IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 6, June 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006165|

Preserving Privacy of Health Care Record using Data Anonymization

Prof. Naresh Patel K M[1], Nivedita K[2], Nishchitha N[3], Madhura S R[4], Nishanth Raj P[5]

Assistant Professor, Department of Computer Science and Engineering, Bapuji Institute of Engineering and Technology, Davanagere, Karnataka, India[1]

B E Student, Department of Computer Science and Engineering, Bapuji Institute of Engineering and Technology Davanagere, Karnataka, India[2]

B E Student, Department of Computer Science and Engineering, Bapuji Institute of Engineering and Technology Davanagere, Karnataka, India[3]

B E Student, Department of Computer Science and Engineering, Bapuji Institute of Engineering and Technology Davanagere, Karnataka, India[4]

B E Student, Department of Computer science and Engineering, Bapuji Institute of Engineering and Technology Davanagere, Karnataka, India[5]

ABSTRACT: In their original form, data have private information about people. If raw data is shared directly, it will be against a person's right to privacy. So, it becomes more and more important to keep published data private. An attacker is likely to be able to figure out who someone is by looking at the public tables and using record linkage, attribute linkage, and table linkage. Algorithms based on data anonymization, i.e., generalisation and suppression, have been proposed to protect sensitive attributes and resist these multiple types of attacks. These algorithms prevent information loss by replacing specific values with more general ones. Here, we suggest a way to protect the privacy of published data by making it anonymous.

KEYWORDS: Data Anonymization, Generalization, Suppression, Sensitive Attribute.

I. INTRODUCTION

The collection of huge amounts of personal information has become a key part of data research. For some purposes, the records of the data that have been collected need to be made public. There may be problems in the data, such as privacy issues, user secrets, and sensitive information. More and more application areas, like marketing, social psychology, and homeland security, use these published data to improve the services they offer to users. The popularity of these kinds of apps also makes it harder to figure out if published data is easy to find. On the one hand, making data public is good for activities like data mining and analysis. For example, many agencies and organisations publish data on public health for demographic research and other reasons. On the other hand, if sensitive information is made public, users' privacy may be at risk if data is made public.

Structured data is data that follows a data model, has a clear structure, is in the same order every time, and is easy for a person or a computer programme to access and use. Most of the time, structured data is kept in well-defined places like databases. It looks like a table with columns and rows that clearly describe what it is. So that its parts can be broken down so they can be processed and analysed more effectively.

A recent study found that a given set of data can be used to identify about 87 percent of the people in the United States. Because of this, it is very important to keep published data, especially sensitive information, private. The original data

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006165|

usually has four types of attributes: explicit identifier, quasi-identifier, sensitive attribute, and non-sensitive attribute. Explicit Identifier is used to identify a person in a unique way, so it is often left out of tables that are made public.

The health industry's technological advances and the digitization of medical records I.e., the switch to Electronic Health Records (EHRs) makes it possible to access health records in real time and from anywhere by using big data. This is done to cut costs and boost profits in the healthcare industry. Malicious attacks, on the other hand, have made it so that there are now risks to the privacy and security of health records.

Sensitive Attribute (SA) is information that is private or unique to each person. Attributes that are not sensitive can be known by anyone without worry. Based on what an attacker knows about QIs or SAs, they are likely to use record linkage, attribute linkage, table linkage, and probabilistic attacks. With a record linkage attack, a specific record in the published tables can be used to figure out who the target user is. Attribute linkage happens when some things about a person, like a disease, become known and can be linked to that person. Table linkage tries to figure out if the victim's record is in the data that has been made public. After looking at the published data, the probabilistic belief about the victim's SA may change in the probabilistic attack.

Data anonymization, also called "generalisation" and "suppression," is the process of replacing values with consistent values that are less specific. Anonymity protects privacy by making sure that each record that is made public is linked to at least k people, even if the records are directly linked to information from the outside world. This paper gives a formal explanation of how to get k-anonymity by using both generalisation and suppression. When you generalise, you replace (or recode) a value with a less specific but still logically correct value. In suppression, a value is not given out at all.

It is hard to protect the privacy of users and keep their identities from getting out through the published tables. To offer these protections, some of the original records should be changed before they are made public. This is called "anonymization." k-anonymity is a well-known way to protect people's privacy when they publish relational data. It was first used in to keep the released data from getting out by making each person anonymous from at least k–1 other people. That is, even if a user's QIs are known, there is no better chance than 1/k of figuring out who they are. Common ways to keep k- anonymous are to generalise and to hide information.

In the generalisation process, some attribute values are swapped out for a broader category, like a parent value in an attribute's taxonomy. During suppression, some of the values of the attributes are changed to special characters like * and #. Another method is perturbation, which changes a set of data by adding noise, swapping values, or making up new data while keeping some statistical properties. Even though these methods help protect the privacy of the user, they often lose a lot of information when they change QIs and SAs. This makes data analysis much less accurate.

It is a way to put human-like thought processes into a control system. It may not be meant to give correct reasons, but it is meant to give good reasons.

II. LITERATURE REVIEW

Existing ways to stop privacy leaks from published data use algorithms to protect privacy when publishing data. These algorithms offer different anonymity operations.

- Gong QY, Yang M, Luo JZ. Data anatomization approach for incomplete microdata. J.Ruan Jian Xuan Bao/Journal of Software: The goal of anatomization and permutation is to break the connection between attributes without changing them. The anatomization approaches break the link between QIs and SAs and make several different tables with attributes that don't overlap.
- Permutation works on the same principle as anatomization. It is used to break the link between a quasi-identifier and a numerically sensitive attribute by splitting the records into groups and switching the sensitive values within each group. All of the pseudo-identifiers and sensitive information are released directly by Anatomy in two separate tables. Both tables share one attribute, the group identifier. Based on the idea of anatomy, anonymized groups are made to keep the social network's structure and tabular data useful. The diversity privacy requirement of sensitive attribute combining with a grouping mechanism in the published social graph is met by a linear-time algorithm for computing anatomized tables.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006165|

- Slicing can make a vertical partition by putting attributes into columns based on how they relate to each other. It can also make a horizontal partition by putting sensitive attributes in each column in a different order to break the link between columns. However, it can't protect the privacy of data that has been made public because it doesn't take into account record linkage attacks and probabilistic attacks. "T-closeness slicing: A new method for publishing transactional data that protects privacy," by M. Wang, Z. Jiang, Y. Zhang, and H. Yang. This method is based on Slicing treats each attribute as a column and doesn't care about how attributes are related to each other. • Slicing is suggested in to protect the privacy of publicly available data. It meets the privacy requirement of diversity by dividing attributes into columns, where each column has a subset of attributes. Slicing also divides the tuples into buckets, and each bucket has a subset of the tuples. Slicing was used to keep the privacy of published data, which has one column for each tribute.
- A new method called "t-closeness slicing" is meant to protect transactional data better from different types of attacks. The anonymization algorithm in uses both anatomization and enhanced slicing to protect the privacy of multiple sensitive attributes while following the principles of k-anonymity and l-diversity. To stop the presence attack, an improved version of anatomy called permutation anonymization divides the original table into groups to meet the diversity requirement. In, a lightweight data privacy method is proposed that uses a pseudo-random permutation to scramble the original data.
- Perturbation changes the data set to protect privacy while keeping some statistical properties. The data are changed by perturbation by adding noise, swapping values, or making up new data. Any method based on differential privacy was used to protect the privacy of the record owner by removing or adding a single record to the published data to stop a probabilistic attack. This was done by adding noise. To get differential privacy, each row of an adjacency matrix was randomly projected into a low-dimensional space, and then random noise was added to the projected matrix.
- A.N.K. Zaman also suggested a randomization algorithm for data sanitization to make the published data in "An improved data sanitization algorithm for privacy-preserving medical data publishing" meet the differential privacy requirements. To protect the privacy of statistical information, sensitive attribute values are swapped between records during data swapping. In synthetic data generation, the original data are replaced with some sample points that come from a statistical model that has already been set up. Random edge perturbation was used to protect the privacy of the publishing data and stop structural identification attacks. The authors wanted to change the length of the shortest path in graphs and keep the structure of the graph the same so that sensitive information could be kept safe. Condensation was another way to make fake data while keeping people's privacy safe.

Literature summary

The main goal is to find out how well anonymization works to protect the privacy of patients. In this study, the important anonymization issues were looked at in four groups: the secondary use of anonymized data, the risk of re-identification, the effect of anonymization on information extraction, and the inadequacy of current methods for different types of documents.

Electronic health records (EHRs) are being used more and more by healthcare providers. This has made it easier for them to talk to each other, get data for other uses, and improve the quality of their services. Since EHRs contain a lot of important information about patients, privacy has become a big issue. Because of this, many ways to protect privacy have been suggested, and anonymization is a common one.

From our review of the literature, we learned that most of the data being made today is structured data and that it is being made in very large amounts. The hospitals and health care centres are where we get our structured data.

Existing System

Existing ways to make sure that published data doesn't leak private information are grouped into three sets of anonymization operations: anonymization, permutation, and perturbation. Most common anonymization operations that

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006165|

are used to protect privacy with k-anonymity. Some techniques of permutation and perturbation try to break the link between two attributes without changing them. The anonymization methods break the link between QIs and SAs and make several separate tables with attributes that don't overlap.

Disadvantages:

1. Existing privacy measures to protect against membership information leaks

2. Makes the current ways of handling data much less useful.

Problem Statement

Processing the data record in a way that makes it impossible to figure out what information is sensitive. To protect a person's privacy, techniques like generalisation and suppression are used to hide information. Also, the privacy of each person needs to be protected while diseases are being ranked.

Proposed Solution

In this system, the goal is to come up with an effective way to make data anonymous so that the published data can be used for accurate analysis and sensitive data labels don't get out. Our plan will protect users' privacy and make sure that data that is made public can withstand record linkage, attribute linkage, and table linkage attacks. We suggest two main ways to reach our goal: generalisation and suppression.

We think about structured, already-processed health care data. After the data has been pre-processed, it will be taken out. After figuring out which attributes are important, we give grades for those attributes.

Advantages

- It can be used to stop attribute disclosure in a good way.
- Our results show that keeps information much more useful than generalisation.

Proposed Objectives

The objectives of the proposed system are as follows:

- Get the structured health record and do some preliminary work on it.
- To identify anonymization techniques available.
- To protect the privacy of Health Care records by making the data anonymous.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006165|

III. SYSTEM DESIGN

SystemArchitecture

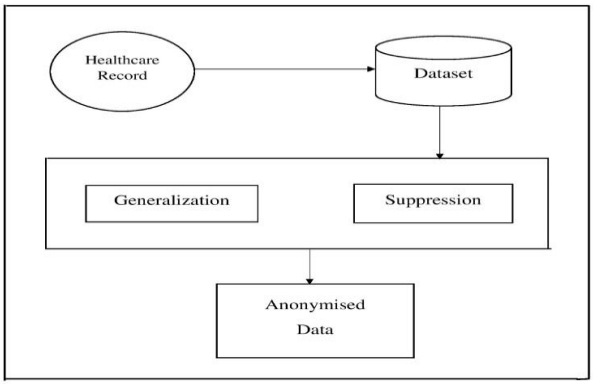


Fig: System Architecture

Methodology Presented

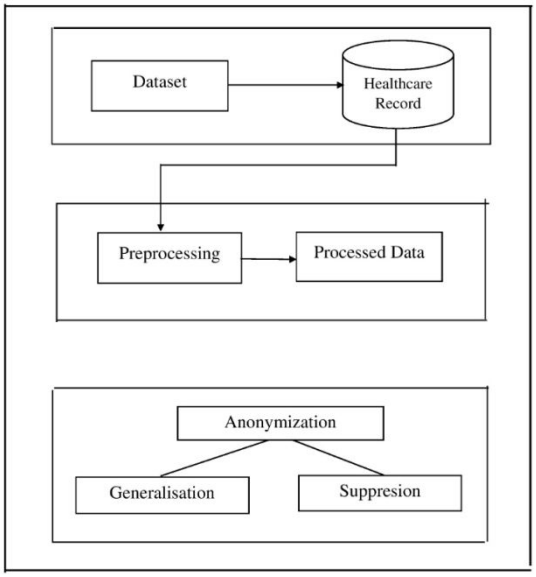


Fig:Methodology Diagram

Detailed description of the methodology

With slicing, the person who collects the data and the person who mines it are two different people. The person who collects the data processes the information from its original owners and then gives it to the person who mines the data. Processing must be done in a way that makes it impossible for the data miner to find out who owns the data and other sensitive information, but still gives the data miner useful information. This is what the project needs most. The anonymization, reconstruction, and modification approaches can help with this.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006165|

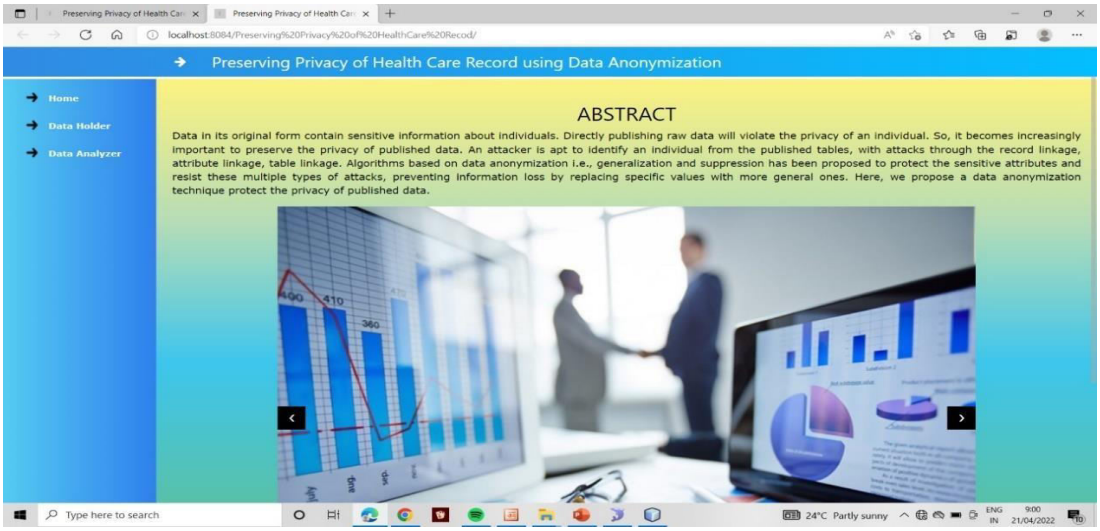
Generalization: It involves replacing a value with a less specific one that still makes sense in terms of meaning.

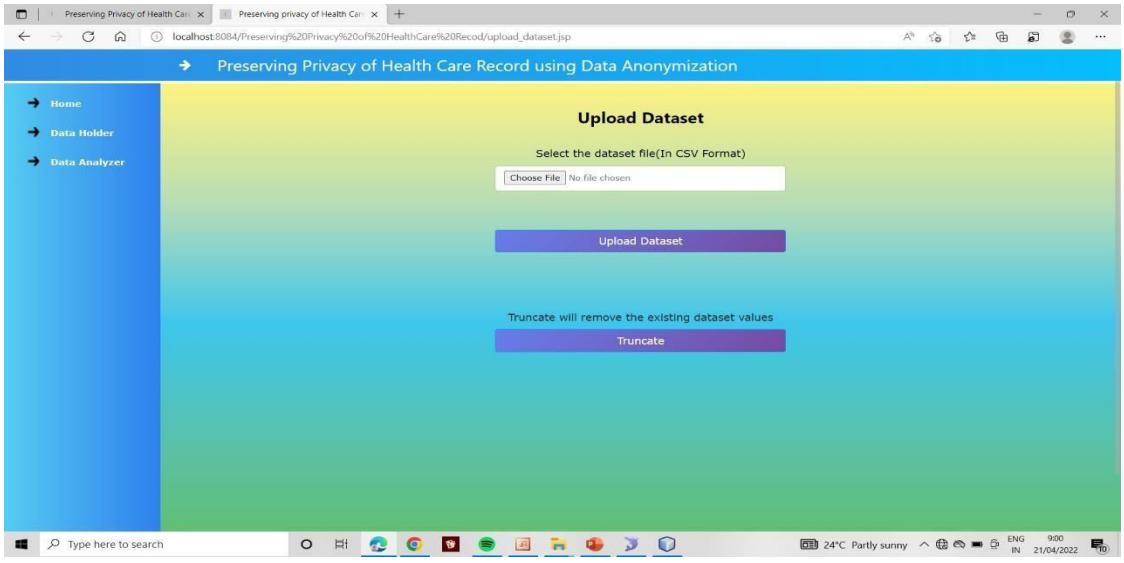
Suppression: It involves changing a value or some information, usually in public reports and data records, to protect people's identities, privacy, and personal information.

Even though there are other methods, combining generalisation and suppression two has a number of benefits.

Dataset: Structured data is data that follows a data model, has a clear structure, is in the same order every time, and is easy for a person or a computer programme to access and use. Most of the time, structured data is kept in well-defined schemas like databases. It is usually in the form of a table with columns and rows that clearly show what it is.

IV. RESULTS AND DISCUSSION





International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006165|

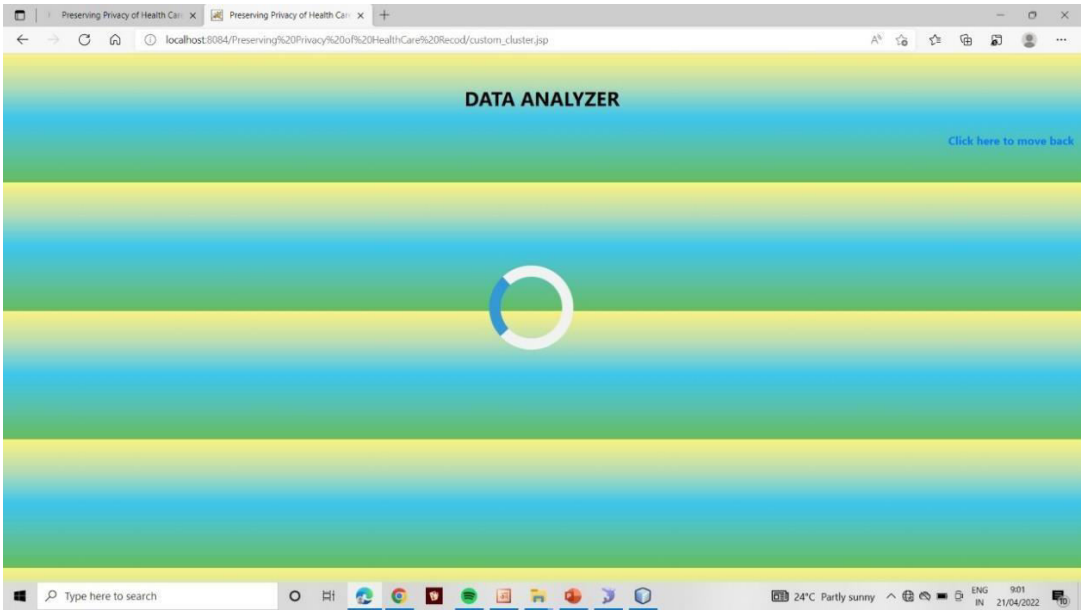


Fig: DataAnalyzing

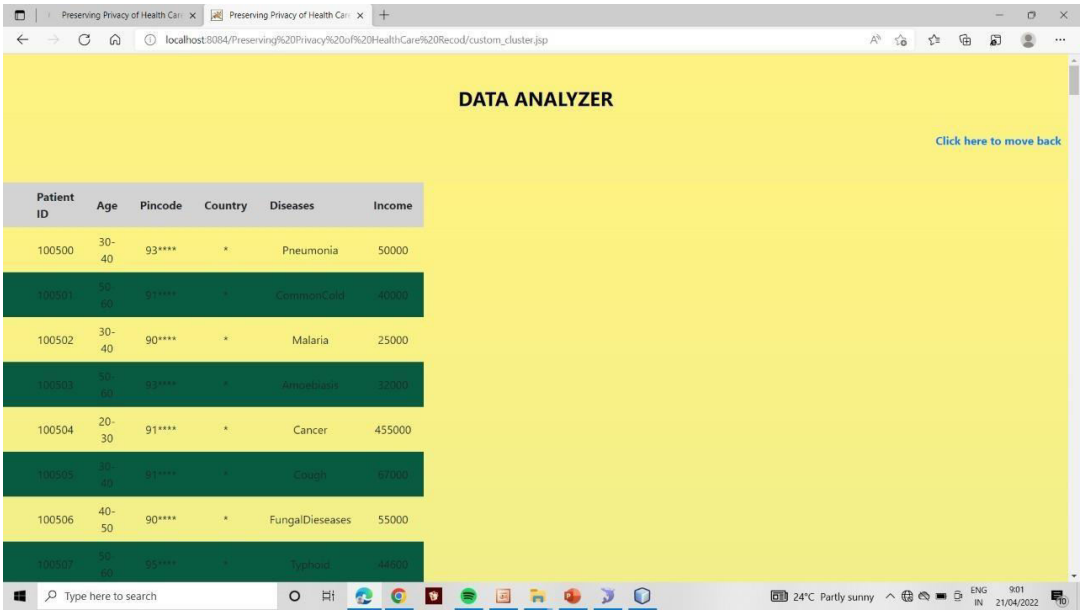


Fig:AnalyzedData

V. CONCLUSION

We propose data anonymization techniques generalisation and suppression. When you generalise, you replace a value with a consistent value. The process of putting together the data into a more general picture. Suppression is the process of replacing a value or selected information, most often in public reports and data records, to protect the identities, privacy, and personal information of individuals.

.

VI. FUTURE SCOPE

Patients' identities, personal information, and privacy must not be put at risk, as this would always cause a lot of physical, mental, and emotional harm. So, medical and health datasets must have a high security priority, and they must

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006165|

be published using privacy-protecting data publishing techniques before they are made available to the public. This is to make sure that a malicious person's actions don't lead to a leak or a probabilistic attack.

REFERENCES

1. M.Wang, Z. Jiang, Y. Zhang,and H.Yang,"T-closeness slicing:A newprivacy-preservingapproachfortransactionaldatapublishing",InformedinJournalon Computing,2018.
2. Jayabalan,M.and Rana,M.E.,"Anonymizinghealthcarerecords:A studyofprivacypreservingdatapublishingtechniques", Advanced ScienceLetters,2018.
3. A.N.K.Zaman,C.Obimbo,andR.A.Dara,"Animproveddatasanitizationalgorithmfor privacy preserving medical data publishing," in Canadian Conference onArtificialIntelligence,2017.
4. V.S.SusanandT.Christopher,"Anatomisationwithslicing:anewprivacypreservationapproachformultiplesensitiveattributes,"Springer Plus,2016.
5. E. G. Komishani, M. Abadi, and F. Deldar, "Pptd: Preserving personalized privacy intrajectorydata publishing by sensitive attribute generalization and trajectory local suppression,"knowledge-BasedSystems,2016.
6. Thakkar, A. A. Bhatti, and J. Vasa, Correlation Based Anonymization Using GeneralizationandSuppression forDisclosure Problems.Springer InternationalPublishing,2015.
7. Poovarasi, V. and Vijay, A.D.," Overlapping slicing: A narrative approach to privacypreservingdatapublishing",Proc.ResearchJournalofComputerSystemsEngineering,2013.
8. Y.Wang,L.Xie,B.Zheng,andK.C.Lee,"Highutilitykanonimizationforsocialnetworkpublishing,"KnowledgeandInformationSystems,2014.
9. Y.Xu,T.Ma,M.Tang,andW.Tian,"Asurveyofprivacypreservingdatapublishingusinggeneralizationandsuppression," AppliedMathematicsandInformationSciences,2014.
10. A.Gkoulalas-Divanis,G.Loukides,andJ.Sun,"Publishingdatafromelectronichealthrecordswhilepreservingprivacy:Asurvey ofalgorithms,"Journal ofBiomedicalinformatics,2014.
11. GongQY,YangM,LuoJZ.Dataanonymizationapproachfor incompletemicrodata.J.RuanJianXueBao/JournalofSoftware,24(12),2883-2896(2013).
12. Faquroddin, M. and Kiran Kumar, G., "A better approach for privacy preserving datapublishingbyslicing",Journalof ScienceandResearch(IJSR),2012.
13. B.Fung,K.Wang,R.Chen,andP.S.Yu,"Privacy-preservingdatapublishing:Asurveyofrecentdevelopments,"ACMComputingSurveys(CSUR),2010.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.165

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details