International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 /doi:10.14445/22315381/IJETT-V68I11P207

Volume 68 Issue 11, 56-62, November 2020
© 2020 Seventh Sense Research Group®

Privacy Preserving Data Mining: Techniques and Algorithms

Ritu Ratra[1], Preeti Gulia[2]

[1.]Ritu Ratra is with Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak(Haryana), India. (e-mail:)
[2.]Preeti Gulia is with Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak(Haryana),India

[1]rituharjai25@gmail.com, [2]preeti.gulia81@gmail.com

Abstract: *There is incredible volume of data that is generated at exponential rate by various organizations such as hospitals, insurance companies, banks, stock market etc. It is done by excellence of digitization of technology. It is well known that very large amount of data is being generated by different electronic devices. This data could be processed to help decision making. However data analytics is prone to privacy violations. There is no doubt that the data analytics is extremely helpful in decision making process, but it will cause some serious privacy concerns. So protect the individual privacy in the process of data analytics became most important and necessary task. In this paper, various threats related to privacy are examined. Techniques and models of privacy preserving are also discussed limitations. Nowadays the role of algorithms of PPDM is very crucial. Today, no doubt a number of PPDM techniques have been grown to preserve the privacy of individual. Some of them are cryptography, secured sum algorithms, perturbation and k-anonymity. Here main focus is on current researches related to PPDM. The paper will enable to understand the different challenges that are confronted in PPDM. It will also help to learn and apply the best applicable technique according to different data circumstances.*

Keywords: *Anonymization, cryptography, Neural Network, Perturbation, Privacy Preserving Data Mining Technique*

I. INTRODUCTION

Data mining is a fundamental processes of discovery of knowledge in databases. Data mining is a branch of research that deals with the obtaining the beneficial information from huge datasets. It can be applicable on different type of applications areas. The datasets that are mined can be of any form like patterns, clusters or may be based on classification. While performing the process of data mining the sensitive data of individual, most of the time, get disclosed to various people. That could be data collectors, data owners, data users and data miners.

There is a massive amount of data that is available to get knowledge about individuals. The source of this is data collected from public. PPDM is a well-known branch data mining that deals with the privacy of individual.

The purpose of this paper is to review the complete literature in current PPDM research. This will very beneficial to get better understanding of existing techniques. There are a number of techniques that can be used to preserve the privacy. To discuss the related techniques this paper is organized. The flow of the further paper is as: Section II provides basics of data mining and privacy concerns. Section III is about the classification of PPDM techniques, Section IV explores the literature survey related to PPDM and explains the research Gap of various studies, and Section V compares the techniques of PPDM. A SECTION VI concludes the whole study.

II. DATA MINING AND PRIVACY CONCERN

There is no doubt that today's world is full of huge amounts of data that are collected on daily bases. To analyse such data is a crucial task. Data mining is a current technology for obtaining knowledge. In this process, there can be disclosure of personal information about organizations and individuals. Confidentiality is the major problem that arises in any massive collection of data. PPDM deals with privacy protection in data mining. Fundamental function of PPDM is to acquire data mining results that are valid and without reveal the corresponding sensitive data. The need for privacy is sometimes due to individual respect or it can be motivated by business interests [3]. Paper [4] defines the PPDM as obtaining valid results of data mining without disclosure of the personal information. The aim of PPDM is to produce valid data mining results through privacy requirements. The study od PPDM can be one of the three approaches: (1) data hiding, personal data can be modified, blocked, in order to secure the privacy. (2) rule hiding, personal data can be drawn out by using some data mining technique. (3) Secure Multiparty Computation (SMC), in this the data that is going to be distributed are encrypted before shared for further uses. PPDM is a growing research area where a various algorithms have been developed, so there is a necessity for integration of different literature that is available to understand the problem, identify implicit research issues, making some novel research ideas, and then make comparison of performance of different techniques and approaches.

A. Privacy Threats in Data Analytics

Major threats in data mining that are regarding to individual privacy are shown in the following table.

TABLE I: PRIVACY THREATS

S. No.	**Name of Threat**	**Description**
1	Surveillance	Surveillance is based on the sentiment analysis. Many opinions about data can be drawn from this. It is possible when somebody continuously monitoring the behaviour of their customer's activities. Surveillance is one of the serious privacy threats because nobody accepts this.
2	Disclosure of data	The data holder shared data to the third party for the purpose of analysis. Anonymization can be used for this purpose. The data analyst at third party can easily map this information with the external data sources.
3	**Discrimination**	Discrimination is also privacy threat which can happen due to inequality. It is some time happens that some personal information of someone is disclosed.
4	Personal embracement	Due to personal embracement private data of a person can be disclosed. It is also a privacy threat.
5	Lack of awareness	Sometimes lack of awareness is one of another reason for various privacy attacks. There are a number of smart phones users who are not aware of this that their privacy is stolen by various applications stored in their mobile. Based on the previous research it can be said that only 17% of mobile phone users are aware of this [9].

III. PRIVACY PRESERVING DATA MINING (PPDM) METHODS & TECHNIQUES

There is no doubt that privacy is very major issue while sharing various types of digital data. Many Privacy preserving techniques were developed like Perturbation, Swapping, and Anonymization etc. According to V. Jane Varamani Sulekha [31] there are four stages of PPDM techniques as given in figure1. At the very first stage, the data or information which taken from different databases are in the form raw data, is more sensitive in nature. At this stage, the raw data is compiled which is collected from various databases. Various PPDM techniques are applied on the raw data for transformations at the second stage. Basically removal of sensitive attribute and quasi identifiers is done at this stage. The PPDM techniques are applied on raw data in second stage as shown in figure1. At the third stage, different data mining algorithms are applied. There could be some modification in algorithms (data mining) for the purpose of protection of privacy. And finally at the fourth stage, output of several data mining algorithms is produced. Different rules and pattern are brought out at this stage.

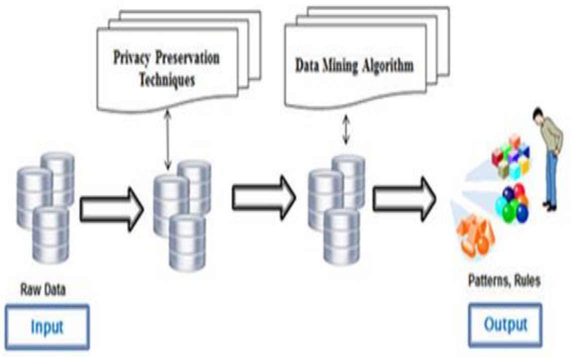

Fig. 1 PPDM Architecture [31]

A. PPDM Techniques

There are a number of techniques are present to preserve the privacy. According to Alpa Shah, the vital classification of PPDM is based on Anonymization, Perturbation, Cryptography, Fuzzy and Neural Networks [5]. These are classified as in this figure2.

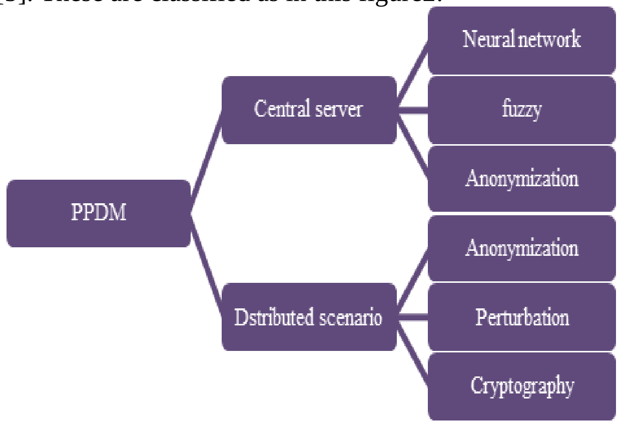

Fig. 2 Classification of PPDM

a) Centralized/Data Publishing Scenario: Centralized scenario is also known as the data publishing scenario. The data can be published publically in its original form. Even though it is also possible that there is no encryption is done in the format. Some types of alterations have to be applied before disclosing the data to maintain the privacy of data of individuals. The techniques used in this scenario like Neural Network based, Fuzzy based and Anonymization based are discussed here. Hayden Wimmer and Loreen Powell [33] provide the framework of PPDM using neural network. They discussed the different PPDM techniques and their effects of on machine learning algorithms. Firstly, they read a data set file and provided the classification using machine learning algorithm. Secondly, they read the same data file and they applied a privacy structure on it and applied the same machine learning algorithm on it. K-anonymization technique of PPDM is employed for the study. The general framework is detailed in Fig.3 described the mentioned framework. In their given work, they used k-anonymity as the PPDM technique

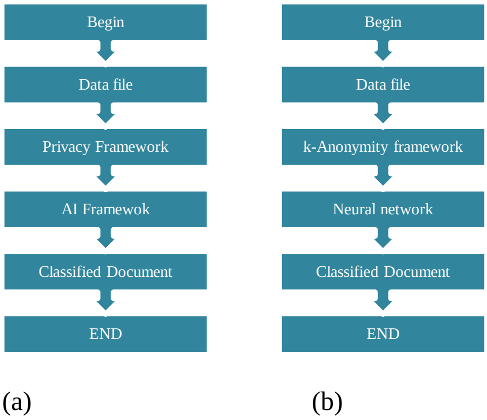

Fig. 3 (a) Machine Learning & Framework of PPDM, (b) ANN Flow

Fuzzy based privacy preserving data mining is well demonstrates by Stanley [18]. He discussed different fuzzy-based mapping techniques in his paper. He also compared various mapping approaches are in terms of PPDM property and checks their potential to maintain the same relationship.

b) Distributed Privacy Preserving Data Mining:

Privacy preserving data mining permits the distribution of personal and private data for the purpose of analysis. Recently there are several techniques and methods have been elaborated for the same. In most of the techniques some type of alteration is performed to the data to provide the protection to the privacy. Three approaches are used in distributed privacy preserving data mining. These are Perturbation model, Cryptographic protocols and Anonymization. Perturbation is also a most popularly used technique in PPDM and it is basically used for electronic health record (EHR). In data perturbation the data is distorting using the noise. Both additive noise and multiplicative noise can be used for the purpose of distortion. The major challenge that is faced in data perturbation is that the balancing the ratio of protection of privacy and the quality of data. Both of these factors are k contradictive factors. D.Kavitha explained various types of Data perturbation approach. The author classified it into the different categories [34]. This is shown in the following figure:

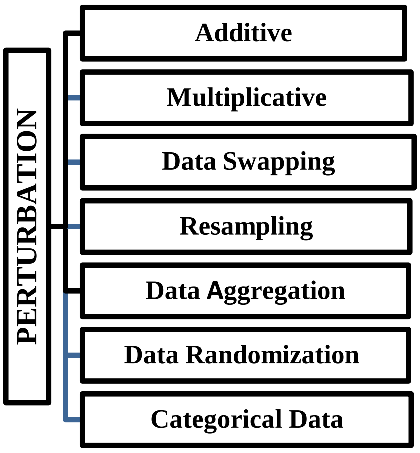

Fig.4: Perturbation Techniques

Data perturbation can be broadly classified in two categories. These are: 1) Probability distribution approach, 2) Value distribution approach. Perturbation approaches can be of three types [13]. These are shown in the following figure

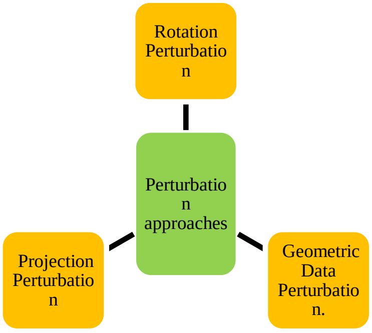

Fig 5: Perturbation approaches

So as shown in the above Figure, data perturbation can be done by several techniques. In data Swapping technique, the values of different records are swapped with each other to perform the privacy preserving. The major benefit of data swapping approach is that only lower order data can be preserved. In randomization, the data is shuffled vertically. The position of the selected record is changed. This will help to hide the exact identification of the record. In Anonymization, privately identifiable data is preserve by generalizing the attributes. K-anonymity is an anonymization method that aims to hide the particular record among a group of records. There are some techniques used in anonymization are listed below in the figure

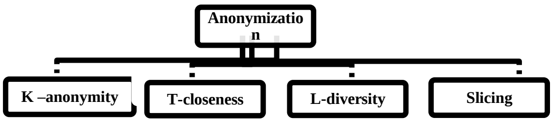

Fig.6: Approaches of Anonymization

K –anonymity is used to apply anonymization method. When anonymization method is used for privacy preserving then it is ensured that the transformed data is true but there is some information loss. P.AnnanNaidu1 [17] stated that in Anonymization technique the original data sets altered using k-anonymous. It is with respect to quasi-identifier attributes. Different approaches are shown in figure 6. The following tables represent the original data set and k-anonymous data.

TABLE II ORIGINAL DATA SET

Emp id	Name	Age	Salary
1201	Jonh	35	40000
1202	Carry	48	45000
1203	Pear	32	35000
1204	Daisy	40	42000

TABLE III K-ANONYMOUS DATA

Emp id	Name	Age	Salary
1#	Jonh	3#	40000
1#	Carry	4#	45000
1#	Pear	3#	35000
1#	Daisy	4#	42000

L-diversity is a technique that can be solution for attribute level privacy problem faced with K-anonymity. In T-closeness, the distribution of a sensitive attributes in any particular class should be close to the distribution of the attributes in the overall table. It should not be more than t threshold. This technique is not effective whenever there is increasing in dimension.

Cryptography is one of the methods that is commonly used to preserve the personal data. In cryptographic technique main focus is on the safety of sensitive attributes. This concept was proposed by authors in [35]. According to them the individual's privacy may be broken at the final stage of data mining. Different types of encryption can apply to preserve the identity and the sensitive information of the records. A number of protocols are used to achieve the security in Cryptography. These are shown in the below Fig.7 [5] [34]. As the authors described about the various protocols like Two party case, Multiparty case, oblivious transfer and oblivious polynomial evaluation.

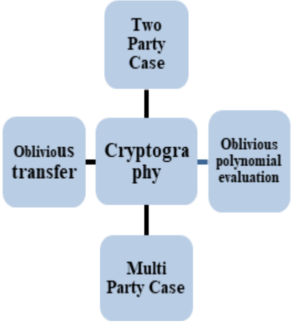

Fig.7 Cryptographic Methods

All the methods and techniques discussed above are successfully proved to achieve the privacy to the certain level but they suffer with some limitations. Data mining algorithms and techniques can be useful if these are efficient enough to achieve the goal of gaining trusts by preserving the privacy of a person.

IV. REVIEW OF LITERATURE BASED ON EXISTING PPDM IN TERMS OF TECHNIQUES & APPROACH

Nowadays there is a need to develop a powerful and strong model that is scalable meet different issues regarding PPDM. Regarding this, the paper recognised the literature gaps of various literatures and analysed them for further more enhancements, significant robust privacy protection, and preservation. There are a number of techniques of preserving the privacy are used. These are based on the requirements in real-world data processing. The following literature is carried out in the topic mentioned. Currently there are a number of data mining techniques that are used to protect the privacy. These techniques are used in PPDM for different research purpose. Intensive research findings revealed that the existing privacy preserving algorithms and approaches are still suffered from incompleteness. The following table 4 outline the different techniques used in PPDM and relevant research gap of current studies.

.

TABLE IV LITERATURE REVIEW AND RESEARCH GAP OF VARIOUS STUDIES

•Author(s)/Year	Techniques	Datasets	Performance	Research Gaps
S.Kaliappan[27]	Ant Colony Optimization, Random Rotation Perturbation, K-means clustering	Real-time healthcare	The proposed algorithm is better in utilization and a combination of K-means clustering algorithm. This method protects the privacy of individual more accurately	Other data hiding techniques can be explored. Furthermore, by choosing suitable QI attributes.
Shengyao Zhou, et al. [37]	MEDICAL AGGLOMERATION BEHAVIORS MINING (MABM)	Medical insurance industry.	MABM algorithm has better scalability, more efficient in running parameter than Eclat algorithm and Apriori algorithm.	A threshold should be determined. And after then the fraud of the person's card can be judged,

Akash Siddhpura Prof. Daxa V. Vekariya [19]	PPDM hybrid algorithm(cryptography and perturbation)	archive.ics.uci .edu (Machine Learning, UCI) (Indian Liver Patient Dataset , Balance Scale Dataset , Ablone Dataset , and Bank Marketing Dataset)	In this, the data is converted in to their respective asci values then apply perturbation techniques and after then cryptography technique was applied on the data.	We can also work with the video and audio data.
Carson K. Leung*, Calvin S. H. Hoi [12]	privacy-preserving item-centric mining algorithm PP-UV-Eclat	--	It is used in the Apache Spark environment to find frequent patterns.	PPDM mining can be improved and publishing step can be shifted from the post processing step to intermediate processing step
Arshveer Kaur[10]	Privacy Preserving Hybrid Technique(Suppression and Perturbation)	Database stored on Ms Access	Information loss is Zero, Execution Time are minimum and Privacy Preserved is maximum	--
Surbhi Sharma, Deepak Shukla[38]	Multi-party privacy preserving data mining for vertically partitioned data	Student data	Performance in the terms of accuracy is high, error rate is low, time consumption high and memory consumption is low as compared to c 4.5	In this algorithm complexity time consumption should be measured as low
David Kenyon , J.H.P Elof [21]	Privacy Preservation method of prediction	Sample of short-term insurance claims	R, Hadoop and Hive are used to test the insurance claims fraud and it is done with help of different PPDM.	--
A. Sheshasayee, S,SusanThomas [16]	Data mining tools which are efficient in detecting upcoding frauds	Healthcare	Efficient Review	semi-supervised method of learning would be highly appreciable in fraud detection

V. COMPARISON OF PPDM TECHNIQUES

Privacy preserving has become crucial in data mining. Today there present a number of various techniques, algorithms and approaches to preserve the privacy. Some of them are discussed in this paper. Table 5 presents a comparison of various PPDM techniques. It also illustrates the methods that are employed by different techniques. The table summarizes the discussion in previous sections on PPDM techniques

TABLE V: COMPARISON OF PPDM TECHNIQUES

S. No.	Techniques	Methods Employed	Scenarios
1	**Anonymization**	Generalization Suppression, Permutation	Central Commodity
2	**Condensation**	Aggregation	Central Commodity
3	**Perturbation**	Adding Noise, Data Swapping,	Central Commodity and Distributed

4	**Randomization**	Adding Noise, Scrambling, Resampling	Central Commodity and Distributed
5	**Neural Network Based PPDM**	Probabilistic NN, Bayesian network, Kohem SOMs	Central Commodity
6	**Fuzzy** **Based PPDM**	K-means clustering algorithm, Fuzzy classifiers, Apriori Algorithm	Central Commodity
7	**Cryptography**	Secure Multiparty Computation, Oblivious Transfer, Digital Envelope.	Distributed

VI. CONCLUSION

The primary aim of the study is to provide the current PPDM techniques used by the researchers in order to get better understanding of existing techniques. An overall overview on privacy preserving techniques based on perturbation, neural network, randomization, cryptography and k-anonymization is presented here. Nowadays PPDM is appeared common because of sharing of private sensitive data. There are some advantages and some disadvantages of recent studies are highlighted in this paper. Currently, there is sharing of Big Data across the various sectors such as insurance, health ,social sites and others. Thus, there is a need of preservation of privacy. The main objective of PPDM is developing algorithm to conceal the individual privacy of sensitive data. In data mining there exist a lot of techniques of preserving the privacy of data. The authors in the paper have tried to classify the PPDM techniques available in the literature with their research gaps. Different techniques are explored which can be still further enhanced to provide better results.

REFERENCES

[1] Mohammed GolamKaosar, Russell Paulet, XunYi, Fully homomorphic encryption based two-party association rule mining, Data & Knowledge Engineering. (2012) 76–78.

[2] W. Lin1 et al., An Ensemble Random Forest Algorithm for Insurance Big Data Analysis, Advances In Computational Intelligence Paradigms For Security And Privacy For Fog And Mobile Edge Computing, 5 (2017) 16568-16575.

[3] Jerry Chun-Wei Lin, PPSF: An Open-Source Privacy-Preserving and Security Mining Framework, IEEE International Conference on Data Mining Workshops (ICDMW). (2018) 1459-1463.

[4] Hemlata, Preeti Gulia, Techniques and Algorithms of PPDM, IJSRD - International Journal for Scientific Research & Development| 3(4) (2015) 3484-3487 ISSN (online): 2321-0613.

[5] Alpa Shah and Ravi Gulati, Privacy Preserving Data Mining: Techniques, Classification and Implications - A Survey, International Journal of Computer Applications (0975 – 8887) 137(12) (2016) 40-46.

[6] K.Naga Prasanthi, A Review on Privacy Preserving Data Mining Techniques, International Journal of Advanced Research in Computer Science and Software Engineering. 6(3) (2016) 35-40.

[7] Charu C.Aggarwal, Applications of Frequent Pattern Mining", from book, " Frequent Pattern Mining, ISBN 978-3-319-07821-2 (eBook) Springer Cham Heidelberg NewYork Dordrecht London, chapter no, 18 (2014) 443-461.

[8] Lei Xu, et al, Information Security in Big Data: Privacy and Data Mining, IEEE Access, The Journal for rapid open source publishing. 2 (2014) 1149-1175.

[9] P.Usha, Shriram, R., & Sathishkumar, S. Sensitive attribute based nonhomogeneous anonymization for privacy preserving data mining. In Information Communication and Embedded Systems (ICICES). International Conference on , (2014) 1-5..

[10] Kaur A., Hybrid Approach of Privacy Preserving Data Mining using Suppression and Perturbation Techniques, International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), (2017).

[11] Vadlana Baby, Dr. N. Subhash Chandra, Distributed threshold k-means clustering for privacy preserving data mining, Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE, (2016) 2286-2289.

[12] Carson K. Leung, Wodi, Privacy-Preserving Frequent Pattern Mining from Big Uncertain Data, IEEE International Conference on Big Data (Big Data), (2018) 5101-5110.

[13] C. Perera, et al, Big Data Privacy in the Internet of Things Era, IT Pro,. (2015) 32-39.

[14] X.Zhang, & Bi, H. Research on privacy preserving classification data mining based on random perturbation. Information Networking and Automation (ICINA),IEEE International Conference on (1) (2010) 171-173.

[15] Jawwad A. Shamsi ,Muhammad Ali Khojaye, Understanding Priva, cy Violations in Big Data Systems, IT Professional Published by the IEEE Computer Society. (2018) 73-81.

[16] Ananthi Sheshasayee, Surya Susan Thomas, Implementation of Data Mining Techniques in Upcoding Fraud Detection in the Monetary Domains, International Conference on 20 Innovative Mechanisms for Industry Applications (ICIMIA). (2017) 730-734.

[17] P.AnnanNaidu, M.Vamsi Krishna, Comprehensive Review on Privacy Preserving Data Mining Techniques and Methods, International Journal of Engineering and Management Research. 7(1) (2017) 121-126.

[18] Samir Patil, Gargi Shah, Aniket Patel, Techniques of Data Perturbation for Privacy Preserving Data Mining International Journal of Advent Research in Computer & Electronics (IJARCE) (1) (2014).

[19] Akash Siddhpura, V. Vekariya, An approach of Privacy Preserving Data mining using Perturbation & Cryptography Technique, International Journal on Future Revolution in Computer Science & Communication Engineering. ISSN: 2454-4248. (4) (2018) 255 – 259.

[20] A.Thomas, J. Rana, A Review on privacy preserving data mining approaches, National Conference on Recent Research in Engineering and Technology (NCRRET) (2015).

[21] Yousra Abdul Alsahib S. Aldeen, Mazleena Salleh and Mohammad Abdur Razzaque, A comprehensive review on privacy preserving data mining, Springer Plus a Springer open journal, (2015) 1-36.

[22] Vipula Rawte, Fraud Detection in Health Insurance using Data Mining Techniques, International Conference on Communication, Information & Computing Technology (ICCICT) (2015) 45-54.

[23] Detail of ICD-10 Available at: https://en.wikipedia.org/wiki/ICD-10

[24] ICD-Classification is Available at http://www.who.int/classifications/icd/en

[25] L.Cranor, T.Rabin, V.Shmatikov, S.Vadhan and D.Weitzner, Towards a privacy research roadmap for the computing community, Comput. Commun. Consortium Committee, Comput. Res. Association, Washington, DC, USA, White Paper, 2015.

[26] Samir Patil, Gargi Shah, Aniket Patel, Techniques of Data Perturbation for Privacy Preserving Data Mining, International Journal of Advent Research in Computer & Electronics (IJARCE), 1(2) 2014.

[27] S. Kaliappan , A Hybrid Clustering Approach and Random Rotation Perturbation (RRP) for Privacy Preserving Data Mining, International Journal of Intelligent Engineering and Systems, 11 (6)(2018) 167-176.

[28] David F. Nettleton et al. Privacy in Multiple On-line Social Networks - Re-identification and Predictability, Transactions On Data Privacy 12 (2019) 29–56.

[29] T. A. Adesuyi, B. Man Kim, A layer-wise Perturbation based Privacy Preserving Deep Neural Networks, CAIIC(IEEE), (2019) 389-394.

[30] S. Scardapane, et al., Privacy-Preserving Data Mining for Distributed Medical Scenarios, Chapter in 'Smart Innovation Systems and Technologies. (2018) 119-128.

[31] V. Jane Varamani Sulekha, Dr. G. Arumugam, A survey on Microaggregation based Privacy Preserving Data Mining Techniques, International Journal of Scientific Research Engineering & Technology (IJSRET), ISSN 2278 – 0882 Volume 7(4) (2018) 268-279.

[32] P.AnnanNaidu, M. Vamsi Krishna, Comprehensive Review on Privacy Preserving Data Mining Techniques and Methods, International Journal of Engineering and Management Research, ISSN (ONLINE): 2250-0758, ISSN (PRINT): 2394-6962, 7(1) (2017) 121-126.

[33] Hayden Wimmer ,Loreen Powell, A Comparison of the Effects of K-Anonymity on Machine Learning Algorithms, (IJACSA) International Journal of Advanced Computer Science and Applications, 5(11) (2014) 155-160.

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

[35] Data Perturbation and Features Selection in Preserving Privacy. Available at: http://ieeexplore.ieee.org/document/6335531/ Date Accessed: 20/09/2012.

[36] Tosin A. Adesuyi1 , Byeong Man Kim2, A layer-wise Perturbation based Privacy Preserving Deep Neural Networks, IEEE, ICAIIC, (2019) 389-394.

[37] Shengyao Zhou, et al., A Novel Method for Mining Abnormal Behaviors in Social Medical Insurance, ISDN 978-1-5386-7266, IEEE, (2018) 744-748.

[38] Surbhi Sharma and Deepak Shukla, Efficient multi-party privacy preserving data mining for vertically partitioned data, Inventive Computation Technologies (ICICT), .1109/INVENTIVE.2016.7824852, © 2017 IEEE,(2017).