Privacy Preserving Classification of Clinical Data using Homomorphic Encryption

Lenat Sughirdha C[1]
M.E. Scholar, Dept. of Computer Science and Engineering
J.J. college of Engineering and Technology
Tiruchirappalli, India

Sumathi R[2]
Professor, Dept. of Computer Science and Engineering
J.J. college of Engineering and Technology
Tiruchirappalli, India

***Abstract*— In recent years, with the explosive development in Internet, data storage and data processing technologies, privacy preservation has been one of the greater concerns in data mining. Decision support systems (DSS) are defined as interactive computer-based systems intended to help decision makers to utilize data and models in order to identify problems, solve problems and make decisions. Clinical decision support system (CDSS) is designed to assist physicians and other health professionals with decision making tasks, such as determining diagnosis of patient data which improve effectiveness. Now-a-days remote outsourcing is employed to reduce the burden of clinical decision support in healthcare. The clinicians can use the health knowledge located in remote servers via the internet to diagnose their patients. But these servers are third party and therefore potentially not fully trusted, raise possible privacy concerns. Thus a privacy preserving protocol is built for diagnosis of patient data where both server and the client are unaware of the internal process. The protocol is built on the WBC dataset with homomorphic encryption as one of its building blocks.**

Keywords*— *Data mining; Classification; Privacy Preservation; Clinical Decision Support System; Secure two party computations; Support Vector Machine; Homomorphic Encryption.

I. INTRODUCTION

Data Mining is one of the most important and motivating area of research with the aim of finding meaningful information from huge data sets. In present era, Data Mining is becoming popular in healthcare field because there is a need of efficient analytical methodology for detecting unknown and valuable information in health data. In health industry, Data Mining provides several advantages such as detection of the fraud in health insurance, availability of medical solution to the patients at lower cost, detection of causes of diseases and identification of medical treatment methods. It also helps the healthcare researchers for making efficient healthcare policies, constructing drug recommendation systems, developing health profiles of individuals *etc*. However, it can also disclose sensitive information about individuals compromising the individual's right to privacy. Therefore, privacy preserving data mining has becoming an increasingly important field of research. The goal of privacy preserving data mining is to develop data mining methods without increasing the risk of misuse of the data used to generate those methods. A number of effective methods for privacy preserving data mining have been proposed. Most methods use some form of transformation on the original data in order to perform the privacy preservation. The transformed dataset is made available for mining and must meet privacy requirements without losing the benefit of mining. We classify them into randomization, anonymization and encryption methods [3]. Randomization method masks the values of the records by adding noise to the original data. The noise added is sufficiently large so that the individual values of the records can no longer be recovered [2]. The randomization method is more efficient. However, it results in high information loss. Anonymization method aims at making the individual record be indistinguishable among a group records by using techniques of generalization and suppression. The anonymization method can ensure that the transformed data is true, but it also results in information loss in some extent. Encryption method mainly resolves the problems that people jointly conduct mining tasks based on the private inputs they provide. These mining tasks could occur between mutual un-trusted parties or even between competitors therefore protecting privacy becomes a primary concern. The encryption method can ensure that the transformed data is exact and secure [2].

A clinical decision support system is a computerized medical diagnosis process for enhancing health - related decisions and actions with pertinent, organized healthcare knowledge and patient data to improve health and healthcare delivery [1]. CDSS are designed to assist physicians and other health professionals with decision making tasks, such as determining diagnosis of patient data which improve effectiveness. The recent advances in remote outsourcing techniques can be exploited in healthcare to reduce the burden of healthcare professionals and provide accurate decision support as a service. This service could be utilized by any clinician in a flexible manner such as on-demand or pay- per use [1]. Within this context, let us consider the following scenario: a third party server builds a clinical decision support system using the existing clinical dataset (i.e., assume that the server has a rich clinical dataset for a particular disease). Now clinicians, who want to verify whether their patients are affected by that particular disease, could send the patient data to the server via the Internet to perform diagnosis based on the healthcare knowledge at the server. This new notion overcomes the difficulties that would be faced by the clinicians, such as having to collect a large number of samples (i.e., a rich clinical dataset), and requiring high computational and storage resources to build their own decision support system. However, there is now a risk that the third party servers are potentially untrusted servers. Hence, releasing the patient data samples owned by the clinician or

revealing the decision to the untrusted server raises privacy concerns. This drawback can affect the adoption of outsourcing techniques in healthcare. Furthermore, the server may not wish to disclose the features of the clinical decision support system even if it offers the service to the clinicians. Hence a privacy preserving clinical decision support system is designed which preserves the privacy of the patient data, the decision and the server side clinical decision support system parameters, using the encryption method. Here we consider a decision support system that is developed using support vector machine (SVM), which is one of the machine learning tools. One of the building blocks of our technique is homomorphic encryption. Homomorphic encryption may be defined as a type of encryption which is processes with the cipher text. [1] For concreteness and without loss of generality, our descriptions are based on the Paillier cryptosystem. The Paillier cryptosystem is an additively homomorphic public-key encryption scheme, whose provable semantic security is based on the decisional composite residuosity problem.

The rest of this paper is organized as follows. We have presented the related work in Section II. Section III describes about the proposed system and Section IV concludes the paper with a summary.

II. RELATED WORK

Clinical decision support system (CDSS) is designed to assist physicians and other health professionals with decision making tasks, such as determining diagnosis of patient data. They are broadly classified into Knowledge based CDSS and Non - Knowledge based CDSS [4]. Fig. 1. depicts about the different methodological branches of CDSS. The knowledge based clinical decision support system contains rules mostly in the form of IF- Then statements. Rule and evidence based systems tend to capture the knowledge of domain experts into expressions that can be evaluated as rules. When a large number of rules have been compiled into a rule base, the working knowledge will be evaluated against rule base by combining rules until a conclusion is obtained. However it is difficult for an expert to transfer their knowledge into distinct rules. The Fuzzy Logic Rule based classifier is very effective in high degree of positive predictive value and diagnostic accuracy. It is a form of knowledge base and has achieved several important techniques and mechanisms to diagnose the disease and pain in patient. Non - Knowledge based CDSS use machine learning technique. Neural Networks have been widely applied to non-linear statistical modeling problem and for modeling large and complex databases of medical information. To derive relationship between the symptoms and diagnosis, neural networks use the nodes and weighted connections. This fulfils the need not to write rules for input. Drawback is that, it consumes much time for training. The generic system goes through an iterative procedure to produce the best solution of a problem.

A. Two - Party Classifier Using ANN

In [5] Smitha Iddalgave et al., presents a new approach on preserving privacy of distributed data for training the neural network. The protocol designed helps to maintain privacy of the network participants working collaboratively on vertically partitioned datasets. The work mainly explains how the parties providing distributed data can jointly train the newly built neural network with their combined data without revealing any of their intricate details, such as their input data and intermediate results generated during the learning process, to one another. Thus it is possible to establish the relation between the data from different providers and classify the combined data in a more meaningful way using the protocol [5]. They propose a two-party classifier algorithm for distributed data scenario. Its features are as follows (i) the network learns over a vertically partitioned data inputs using artificial neural network. The learning technique followed is supervised back propagation with preserving the privacy of data owners. (ii)The algorithm is semi honest in nature i.e. each party learns from its own input and output without acquiring any knowledge about other participant's data. (iii)Since privacy is a major objective here, it needs to use certain cryptographic tools to maintain secure computation of activation function. However, this task is more challenging as most of the cryptographic tools are defined on finite field whereas the activation functions are infinite in nature. Elgamal is the cryptographic tool utilized in the algorithm. It is a public-key encryption scheme defined on any cyclic group. Elgamal scheme is also probabilistic, which means that the encryption operations requires a random number as input along with the plain text. A single plaintext can be encrypted to many possible cipher texts. The drawback is that the encryption scheme has low speed up and high computation cost and consumption time.

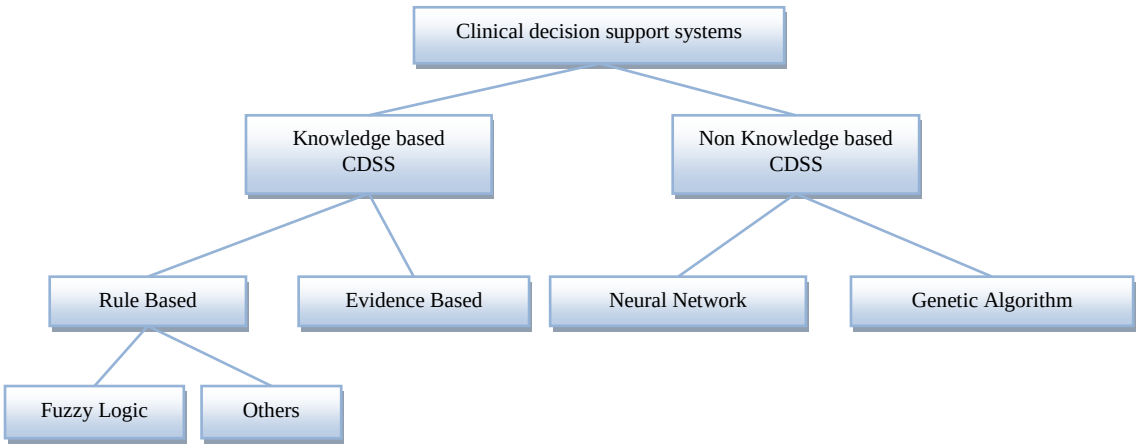

Fig. 1. Different methodological branches of Clinical Decision Support System

B. Gradient Descent Method

In [6] Shuguo Han et al. focus is on protecting the privacy in significant learning model i.e. Multilayer Back Propagation Neural Network using Gradient Descent Methods. For protecting the privacy of the data items, a semi honest model was formulated. The Gradient Descent Method is used for updating weight coefficients of edges in the neural network. In [6] they propose a generic formulation of gradient- descent methods for secure computation by defining the target function f as a composition of $g \circ h(x)$. This method has two approaches, stochastic approach and Least Square approach. In stochastic approach, g is any differentiable function and $h(x) = \sum_{j=1}^{m} h_j (x_j, w_j)$ and in the least square approach, g is an invertible function i.e., g has an inverse function and $h(x) = \sum_{j=1}^{m} x_j w_j$. It can train the neural network by using distributed datasets for solving classification problems. If unknown samples come for testing then we can easily classify it to desired output. Four protocols are proposed for the two approaches incorporating various secure building blocks for both horizontally and vertically partitioned data. The drawbacks of this method are it is not a sufficient method for providing security and proper classifier cannot be implemented.

C. Complementation Approach

In [7] Madhusmita Sahu et al. provide privacy preservation via dataset complementation. It is a data perturbed approach that substitutes each original dataset with an entire unreal dataset. Unlike privacy protection strategies, this new approach preserves the original accuracy of the training datasets without linking the perturbed datasets to the information providers. In other words, dataset complementation can preserve the privacy of individual records and yield accurate data mining results. However, this approach is designed for discrete-value classification only, such that ranged values must be defined for continuous values. The privacy preserving process sometimes reduces the utility of training datasets, which causes inaccurate data mining results. Privacy preservation approaches focuses on different areas of a data mining methods also vary. Madhusmita Sahu et al. in [7] focus on privacy protection of the training samples applied for decision tree data mining. The original data set cannot be reconstructed if some portion of the dataset is stolen by unauthorized party and it only focuses on privacy of training data.

D. PCDS Method

Ching-Ming Chao et al. in [8] have described privacy preserving classification of data streams (PCDS) process is divided into two stages, which are data streams pre-processing and data streams mining, respectively. The primary objective of the first stage, which is handled by the data streams pre-processing system (DSPS), is to perturb data streams to preserve data privacy. The primary objective of the second stage, which is handled by the online data mining system (ODMS), is to mine perturbed data streams to construct an accurate classification model. Data streams continuously flow in DSPS and the arriving time of data is predictable. If DSPS processes data streams immediately upon arrival of the data, this will consume a lot of system resources. Therefore, DSPS adopts the batch processing mode to process incoming data streams. Not only system resources can be more effectively utilized, but also data mining can be more efficiently performed. Whenever accumulating a sufficient amount of data, DSPS uses the data splitting and perturbation algorithm to perturb confidential data as well as computes the error rate resulted from data perturbation. Then DSPS passes perturbed data and the error rate to ODMS. ODMS uses the weighted average sliding window algorithm to mine perturbed data streams to construct a classification model. Because only partial data are available for data mining, ODMS utilizes the sampling method to construct the classification model. In addition, ODMS adopts the sliding window mode to store and process received data streams. There are two reasons for adopting the sliding window model. First, the amount of data streams is enormous and hence it is impossible to store all data. Second, users are usually more interested in more recent data. When data distribution results in a significant change, ODMS reconstructs the classification model to keep it accurate.

E. Anonymization Methods

Benjamin C.M. Fung et al. in [9] says classification is a fundamental problem in data analysis. Training a classifier requires accessing a large collection of data. Releasing person-specific data, such as customer data or patient records, may pose a threat to an individual's privacy. Even after removing explicit identifying information such as Name and SSN, it is still possible to link released records back to their identities by matching some combination of non identifying attributes such as Sex, Zip and Birth date. A useful approach to combat such linking attacks, called k-anonymization is anonymizing the linking attributes so that at least k released records match each value combination of the linking attributes. K-anonymization preserves the classification structure. One way to prevent such linking is masking the detailed information of the attributes by generalizing, suppressing, discretizing etc. By applying such masking operations, the information on birth place, birth year and sex is made less specific, and a person tends to match more records. Thus, the masking operation makes it more difficult to tell whether an individual actually has the diagnosis in the matched records.

III. PROPOSED SYSTEM

A privacy preserving clinical decision support system which preserves the privacy of the patient data, the decision and the server side clinical decision support system parameters is proposed so that the benefits of the emerging outsourcing technology can also be enjoyed in healthcare sector. A decision support system is developed using support vector machine (SVM), which is one of the machine learning tools to compute the decision value. In order to preserve privacy, the SVM algorithm is re – designed as an encrypted domain algorithm using the Paillier homomorphic encryption technique as one of its building blocks. Initially the input for the system is selected which is then followed by dataset loading. Then the dataset is being classified according to class label and normalized. After this stage the data is being encrypted using homomorphic encryption technique. The encrypted data is then stored to the third party server where the decision value is estimated. The results are sent back to the clinician where it is decrypted and the accuracy results are obtained. The architecture of the proposed system is described in fig. 2.

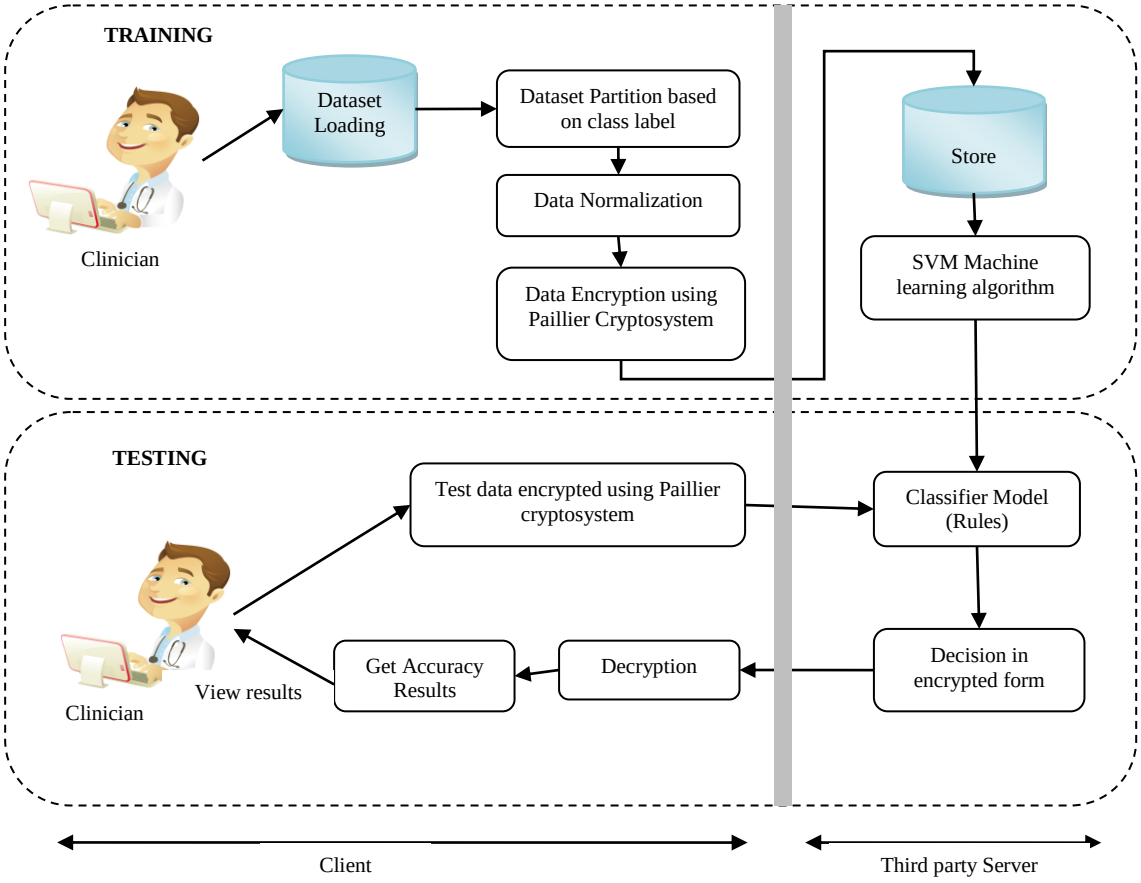

Fig. 2. System Architecture

A. *Data Normalization*

Database normalization is the process of organizing the fields and tables of a relational database to minimize redundancy. Normalization usually involves dividing large tables into smaller (and less redundant) tables and defining relationships between them. The objective is to isolate data so that additions, deletions, and modifications of a field can be made in just one table and then propagated through the rest of the database using the defined relationships. The input for the process is selection of dataset for our application. Here we are using the dataset named Wisconsin Breast Cancer (WBC) dataset. It is one of the medical dataset. It contains patient diabetes details such as age, pedigree level, insulin level etc. In this class attributes are also available. After the dataset has been selected, dataset has been inserted into the database. After the dataset has been loaded, it is preprocessed. That means eliminating the unwanted values or characters in the dataset. Based upon the class attribute in the dataset, we can classify the dataset values. Classification parameters are malignant and benign.

Here we normalize the dataset by using the training and testing datasets. Mean (x^*) and Standard deviation (σ) of the table is computed after pre-processing of the dataset. Normalization is followed by eliminating or compressing the double values. After normalizing the data, we extract the original training data for data encryption. Normalization keeps the numeric values of training samples on the same scale and prevents samples with a large original scale from biasing the solution. Let us denote the normalized training data samples as $x'_i \in R^n$, training set of samples as $x_i \in R^n$ and $i = 1 \ldots\ldots N$ where,

$$x'_i = \frac{x_i - x^*}{\sigma^2}, \forall i \quad (1)$$

Depending on the separability of the training data, this problem is further divided into either a linear classification problem or a non-linear classification problem.

B. *Homomorphic Encryption*

Homomorphic encryption [10] is the encryption on the already encrypted data rather than on the original data with providing the result as it is done on the plain text. The complex mathematical operations can be performed on the cipher text without changing the nature of the encryption. Homomorphic Encryption H is a set of four functions.

H = {Key Generation, Encryption, Decryption, Evaluation}

- *Key generation:* client will generate pair of keys public key pk and secret key sk for encryption of plaintext.
- *Encryption:* Using secret key sk client encrypt the plain text PT and generate Esk(PT) and along with public key pk this cipher text CT will be sent to the server.

- *Evaluation:* Server has a function f for doing evaluation of cipher text CT and performed this as per the required function using pk.

- *Decryption:* Generated Eval(f(PT)) will be decrypted by client using its sk and it gets the original result.

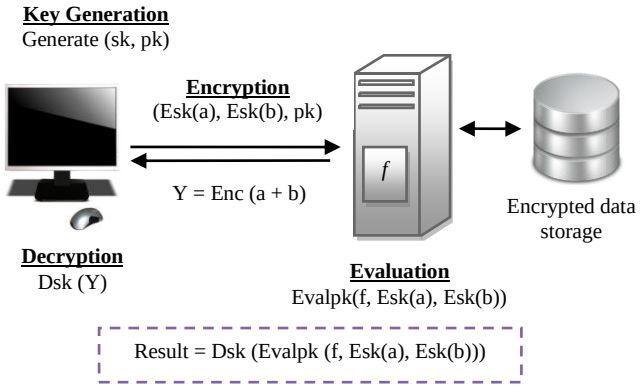

Fig. 3. Homomorphic Encryption Functions

A homomorphic encryption is additive. Here we use Paillier cryptosystem which has the properties of homomorphic encryption. The Paillier cryptosystem is described as follows.

1) Key Generation:

Step 1: n = pq, the RSA modulus

Step 2: $\lambda = lcm (p - 1, q - 1)$

Step 3: $g \epsilon Z / n^2 Z$ s.t. n|or $d_n^2(g)$

Step 4: Public-key: (n, g), secret key: λ, μ

2) Encryption of m:

Step 1: $m \epsilon \{0, 1... n - 1\}$, a message

Step 2: $h \epsilon R Z/n Z$

Step 3: $c = g^m h^n \bmod n^2$, a cipher text

3) Decryption of c:

$m = L (c^{\lambda} \bmod n^2) L(g^{\lambda} \bmod n^2)^{-1} \bmod n$

The constant parameter, $L(g^{\lambda} \bmod n^2)^{-1} \bmod n$ or $L(g^{\alpha} \bmod n^2)^{-1} \bmod n$ where $g=1+ n \bmod n^2$ can also be recomputed once for all.

C. Support Vector Machine

Support vector machines (SVMs), [11] a method for the classification of both linear and nonlinear data. In a nutshell, an SVM is an algorithm that works as follows. It uses a nonlinear mapping to transform the original training data into a higher dimension. Within this new dimension, it searches for the linear optimal separating hyperplane (i.e., a "decision boundary" separating the tuples of one class from another). With an appropriate nonlinear mapping to a sufficiently high dimension, data from two classes can always be separated by a hyperplane. The SVM finds this hyperplane using support vectors ("essential" training tuples) and margins (defined by the support vectors). Although the training time of even the fastest SVMs can be extremely slow, they are highly accurate, owing to their ability to model complex nonlinear decision boundaries. They are much less prone to overfitting than other methods. The support vectors found also provide a compact description of the learned model. SVMs can be used for numeric prediction as well as classification. They have been applied to a number of areas, including handwritten digit recognition, object recognition, and speaker identification, as well as benchmark time-series prediction tests.

Server receives the encrypted file from the user and SVM classification process begins for estimating the decision function value. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier.

IV. CONCLUSION

This paper carries out a wide survey of the different approaches for privacy preserving data mining, and analyses the major algorithms available for each method and points out the existing drawback. While all the methods are only approximate to the goal of privacy preservation, we need to further perfect those approaches or develop some efficient methods. To address this issue, we advise that the following problems should be widely studied:

1) Privacy and accuracy is a pair of contradiction; improving one usually incurs a cost in the other.

2) In distributed privacy preserving data mining areas, efficiency is an essential issue. We should try to develop more efficient algorithms and achieve a balance between disclosure cost, computation cost and communication cost.

3) How to deploy privacy-preserving techniques into practical applications is also required to be further studied.

The proposed algorithm is a potential application of emerging outsourcing techniques such as cloud computing technology, rich clinical datasets (or healthcare knowledge) available in remote locations could be used by any clinician via the Internet without compromising privacy, thereby enhancing the decision making ability of healthcare professionals. We use the homomorphic properties of the Paillier cryptosystem within the proposed algorithm which maintains the privacy.

REFERENCES

[1] Yogachandran Rahulamathavan, Suresh Veluru, Raphael C.W. Phan, Jonathon A. Chambers and Muttukrishnan Rajarajan, "Privacy Preserving Clinical Decision Support System Using Gaussian Kernel-Based Classification", *IEEE Journal of Biomedical and Health Informatics*, Volume 18, No. 1, pp. 56 – 66, January 2014.

[2] Pingshui WANG, "Survey on Privacy Preserving Data Mining", *International Journal of Digital Content Technology and its Applications*, Volume 4, Number 9, December 2010.

[3] Divya Sharma, "A Survey on Maintaining Privacy in Data Mining", *International Journal of Engineering Research and Technology* ISSN: 2278-0181 Volume1 Issue 2, April – 2012.

[4] M.M.Abbasi, S. Kashiyarndi, "Clinical Decision Support Systems: A discussion on different methodologies used in Health Care".

[5] Smitha Iddalgave, Sumana M, "Privacy Preserving Protocol for Two-Party Classifier Over Vertically Partitioned Dataset Using ANN", *International Journal of Science and Research ISSN (Online): 2319-7064 Impact Factor (2012): 3.358* pp. 1654 – 1660, Volume 3 Issue 6, June 2014.

[6] Shuguo Han, Wee Keong Ng, Member, Li Wan and Vincent C.S. Lee, "Privacy-Preserving Gradient-Descent Methods", *IEEE Transactions On Knowledge And Data Engineering*, Volume 22, No. 6, pp. 884 – 899, June 2010.

[7] Madhusmita Sahu1, Debasis Gountia and Neelamani Samal, "Privacy Preservation Decision Tree Based On Data Set Complementation", *International Journal of Innovative Research in Computer and Communication Engineering*, Volume 1, Issue 2, pp. 197 – 207, April 2013.

[8] Ching - Ming Chao, Po - Zung Chen and Chu – Hao Sun, "Privacy-Preserving Classification of Data Streams", *Tamkang Journal of Science and Engineering*, Volume 12, No. 3, pp. 321-330 (2009).

[9] Benjamin C.M. Fung, Ke Wang and Philip S. Yu, "Anonymizing Classification Data For Privacy Preservation", *IEEE Transactions On Knowledge And Data Engineering*, Volume 19, No. 5, pp. 711-724, May 2007.

[10] Payal V. Parmar, Shraddha B. Padhar, Shafika N. Patel, Niyatee I. Bhatt and Rutvij H. Jhaveri, "Survey of Various Homomorphic Encryption algorithms and Schemes", *International Journal of Computer Applications (0975 – 8887)*, Volume 91, No.8, pp. 26-32, April 2014.

[11] Han Jiawei, Micheline Kamber and Jian Pei, *"Data Mining: Concepts and Technique"*, Waltham, USA: Morgan Kaufmann, Elsevier, 2012, pp 408 - 415.