iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VI **Month of publication:** June 2023

DOI: https://doi.org/10.22214/ijraset.2023.54212

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VI Jun 2023- Available at www.ijraset.com

Health Care Data Privacy Using the Slicing Technique

D Vandana [1], K Likhitha[2], M Priya Latha[3],Aiesha Sidiqa[4], U Bhavani[5]
[1]*Assistant Professor, Department of Information Technology, GNITS, Hyderabad, India*
[2, 3, 4, 5]*Under Graduate Student, Department of Information Technology, GNITS, Hyderabad*

Abstract: ***Data security is essential because it protects an organization's information from fraud, hackers, and even identity theft. To safeguard the security of its information, any business that aspires to run efficiently must have a data protection plan. Data security becomes increasingly crucial as the amount of data created or stored grows. Data leaks and cyberattacks can have disastrous repercussions. Organizations must actively protect their data and regularly upgrade their security policies. A number of solutions, including generalization and bucketization, are proposed to cope with privacy preservation. To preserve privacy when publishing microdata, numerous strategies, including generalization and bucketization, have been devised. According to various studies, generalization leads in some information loss, particularly for high-dimensional data. As a result, it is ineffective when dealing with high-dimensional data. Bucketing is also worthless in this circumstance because there is no discernible separation between sensitive and quasi-identifying aspects in the data, and thus does not prevent attribute membership disclosure. Because of the dimensionality of huge dimensional data, generalization fails, resulting in information loss due to uniform distribution. Membership disclosure, on the other hand, is not conceivable with bucketization. The fundamental purpose of this project is to protect the privacy of users. According to various studies, generalization leads in some information loss, particularly for high-dimensional data. As a result, it cannot be used for high-dimensional data. Bucketing is useless in this circumstance for preventing the publishing of attribute membership data, and it is also useless for data that does not distinguish between sensitive and quasi-identifying traits. Generalization fails for large dimensional data due to the dimensionality and information loss of the uniform distribution. Bucketing, on the other hand, bans membership disclosure. The primary goal of this initiative is to protect users' privacy. Using data slicing also prevents attribute disclosure and creates an efficient method for computing the sliced data that meets the l-diversity criteria. Data slicing, according to experimental results, is more successful than bucketization for sensitive aspects and protects data utility better than generalization. Experiment results show how effective this method is. HTML, CSS, JSP, and MySQL were all used in this project.***
Keywords: Data privacy, Attributes, Generalization, Bucketization, Slicing

I. INTRODUCTION

Publishing microdata while protecting privacy has received a lot of attention lately. Today, the majority of organizations must publish microdata. Microdata is made up of records, each of which provides details about a distinct entity, such as a person. Although many microdata anonymization strategies have been developed, generalization with k-anonymity and bucketization with l-diversity are the most often used. Both techniques divide qualities into three groups; some of them are identifiers that may be used to identify anything specifically, such a name or a security number, while others are quasi-identifiers. Birth date, sex, and zip code are examples of quasi-identifiers, while diseases and salaries are examples of sensitive attributes that are kept secret from competitors. The quasi-identifiers are a group of attributes that, when combined with outside data, can be used to reidentify a person. In both methods, identifiers are first eliminated from the data before the tuples are divided into buckets. The quasi-identifying values in each bucket are changed by generalization into less precise and semantically constant values, making it impossible to distinguish between tuples that are in the same bucket based on their QI values. By randomly permuting the SA values in the bucket, bucketization separates the SA values from the QI values. A collection of buckets with permuted sensitive attribute values make up the anonymized data. When patient data is exchanged, patient identities must be safeguarded. In the past, we employed methods based on kanonymity and l-diversity. While patient data includes numerous sensitive features, including diagnosis and treatment, existing studies mostly focus on datasets with a single sensitive variable. Therefore, neither method is very effective in maintaining patient data. So, by dividing the data both horizontally and vertically, we are offering a novel method for maintaining patient data and publishing. Data slicing is effective for high dimensional data and maintains improved data utility. It can also be used to prevent membership disclosure.

This study introduces the unique technique known as "slicing." To make the data more useful, slicing divides it into horizontal and vertical portions.By placing the attributes in a column that are highly connected, vertical partitioning is accomplished.The process of horizontal partitioning involves bucketing the tuples.

II. RELATED WORK

Sliced data is more useful than generalization and bucketing for improved patient data sharing and data usefulness preservation. Generalization has been demonstrated to result in significant information loss, especially when dealing with high-dimensional data. The data analyst must make the uniform distribution assumption when doing data analysis or data mining operations on the generalization table. This assumption states that any value in a generalized set is equally possible. A different distribution assumption is not possible. As a result, the data value of generalized data is severely diminished. Because each property is generalized separately, the linkages between the attributes in the generalized table are lost. This is a natural limitation of generalization. Although bucketization does not prohibit membership disclosure, it does provide more valuable data than generalization. Second, by employing bucketization, which broadcasts the QI values in their original forms, an opponent can quickly determine whether a certain person has a record in the leaked data. This implies that the bucketization table can be used to ascertain the majority of people's membership information. Furthermore, bucketization needs a clear distinction between QI and SI values.Bucketing breaks down the attribute link between QIs and SAs by segregating sensitive characteristics from quasi-identifying features. In many data sets, it is not always evident which attributes are QIs and which are SAs. As a result, bucketization is ineffective for distributing and storing microdata.

III. LITERATURE SURVEY

1) Ninghui Li, Senior Member, IEEE, Jian Zhang, Member, IEEE, and Ian Molloy have illuminated certain strategies like generalization and bucketization in the study of Tiancheng Li. It is difficult to do generalization on highly dimensional data, and membership disclosure is a problem with bucketization, thus how to safeguard the data is also illustrated. The tuples in the table are divided into buckets using bucketization, and the sensitive attribute and non-sensitive attribute are then separated by randomly permuting the sensitive attribute values within each bucket. The bucket with the permuted sensitive values is now present in the sanitized data. When compared to generalization, data usage in bucketization is high. Bucketization did not prohibit membership disclosure, but that is one of the limits. Since the QI values are provided in their original formats, a hacker can quickly determine whether a particular person's record is present in the public data. A distinct division between QI and Sensitive Attribute (SA) is necessary for bucketization. Many datasets make it difficult to distinguish between the QI and SA. Finally, bucketization violates the attribute correlations between the QI and the SA by segregating the sensitive characteristics from the QI attributes.

2) Dr. T. Christopher noted in the work by V. Shyamala Susan that the generalization loses some information, especially for high dimensional data. As a result, it is ineffective for high-dimensional data. Generalization is the process of substituting a value for one that is less precise yet semantically compatible. This approach is applied at the cell level, where some original values are retained but with greater ambiguity. This makes it more difficult for the attacker to deduce sensitive information. Because there is no visible separation between sensitive attributes and quasi-identifiers in bucketized data, attribute membership disclosure is not avoided. At the cell level, where there is additional uncertainty added to the initial values, this method is applied. The attacker will find it more challenging to determine sensitive information as a result. Data that has undergone bucketization does not clearly distinguish between sensitive characteristics and quasi-identifiers, therefore attribute membership disclosure is not stopped.

IV. PROPOSED SYSTEM

The proposed technology, known as "slicing," retains more data utility while maintaining user data privacy and functioning effectively with high-dimensional data. The unique privacy-preserving approach known as "slicing" offers significant advantages over generalization and bucketization. The software solution aims to use the "slicing" technique to focus on protecting sensitive data in individual medical records. This method separates the data horizontally as well as vertically. Vertical partitioning is accomplished by categorizing qualities into columns based on how they relate to one another. Each column contains a portion of the traits that are highly connected. During the horizontal partitioning operation, the tuples are bucketed.

Finally, the data in each bucket are sorted at random to undermine the relationship between the various columns. The fundamental idea of data slicing is to break linkages between cross-columns while maintaining associations within individual columns.

In comparison to bucketization and generalization, this decreases the dimensionality of the data while maintaining improved data utility. Every tuple has several matching buckets thanks to the slicing procedure. Slicing first divides the qualities into columns. A subset of attributes are present in each column. The tuples are also partitioned into buckets by slicing. A subset of tuples make up each bucket. The table is divided horizontally as a result. The linkage between various columns is broken by randomly permuting the values in each column. The slicing process ensures that every tuple has several matching buckets. The attributes are initially divided into columns by slicing. There is a subset of attributes in each column. Additionally, the tuples are cut into buckets. Each bucket consists of a subset of tuples. As a result, the table is split horizontally. By randomly rearranging the values in each column, the connections between various columns are severed. There have been many privacy-preserving methods put forth, such as generalization and bucketization, but they all result in attribute exposure. The proposed solution, called slicing, consists of three phases: attribute partitioning, column generalization, and tuple partitioning.

A. Slicing Algorithm

1) *Step 1:* We consider a queue of buckets Q and a set of sliced buckets SB in this step. SB is empty, and Q initially has only one bucket, which holds all of the tuples. As a result, Q=T; SB=.
2) *Step 2:* After each iteration, the algorithm divides each bucket into two buckets and eliminates a bucket from Q. For the l-diversity check, Q=Q-B (T,Q,B1,B2,SB,l);The tuple splitting method's primary purpose is to determine whether a sliced table meets the l- diversity criteria.
3) *Step 3:* Using the diversity check process, the statistics for each tuple are recorded. L[t] contains statistics for a single matched bucket B.D(t,B) is the expected sensitivity value distribution and the likelihood of matching.
4) *Step 4:* Using the equation Q=QB1,B2, two buckets are shifted to the end of the Q.
5) *Step 5:* If SB=SBB, the bucket is shifted to SB because there is no further splitting that can be done at this stage.
6) *Step 6:* After computing the sliced table, we return SB if Q is empty. The term "collection of separated buckets" is abbreviated as SB.Finally, return SB.

B. Advantages

1) Slicing is an effective strategy for dealing with high-dimensional data.
2) Slicing can be used to effectively restrict attribute disclosure based on the l-diversity privacy requirement. The basic notion behind l-diversity is that each group must have an acceptable representation of the values associated with the sensitive qualities.
3) Developing a fast approach for computing the sliced table while preserving l variety. The proposed method partitions attributes into columns, employs column generalization, and buckets tuples. The characteristics with the highest association are listed in the same column.

V. SYSTEM ARCHITECTURE

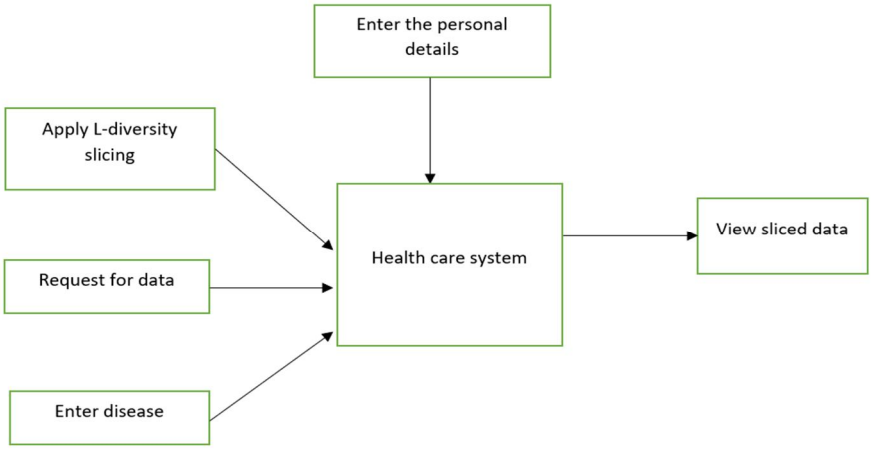


Figure 1: System Architecture

In this system, there are three active actors: administration, a doctor, an employee, and a patient. There is also one system that works together. The staff enters all patient data. He can also look up a patient's details by ID or name. The doctor can examine the new patient and diagnose the ailment based on the symptoms provided by the patient. He can also look up a patient's details by ID or name. The administrator attempts to change the patient's sensitive information by using l-diversity slicing on the data. Slicing is the process of dividing data into both vertical and horizontal partitions and randomly organizing the tuples so that, if a doctor were to try to retrieve patient information, he or she would only receive the information necessary to study the patient's health and would not see the other personal information. The patient has access to his or her own data, or personal profile, and can find out what condition the doctor has identified based on the reported symptoms.

VI. IMPLEMENTATION

Here, we may observe how patients register by providing information based on semi-identifying and sensitive features. The original information is sent to the physician so that he may recognize the patient using the patient ID and identify the ailment based on the symptoms the patient has described. To protect privacy, the data is published in a sliced format.

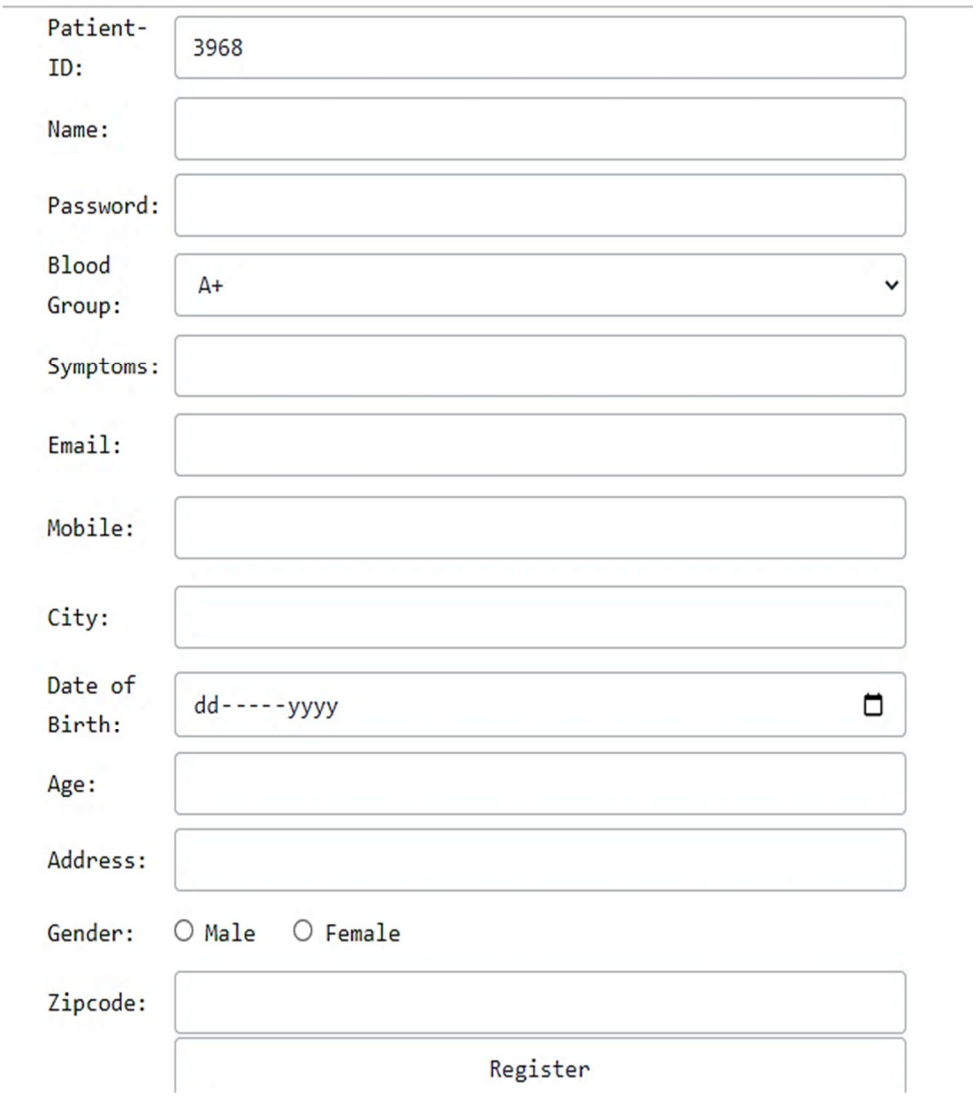


Figure 2: Patient Registration Form

The original data is displayed without any of the patient's personal information, such as name, phone number, and so on.The doctor can also use their patient ID to find a patient. Data is separated into two columns, one for zip code, age, and disease, and the other for location and gender, and then presented.Any user other than the organization's user can view the sliced data.

Patient-ID	Postal Code	Place	Age	Gender	Disease
1062	54368	bangalore	12	Female	covid
1887	500104	Mumbai	20	Female	cold
2565	565432	pune	9	Male	dry cough
3582	504003	delhi	17	Female	high bp
3592	500008	Hyderabad	20	Female	Fever
4219	54543	pune	16	Male	high fever
8525	54541	vizag	10	Male	Viral fever

Figure 3: Original Data

Sno	Place,Gender	Zipcode,Age,Disease
1	(bangalore,Female)	(500008,20,Fever)
2	(Hyderabad,Female)	(54541,10,Viral fever)
3	(delhi,Female)	(54368,12,covid)
4	(pune,Male)	(504003,17,high bp)
5	(vizag,Male)	(54543,16,high fever)

Figure 4: Sliced Data

VII. CONCLUSION & FUTURE SCOPE

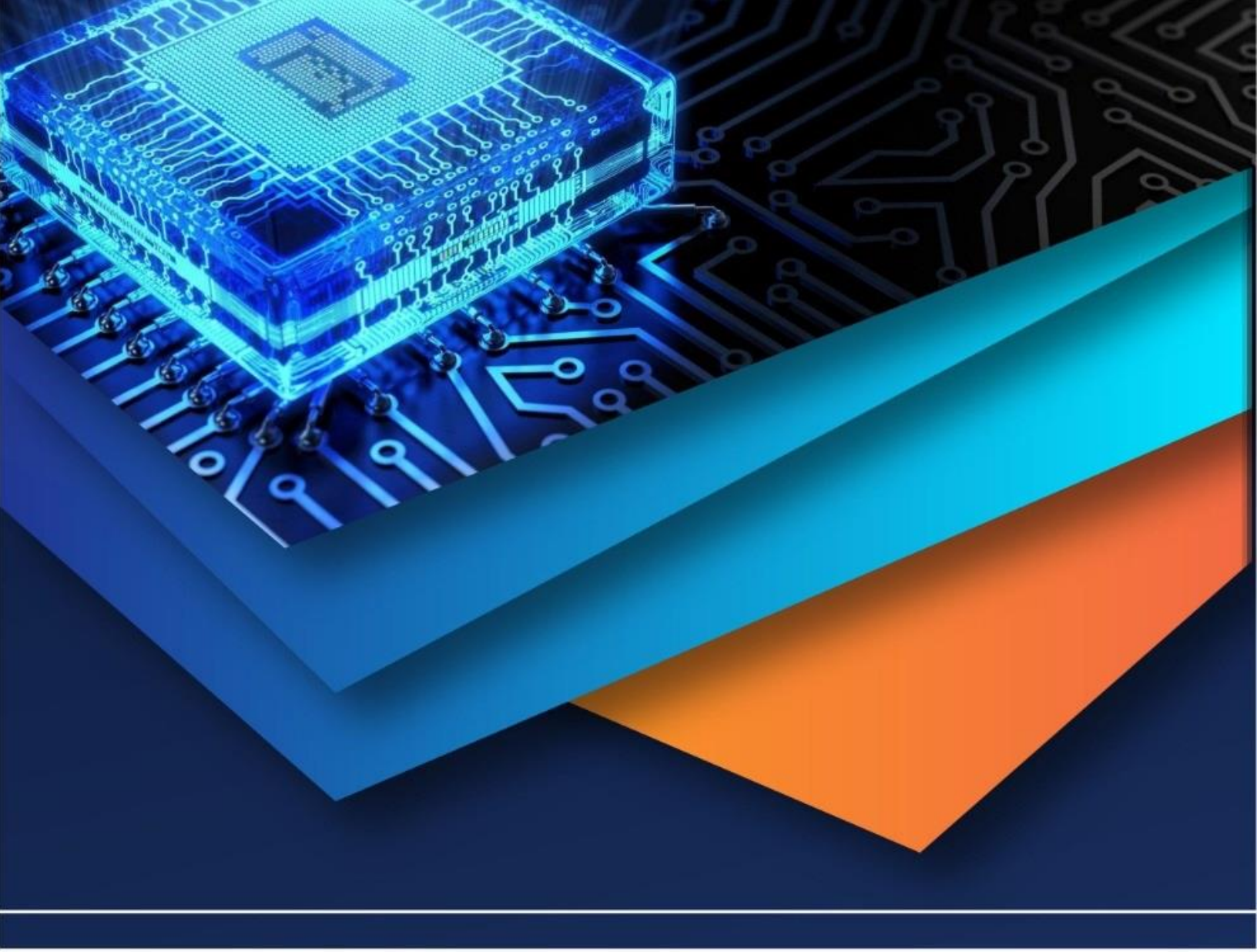
Data slicing, a cutting-edge anonymization technique for data distribution and privacy protection, is presented in this paper. Data slicing maintains improved utility and protects individual privacy while avoiding the drawbacks of generalization and bucketization. We provide an example of how attribute disclosures might be avoided using slicing. The fundamental premise is that when we fully comprehend the data, we can easily create better anonymization methods. Data slicing currently performs better than generalization and bucketization, as we have demonstrated. Data slicing is one potential approach to handling high dimensional data. By categorizing attributes into columns, privacy is maintained. This work's overall technique suggests that one should evaluate the data attributes before anonymizing the data and then make use of these characteristics when doing so. This is based on the idea that enhanced data understanding would enable the development of more effective data anonymization techniques.We provide examples of how attribute correlations can be used to violate privacy. This study has served as an inspiration for many other studies. We first take into account slicing in this work, where each characteristic is in exactly one column. It is a development of the idea of overlapping slicing, which copies an attribute across several columns. More attribute connections are consequently made accessible. For instance, the first column can also have the Disease property. To be more precise, there are two columns: "Place; Gender; Disease" and "Zipcode; Age; Disease." Better data utility might result, but the privacy consequences need further research and understanding. We also want to analyze membership disclosure protection in further detail. According to our research, random grouping is not extremely effective. More effective tuple grouping methods are what we're aiming for.

Third, slicing is one possible approach to handling massive amounts of data. We protect privacy by removing the correlation between uncorrelated variables, while data utility is maintained by keeping the correlation between highly linked features. Second, we want to examine membership disclosure protection in more detail. According to our research, random grouping is not extremely effective. More effective tuple grouping methods are what we're aiming for.

Third, slicing is one possible approach to handling massive amounts of data. We protect privacy by removing the correlation between uncorrelated variables, while data utility is maintained by keeping the correlation between highly linked features.

REFERENCES

[1] Tiancheng Li, Ninghui Li, Senior Member, IEEE, Jian Zhang, Member, IEEE, and Ian Molloy, "Slicing: A New Approach for Privacy Preserving Data Publishing", IEEE Transactions on Knowledge and Data Engineering, vol. 24, NO. 3, March 2012.
[2] V. Shyamala Susan1, Dr. T. Christopher, "A Survey on Privacy Preservation in Data Publishing", IJCSMC, Vol. 3, Issue. 3, pp.188 – 193,March 2014.
[3] Gokila, S, and P. Venkateswari, "A Survey on Privacy Preserving Data Publishing," International Journal on Cybernetics and Informatics (IJCI) Vol. 3, No. 1, February 2014.
[4] Manjusha S. Mirashe, Kapil N. Hande , "Survey on Efficient Technique for Annonymized Microdata Preservation using slicing," International Journal of Emerging Trends in Engineering and Development Issue 5, Vol.2 (Feb.-Mar. 2015) .
[5] SabaYaseen, Syed M. Ali Abbas, Adeel Anjum, Tanzila Saba, Abid Khan, Saif U. R. Malik, Naveed Ahmad, Basit Shahzad, Ali Kashif Bashir, "Improved Generalization for Secure Data Publishing", IEEE 2018.
[6] Sesha Bhargavi, V., Spandana, T. (2017). Recommendation Based P2P File Sharing on Disconnected MANET. In: Deiva Sundari, P., Dash, S., Das, S., Panigrahi, B. (eds) Proceedings of 2nd International Conference on Intelligent Computing and Applications. Advances in Intelligent Systems and Computing, vol 467. Springer, Singapore. https://doi.org/10.1007/978-981-10-1645-5_18
[7] Velagaleti, Sesha & Seetha, M. & SOmalaraju, Viswanadha raju. (2013). A Simulation and analysis of DSR Protocol in Mobile ad hoc Networks. INTERNATIONAL JOURNAL OF MANAGEMENT & INFORMATION TECHNOLOGY. 8. 1279-1286. 10.24297/ijmit.v8i1.692.
[8] Suntherasvaran Murthy, Asmidar Abu Bakar, Fiza Abdul Rahim, Ramona Ramli, "A Comparative Study of Data Anonymization Techniques", IEEE 5th Intl Conference 2019.
[9] I.Ravi Prakash Reddy, "Privacy Preserving Intrusion Detection System for Low power Internet of Things", International Conference on Computational Intelligence & Data Analytics (ICCIDA-2022), 8-9 Jan-2022.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY