A hybrid decision tree model for high dimensional privacy preserving process

Aaluri Seenu[1], Dr. G Samba Siva Rao[2]
[1]Research Scholar, [2]Professor
[1,2] Department of CSE, Acharya Nagarjuna University (University College of Engineering and Technology) Nagarjuna Nagar-522510, Guntur, A.P, India
[1]Email: cnuaaluri@gmail.com

ABSTRACT
The data mining system can help to identify the important hidden patterns for decision-making in large datasets. Privacy preserving data mining (PDDM) has emerged as a critical area for the sharing, decision-making and dissemination of confidential data. Preserving privacy is a common data security model for protecting unauthorised access to individual decision patterns. Because the distributed data of the individuals is processed by the third Party, the information in digital networks is misused. Such information on privacy about businesses , industries and persons must be encoded prior to publication or published. During the processing of data from various sources, decision patterns based on standard data security protection models such as Naïve Bayes, SVM and the models for the decision tree are very difficult to maintain. In addition, the use of traditional models to fill sparse values is inefficient and inadequate for the protection of privacy. A novel data security model was developed and applied on large data sets in this paper. In this model a philtre based data protection scheme is used to cover decision patterns with homomorphic encoding and decryption algorithm using the decision-tabo classifier. In this scheme. Experimental findings demonstrated the high processing efficiency of the proposed model relative to conventional data protection approaches of large-scale datasets.

1.Introduction

The Data Mining Privacy-Preservation framework (PPDM) combines data mining principles with the possible security of privacy[1]. Stable Multi-Party Computing (SMC) is entirely the responsibility of PPDM. This technique is widely used in the field of the defence of privacy. PPDM is in charge of expanding protection to sensitive data mining information. Corporate companies are worried about their confidential intruders' data disclosure. PPDM is therefore very important to be applied in different data mining applications. Over decades , a large number of research projects in the field of PPDM have been carried out and many researchers have also proposed several algorithms. Today, a safe channel is very important to build to exchange complex data. -- transaction is saved in the database during safe sharing. Either terabytes or petabytes are the data sets resulting. The performance of mining algorithms shall be determined by productivity, scalability and confidentiality measures. In order to mine vast amounts of information[1], several methods for the analysis of privacy are deployed in various data mining algorithms. Maintaining privacy in the field of data mining is very important and important.

The data mining method can be described as the data extraction tools that turn enormous data volumes into interesting information. To extract valid and functional rules, decision bodies or clusters from large datasets, extracting algorithms are created. It also offers a way to find fascinating trends and their connexions between various datasets. An expanded protection framework for sensitive information as well as sensitive information for users is therefore necessary. Many algorithms have been suggested for years in order to address these privacy problems in the area of data mining. In order , for example, to mask sensitive data by discarding stored information or random noise input, several sanitization approaches are created. The primary objective of privacy protection of classification techniques is to create a classificator that can accurately predict confidential information. Moreover, it is responsible for distorting critical attributes and retaining fundamental characteristics of clustering that privacy maintains clustering approaches.

In three steps, all confidentiality algorithms can be performed:- interruption, encoding or restricted queries. Data disturbances include data exchange processes, noise addition and the initial signal disturbance. Various encryption algorithms are developed with the intention of encoding the message like SMC. The limitation query is described as preventing raw data mining. Some common examples of restriction inquiries include:- delete, generalise, analyse, and divide the data. All approaches previously developed emphasise that private information must be maintained during the entire data mining process.

Data-mining approach to the protection of privacy can be widely categorised in 2 sub-classes, as mentioned below [2-5].

1. The initial phase of the PPDM strategy focuses on confidential information security. For example, names, IDs, addresses and other private information are:.

2. The second step begins with the information concealed inside the database after the completion of the initial step of PPDM. It shall incorporate safety measures to protect confidential information. This step has the primary objective to evaluate the exact information secret mechanism without impacting non-sensitive and useful laws.

These methods are responsible for manipulating data with complex associations to safeguard private data. The accuracy of the results is therefore responsible for determining the efficiency of the algorithm. The quality assessment depends also on the expected use of data instances. As the data volume grows, various current methods are made difficult due to the high cost of computing. The main aim of the SDC is to secure any single instance of data in the training data.

The security of communications and encryption are two key issues with the distributed data mining approach. Divided applications also incorporate privacy protection along with encryption schemes. The distributed application is able to store data using two models-vertically partitioned and horizontally divided. Vertically separated data can be described as data collected on different sites, and the data stored is segregated so that no overlap of data is observed. Horizontally partitioned data is defined as information stored on the basis of records in numbers of locations. No information on history of other sites is given on the pages. A consideration of the horizontally partitioned data[6,7] produces several useful and true association laws. A great deal of research in the field of cryptography has been developed to address privacy issues. Take the Stable

Multi-Party Computing (SMC) example for consideration. In the multi-user networks where each user is responsible for collaborating with other users to perform computer tasks and to safeguard privacy[8][9], SMC will always be introduced.

Present research methods include numerous techniques of privacy protection and process them in order to discover decisions using data mining models. In order to obtain useful expertise in confidential information the key aim of the privacy protection system is to build a data mining model. The above-mentioned PPDM method involves two major problems:

1. The most frequent problem is to protect sensitive information including name, ID number , address and income level.
2. The second issue concerns the protection of confidential information through the implementation of knowledge communication in a database which is often called a database of knowledge (KD hidden).

The key constraints of conventional models of privacy are:

1. Loss of information: If the original information has been saved, the data recovery process will be considered to be loss of information.
2. Knowledge is obscured or distorted here. Confidentiality preserved: The algorithm would be considered effective if the PPDM results in a higher value for privacy protection.
3. Computing time: computing time is an important factor in assessing the success of the strategy for privacy protection. If and only with less computing time, a data conservation approach is successful. In other words, reduced computing time would make privacy protection strategies more efficient and reliable.
4. Complexity: The PPDM improves reliability and performance when applied to large data sets with a minimum complexity.
5. Dependence on data size: The efficacy of the algorithm decreases by many days with the exponential growth of data.

2. RELATED WORK

[7] developed a simple algorithm to build up a powerful decision-tab classification known as the ID3 algorithm. The database administrator maintains trends to inhibit values on the class mark in the proposed data mining privacy security system. The information is reduced from high (security) to low (public), and inferences in this mining model are also taken into account. The possible downgraded information in the above-mentioned system is combined with the not initialised costing method. The data mining platform also strengthens the confidentiality mechanism. For horizontally partitioned data[8], the exchange of attributes between the parties includes all pre-existing grading techniques. In the case of vertically divided data, conventional approaches must exchange information from the communicating parties and each instance. In the area of Secure Multiparty Computing, several cryptographic approaches to accurate privacy and security are implemented. This method is an extension of research in Lindell[9] involving the combination of rules of mining, classification systems[10], and clustering approaches. There are few efficient approaches to privacy, while most algorithms are unable to preserve privacy and limit the disclosure of their information.

Models for privacy protection are the main applications of Tree Classification Decision: radar signal detection, character recognition, remote sensing, medical diagnosis, expert systems, speech recognition etc. Many extended models to increase the range of features of the present model decision tree have been introduced.

Decision trees are usually seen as graphs of nodes and edges. The root node and the intermediate nodes represent tests, while the outgoing edges show the results of the tests. These intermediate nodes were called inner nodes by researchers. In addition, the leaf nodes represent different labels of class that define patterns of privacy in large databases.

In [11] they evaluated the conventional data protection strategy and suggested an innovative alternative on small datasets. The standard privacy algorithm, which includes horizontally partitioned databases, was introduced. [7] implemented the new ID3 algorithm privacy approach by taking horizontal partitioning data into consideration. Instead of considering two parties such as vertically partitioned data methods, they considered multiple parties. Instead of entropy for the ID3 algorithms, Gini feature selection steps are implemented to create decision tree patterns. Both parties may assess each other's attributes gain value. This method works well if the database is split into two or more groups. Entropy helps to create healthy trees in general.

In certain cases, statistical interventions such as the rules of association and classification of data mining techniques are unsuccessful in the application of data security. In this case, association rules are often concealed in order to protect the privacy of data mining. A new PPM is created to protect the privacy of a respondent, known as the Randomized Response Scheme[8]. It follows the underlying principle of decomposition, in which the data owner separates and store the original data in various locations. It provides a hindrance to the attacker's original data and often often contains false information to deter attacks. The new randomised response method for various data mining computations has been developed [6]. This data mining approach preserves the privacy of the respondent by random queries. The data mining approach Reasonable distributed probability is calculated by the proposed approach and distributed probability for certain characteristics also evaluated. [9] has created a data hiding method focused on randomised response strategies to the law of the privacy protection association.

[10] has introduced the so-called Randomized Partial Hiding Response (RRPH) technique of extended pre-treating data. They have also developed an advanced version of the RPH Rule for the security of privacy. The processing of categorical data is based on this type of technique.

Disruption of data

In order to protect the protection of sensitive data and to perform data analysis , data disturbance techniques are used. The original data can not be recreated so it is only possible to restore data distributions. Stanley R. M. Oliveira [5] proposed to safeguard the privacy of clustering, which requires centralised delivery and rotational transformation technology. This method was applied in a medical dataset where the data server is unable to learn or retrieve important data. This is one of the core limits of disruption approaches. All conventional approaches to data security can not be applied to massive data sets. Until dissemination of data in multi-cloud servers, it is very important to incorporate privacy controls. The most positive

aspect of all approaches to data security is that confidential consumer information is correctly detected before sharing with others.

Nowadays, the most common solution to privacy differential protection is to provide improved safety in statistical databases. It reduces the chance to classify information for privacy enhancement. In comparison, the data protection mechanism helps safeguard data against unauthorised access when interacting through a network. If an approved and certified user receives the data, no additional restraint from gathering sensitive information is placed on that user. Many studies have therefore been attempted to identify the connexion between data safety and data privacy.

[10], different approaches to data mining protection in privacy have been studied and analysed. Implementing certain approaches can limit certain basic data mining functions. They created a new way to safeguard secrecy and referred to it as a Record Protection Link (PPRL). This technique makes multiple database connexions to organisations along with sensitive data privacy.

Traditional methods to data mining to maintain confidentiality. Such approaches rely fundamentally on the distribution, distortion, mining and data hiding. [10] A novel model for the safeguard of privacy using the conventional digital differential method for homomorphic encryption was introduced. This model is based on the communication protocol for multiparties. This protocol also relies on homomorphic encoding to ensure that the privacy of the decision tree building is maintained. [11] proposes a new approach that considers k to be the closest neighbour classification system. It depends on the closest approach to the neighbour to address the problem in the following two steps:

1. The closest neighbour to the security of privacy classification is initially chosen. In terms of precision, performance and privacy, the algorithm is balanced. In order to achieve optimisation, the method is multi-use.
2. The most widely used classification mining methods are vertically distributed data classification and decision tree classification mining. They normally use the template ID3 and C4.5.

On small dimensional datasets, a new technique called the lazy decision tree algorithm was introduced. This includes a score feature, lazy learning and a data classification decision-making tree. The decision trends in the decision tree are expected. Generalization or deletion methods enforce the privacy anonymity, which can contribute to tremendous loss of original details. The loss of knowledge also increases with the increase of k. In the multidimensional dataset k-anonymity strategy. The record can not be distinguished from at least k-1 records because information about k-1 records are also available in the dataset. For generalisation and containment, K-anonymity technique is usually introduced. The approach for preserving privacy of high-dimensional data due to homogeneity attacks is not applicable.

The issue occurs in the estimation of the help value in the case of vertically partitioned data mining approaches. The implementation of secure scalar product or safe set intersection will address this problem[12]. In the case of multi-classification mining technologies[13], an extended safety classification scheme was developed, based on vertically separated data and the safe scalar product. [14] used the horizontally partitioned data horizontal encryption method by a special form of decision tree. The data server is unable to learn or retrieve valuable records

in the disturbance technology. The approach can not restore the original data, but it can only recreate its distributions. [15] a Random Decision Trees approach (RDT) on small dimensional data sets has been established. The trees are used to achieve a paradigm of privacy protection using association techniques. In this segment, a difference in mean zero and threshold is used in randomised noises. In the evaluation of data mining models, they attempted to suggest a new encryption method with the goal of protecting privacy.

[16] A new method for privacy protection using symmetrical key encryption has been introduced. In this case, symmetrical keys contain the data. Symmetric keys from the Key Distribution Centre (KDC), are created from the trusted. After the encryption the intruder finds it very difficult to guess the original data. Encrypted information is absolutely secure until the keys are encrypted. For large datasets, this approach is unsuccessful.

3.PROPOSED METHODLOGY

In this approach, the issue of sparsity values and high dimensionality data protection issues was tackled using a novel privacy-preserving decision tree classifier. The main aim of this model is the use of the protected model of privacy protection to protect the decision pattern in distributed environment in the supervised decision tree classification. The model proposed is generally aimed at improving privacy in data mining privacy and at predicting correct decision-making rules on high dimensional data. The model implementation increases the security level in terms of true positives, false positives and accuracy in comparison with conventional data protection based data mining models, such as C4.5 and ID3 models.

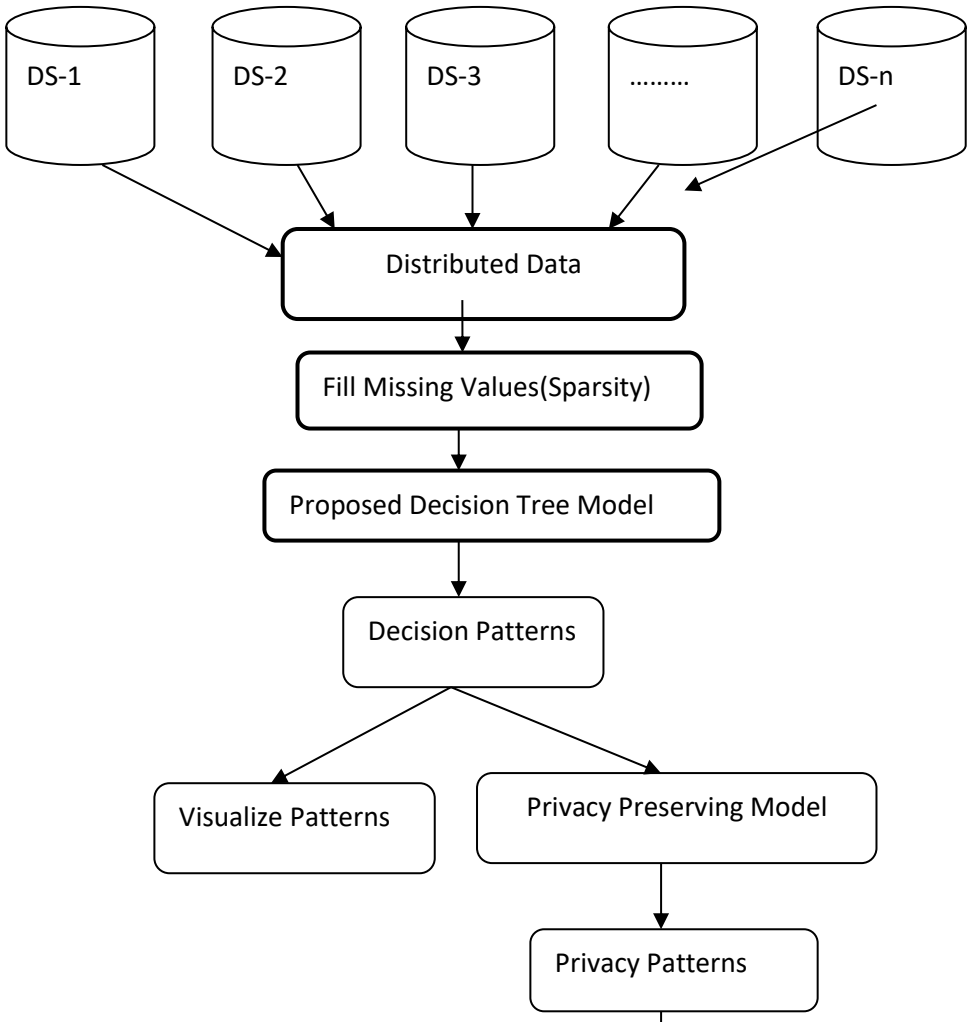

Figure 1:ProposedPrivacy Preserving Framework

Figure 1 illustrates the use of a high-dimensional private data collection privacy protection model. At first, a central database for privacy protection is generated from multiple sources. As a pre-processing input for data the proposed model considers high-dimensional distributed data. The proposed model was implemented in two main stages, firstly through data filtering, and secondly through the use of decision tree approaches, through the protection of the privacy model.

Privacy Preserving on Decision Patterns:

The homomorphic approaches are developed for privacy preservation of users through hiding some query information of the data mining models. The updated version of homomorphic encryption technique includes a searching strategy for encrypted data. Hence, it is implemented in privacy preservation applications in the fields like finance, biomedicine and military database. Achieving the same level of security and performance along with other conventional techniques are more costly in terms of computational cost, storage cost and communication overheads. Homomorphic encryption checks the data confidentiality in order to resolve the security issue of storage or processing data by an untrusted third party.

Encryption Process:

Additive Homomorphic Encryption

$$EncD(\mathrm{I}_1 + I_2) = EncD(\mathrm{I}_1) + EncD(\mathrm{I}_2);$$

Multiplicative Homomorphic Encryption

$EncD(I_1 . I_2) = Enc(I_1).Enc(I_2)$;

$\psi(i)_0$ =EncD(I_1):=EncD(I_1)=($I_1 + \gamma * \beta$)mod n where n= $\alpha * \beta$;

$\psi(i)_0^{'}$ =EncD(I_2):=Enc(I_2)=($I_2 + \gamma * \beta$)mod n where n= $\alpha * \beta$;

$EncD(I_1 + I_2)$:= Enc($\psi(i)_0$)+Enc($\psi(i)_0^{'}$);

$EncD(I_1 + I_2)$:=($I_1 + \gamma * \beta$)mod(n)+($I_2 + \gamma * \beta$)mod n

EncD($I_1 . I_2$):=EncD($\psi(i)_0$).EncD($\psi(i)_0^{'}$);

:=($I_1 + \gamma * \beta$)mod n. +($I_2 + \gamma * \beta$)mod n;

Where α , β , γ are the random numbers taken from cyclic group generator from G.

Decryption Process:

$EncD(I_1 + I_2)$= Enc($\psi(i)_0 + \psi(i)_0^{'}$)=Enc($\psi(i)_0$)+Enc($\psi(i)_0^{'}$);

:=($\psi(i)_0 + \gamma * \beta$)mod n+($\psi(i)_0^{'} + \gamma * \beta$)mod n

EncD($I_1 . I_2$):=Enc($\psi(i)_0 . \psi(i)_0^{'}$)

:=Enc($\psi(i)_0$).Enc($\psi(i)_0^{'}$);

:=($\psi(i)_0 + \gamma * \beta$)mod n. +($\psi(i)_0 + \gamma * \beta$)mod n;

Dec($EncD(I_1 + I_2)$):=(EncD($\psi(i)_0 + \psi(i)_0^{'}$)) mod α

:= (($\psi(i)_0 + \gamma * \beta$)mod n+($\psi(i)_0^{'} + \gamma * \beta$)mod n) mod α

:= $I_1 + I_2$ ----- > (1)

Dec(EncD($I_1 . I_2$)):=(EncD($\psi(i)_0 . \psi(i)_0^{'}$):=EncD($\psi(i)_0$).Enc($\psi(i)_0^{'}$))mod α ;

:=(($\psi(i)_0 + \gamma * \beta$)mod n. +($\psi(i)_0^{'} + \gamma * \beta$)mod n) mod α ;

:= $I_1 . I_2$ --------- > (2)

Solving eq-(1) and eq (2) we will get decrypted values of I_1 , I_2 .

4.E**xperimental Results**

Intel core i5 2.4GHz with minimum 2 GB of RAM has implemented our proposed model. For the proposed model to protect privacy, Jama, JUnit third-party libraries were used. For testing purposes, we used the complete KDD99 and 10 fold cross validations. With 10-fold cross validation, whole training data is divided randomly into 10 equal subparts. The proposed model accuracy can be tested using these test results.

Definition of the dataset

Tables 1 and 2 define the diabetes and KDD data sets and their classification attributes, size and number. In 1999, it accepted and approved DARPA data, which is available in http:/www.kdd.ics.uci.edu / databases / kddcup99, as the traditional ISD-Benchmark database called KDDCup99. In data KDD99, different attribute values are associated with named attack names in each instance. These labels are categorised in five types: ordinary, sample, DOS attack, U2R, and R2L. A full KDD99 dataset includes nearly 4 million instances with 41 features, which are broken down into 22 forms of attacks and summed up in Table 1.

Table1:Detailsoflabelledattacks

AttackCategory	**AttackName**
DenialofService	Smurf,Neptune,back,land,pod,teardrop
Probe	IPsweep,satan,portsweep,satan
R2L	Ftpwrite,Imap,Guesspassword,warezclient
U2R	Perl,Bufferoverlow,Rootkit

Correctly Classified Instances 5087 96.1444 %
Incorrectly Classified Instances 204 3.8556 %
=== Confusion Matrix ===

```
   a    b    c    d    e    f    g   <-- classified as
2690   26    1    1    0    0    0 |    a = normal
  32 2397    0    0    0    0    0 |    b = DDOS
  10   19    0    0    0    0    0 |    c = DOS
  19   22    0    0    0    0    0 |    d = teardrop
  17    6    0    0    0    0    0 |    e = back
  14   11    0    0    0    0    0 |    f = land
  17    9    0    0    0    0    0 |    g = smurf
```

Table 2: Performance analysis of Proposed model to the existing models on KDD Dataset

Model	**Patterns**	**ExecutionTime**
NaiveBayes+ABE	523	8245
SVM+ABE	312	8973
Logistic+ABE	353	8243
DecisionTree+ABE	643	7567

Proposed+Homomorphic	735	4673

Table 2 summarises the current application performance analysis for the existing KDD data set privacy protection application. From this table, we can note that the model proposed is less time to measure than conventional models.

Table 3: Performance of the proposed model to the existing model using homomorphic PPDM

Model	**Patterns**	**PrivacyTime**	**RecoverTime**
NaiveBayes+ABE	523	8245	7638
SVM+ABE	312	8973	7245
Logistic+ABE	353	8243	7284
DecisionTree+ABE	643	7567	6583
Proposed+Homomorphic	735	4673	5183

Table 2 summarises the current application performance analysis for the existing KDD data set privacy protection application. This table indicates that the new model is less sensitive and has less time to recover compared to conventional models.

5.CONCLUSION

As decision trends for the large-scale datasets increase, the need to use privacy models exponentially increases. In this article we used a new, high-dimensional dataset model focused on philtre data security. In this model a philtre based data protection scheme is used to cover decision patterns with homomorphic encoding and decryption algorithm using the decision-tabo classifier. In this scheme. Experimental findings demonstrated the high processing efficiency of the proposed model relative to conventional data protection approaches of large-scale datasets. Based on studies, the conventional efficiency of PPDM in large dimensional datasets such as KDD and diabetes has been improved by at least 10 percent.

REFERENCES

[1] Mahesh, R., &Meyyappan, T. (2013, February). Anonymization technique through record elimination to preserve privacy of published data. In Pattern Recognition, Informatics and Mobile Engineering (PRIME), 2013 International Conference on (pp. 328-332).

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

[3] Prakash, M., &Singaravel, G. (2015). An approach for prevention of privacy breach and information leakage in sensitive data mining. Computers & Electrical Engineering, 45, 134-140.

[4] J. Le Ny and G. Pappas, "Differentially private filtering," Automatic Control, IEEE Transactions on, vol. 59, no. 2, pp. 341–354, Feb 2014.

[5]M. Bae, K. Kim, and H. Kim, "Preserving privacy and efficiency in data communication and aggregation for AMI network," Journal of Network and Computer Applications, vol. 59, pp. 333–344, Jan. 2016, doi: 10.1016/j.jnca.2015.07.005.

[6]A. Barenghi, M. Beretta, A. Di Federico, and G. Pelosi, "A privacy-preserving encrypted OSN with stateless server interaction: The Snake design," Computers & Security, vol. 63, pp. 67–84, Nov. 2016, doi: 10.1016/j.cose.2016.09.005.

[7]T. P. Bhat, C. Karthik, and K. Chandrasekaran, "A Privacy Preserved Data Mining Approach Based on k-Partite Graph Theory," Procedia Computer Science, vol. 54, pp. 422–430, Jan. 2015, doi: 10.1016/j.procs.2015.06.049.

[8]D. Chandramohan, T. Vengattaraman, D. Rajaguru, and P. Dhavachelvan, "A new privacy preserving technique for cloud service user endorsement using multi-agents," Journal of King Saud University - Computer and Information Sciences, vol. 28, no. 1, pp. 37–54, Jan. 2016, doi: 10.1016/j.jksuci.2014.06.018.

[9]D. Chandramohan, D. Sathian, D. Rajaguru, T. Vengattaraman, and P. Dhavachelvan, "A multi-agent approach: To preserve user information privacy for a pervasive and ubiquitous environment," Egyptian Informatics Journal, vol. 16, no. 1, pp. 151–166, Mar. 2015, doi: 10.1016/j.eij.2015.02.002.

[10]G. Drosatos, P. S. Efraimidis, I. N. Athanasiadis, M. Stevens, and E. D'Hondt, "Privacy-preserving computation of participatory noise maps in the cloud," Journal of Systems and Software, vol. 92, pp. 170–183, Jun. 2014, doi: 10.1016/j.jss.2014.01.035.

[11]A. Fahad, Z. Tari, A. Almalawi, A. Goscinski, I. Khalil, and A. Mahmood, "PPFSCADA: Privacy preserving framework for SCADA data publishing," Future Generation Computer Systems, vol. 37, pp. 496–511, Jul. 2014, doi: 10.1016/j.future.2014.03.002.

[12]S. Fletcher and M. Z. Islam, "An anonymization technique using intersected decision trees," Journal of King Saud University - Computer and Information Sciences, vol. 27, no. 3, pp. 297–304, Jul. 2015, doi: 10.1016/j.jksuci.2014.06.015.