re-identification attack, Published in Data Mining and Knowledge Discovery https://doi.org/10.1007/s10618-024-01066-3, September 2024.
35. Jiuyong Li a, Jixue Liu, Muzammil Baig, Raymond Chi-Wing Wong, Information based data anonymization for classification utility, Data & Knowledge Engineering 70 (2011) 1030–1045.