# APPLICATION OF PSO OPTIMIZATION TECHNIQUE ON MEDICAL DATA TO UPHOLD DATA PRIVACY

**[1]ASHA KIRAN GRANDHI, [2]MANIMALA PURI, [3]S. SRINIVASA SURESH**

[1]Research Scholar, University of Pune, Pune, Maharashtra
[2]Director, JSPM Group of Institutions, Pune, Maharashtra
[3]Associate Professor, CSE Department, KMIT, Narayanaguda, Hyderabad
E-mail: [1]ashakiran45@rediffmail.com, [2]manimalap@yahoo.com, [3]sssuresh74@gmail.com

**Abstract—** Data privacy is an emerging research topic, as security issues are on high rise everywhere. Even though, security guidelines, policies, laws and acts are in force, data hackers are on their way to steal the data. Data hackers are not only competent in thefting the data but also experts in analysis. At the same time, there is advancement in technology, to counter data hacking. Generally, medical data contain sensitive information. Hackers use this underlying sensitive information and do fraudulent activities To protect the confidentiality of the data, the researchers have undertaken design and development of an algorithm with the optimization technique to improve the accuracy of privacy on medical data. As per the researchers view, data privacy can be protected through cryptography, and data access controls. Optimization techniques are outweighing the data privacy implementations. The current paper describes self-defined algorithm on medical data (heart data sets), to improve the privacy accuracy on three openly available data sets, applying PSO optimization technique. The current experiment includes standard privacy techniques: clustering, encryption and distribution. In this work, accuracy is measured on varying number of clusters ranging from 2 to 10, on each of the data sets. The paper describes methodology, PSO optimization, and implementation and result evaluation.

**Index Terms—** Data privacy, Privacy Preserving Data Mining (PPDM), PSO

## I. INTRODUCTION

The need for data privacy is increasing day-by-day. So, research in this area is also gaining importance. Data privacy is concerned with confidentiality. According to Reiskind [1], Privacy is about, how you use personal data. It is further concerned with collection of data, the use of data, and the disclosure of the data – to whom, you are giving the data. People do not like to disclose their private data to others for security reasons. Disclosure may impact their personal life. Privacy is multidimensional concept. Privacy can be perceived in many aspects: Informational privacy (e.g., confidentiality, anonymity, secrecy and data security), communicational privacy, territorial privacy (e.g assests), physical privacy (e.g., modesty and bodily integrity), associational privacy (e.g. intimate sharing of death, illness and recovery), proprietary privacy (e.g., self-ownership and control over personal identifiers, genetic data, and body tissues); and decisional privacy (e.g., autonomy and choice in medical decision-making) [2]. In short, privacy has wide spread meaning. Scope of privacy varies from person-to-person, domain-to-domain, and organization-to-organization. There are different opinions between countries on the privacy topic: US refer it as "privacy policy", whereas Europe refers it as "Data Protection" [1]. However, whatever the granularity of the data, from the data owner point view, privacy preservation is important.

Health care industry stay forefront in maintaining the data privacy. Often, health care data includes patient medical records, which need to be secured from various attacks. If effective controls (privacy controls) are in place, asset (i.e data) safe guarding becomes simple. Organizations like hospitals and medical companies majorly focus on data safety and confidentiality. Simultaneously, the concept of medical data sharing is on high rise, for gaining business competency and data insights. Data privacy is concern with usage of the data. This concept is not limited to health care. It is applicable in all service domains. For example, Google offers various services online, and maintains its own privacy policy. Users accept Google privacy policy before registering for their services [3].

Authorized data access does not focus much on privacy, as the producer and consumer agree upon privacy policy. The real problem lies in un-authorized access of the data. Automated tools & technologies, hacker competency, intercontinental are the main elements causing trouble to data banks.

In spite of emergence of new safe guarding technologies, the severity of data maintenance is increasing day-by-day. For example, data mining technologies are threat to data privacy [4]. The data mining tools embeds advanced tools and techniques (Artificial Neural Networks, AI etc.,), having power to extract useful associative information. The extracted information might leak sensitive information, leading to confidentiality leak.

Due to the importance of the data privacy, the need for optimization techniques has gained significance. In this work, we employ PSO (Particle Swam Optimization) to uphold privacy of the data by

adopting cryptography, clustering, data modification and reconstruction techniques. In cryptography, computation is secured. No party knows anything except its own input and the results. Whereas, in reconstruction based techniques, the original distribution of the data is re-constructed from the randomized data. We employ clustering techniques (K-means and Fuzzy-C means) for applying filter based coefficients using PSO optimization. Clustering is the subject of active research in machine learning. We consider health data sets from UCI repository namely Cleveland, Switzerland, and Hungarian containing heart measurements [5], suitable for machine learning applications.
In our earlier publication, details of our experiment were explored at abstract level [6]. The current paper describes detailed PSO modeling on UCI data sets.
The remaining paper is organized as follows:
Section II describes literature review, section III describes methodology, and section IV describes Implementation: PSO optimization and section V describes results followed by conclusion.

## II. LITERATURE REVIEW

W. K. Wong et al. [7] have proposed substitution cipher techniques in the encryption of transactional data for outsourcing association rule mining. After identifying the non-trivial threats to a straightforward one-to-one item mapping substitution cipher, they propose a more secure encryption scheme based on a one-to-n item mapping that transforms transactions non-deterministically, yet guarantees correct decryption. They developed an effective and efficient encryption algorithm based on this method. The algorithm performed a single pass over the database and suitable for applications in which data owners send streams of transactions to the service provider. The results showed that the technique was highly secure with a low data transformation cost.

Ling Qiu et al. [8] have proposed an approach for preserving privacy in association rule mining. The main idea is to use keyed Bloom filters to represent transactions as well as data items. The approach can fully preserve privacy while maintaining the precision of mining results. The tradeoff between mining precision and storage requirement was investigated. They also proposed $\delta$ - folding technique to further reduce the storage requirement without sacrificing mining precision and running time.

Jun-Lin Lin and Yung-Wei Cheng [9] investigated the problem of privacy-preserving mining of frequent itemsets. They presented a procedure to protect the privacy of data by adding noisy items to each transaction. An algorithm is proposed to reconstruct frequent itemsets from these noise-added transactions. The experimental results indicated that the method could achieve a rather high level of accuracy. The method utilizes existing algorithms for frequent itemset mining, and thereby takes full advantage of their progress to mine frequent item sets efficiently.

In 2010, Maneesh Upmanyu, Anoop M. Namboodiri, Kannan Srinathan, and C.V. Jawahar [10] published “Efficient Privacy preserving K -means clustering”. The authors suggest a method for privacy. Which overcomes problems of traditional cryptographic protocols such as Secure Multiparty Computation (SMC) and/or homomorphic encryption schemes like Paillier’s encryption. In this work, authors focused on paradigm of secret sharing, which allows the data to be divided into multiple shares and processed separately at different servers. With the idea of paradigm of secret sharing, the authors designed secured, cloud computing based solution which has negligible communication overhead compared to SMC and proved that the proposed method is a million times faster than similar SMC based protocols.

In 2013, D.Jaya Kumari et al. [11] have categorized sensitive data and non-sensitive data and the sensitive data was hidden from the third party users and the non-sensitive data was made public to the third party users. The data was categorized into two sorts like high sensitive data and low sensitive. The categorization was done mainly to guard the confidential data from the third party users. This process of mining is supposed to be privacy preserving data mining. For categorizing the data they have used fuzzy logic technique. For testing, input was given to the fuzzy rule and received the corresponding output. They have simulated this technique in MATLAB and it shows an efficient result by hiding the confidential data from the third party users.

In 2013, Jaya Kumari, et al. [12] designed MMIS algorithm. The proposed MMIS algorithm is a modified version of existing MSICF algorithm. The MMIS algorithm computes the sensitive itemsets by utilizing the user defined utility threshold value. In order to hide the sensitive itemsets, the frequency value of the items is changed. If the utility values of the items are same, the MMIS algorithm selects the accurate items and then the frequency values of the selected items are modified. The proposed MMIS reduces the computation complexity as well as improves the hiding performance of the item sets. The algorithm is implemented and the resultant item sets are compared against the item sets that are obtained from the conventional privacy preserving utility mining algorithms.

Maheswari et al. [13] proposed a technique Single Value Decomposition method for privacy preserving based on clusters. As per the authors view, data

distortion play important role for preserving the privacy. Based on the data distortion, the author proses a method called Clustering Singular Value Decomposition for confidential numerical attributes. At the same time the experiments preserves general features for clustering. Finally, the performance is evaluated based on the differences between distored databases to undistored database. The difference obtained is 0% for K=3, 4, and 5 with records 6 and 200.

Mithun karmakur et al [14] proposed a scheme based on ENCA (Extended Neighborhood cellular automata). The authors has addressed upon individual adoptability in preferring the privacy levels for dissimilar attributes during perturbation.
Stanley R.M. Oliveira et al [15] have deformed numerical attributes by translations, scaling, rotations, and other geometric transformations. The proposed method has endured that not only the privacy of the data is preserved but also guarantee valid clustering results.

Dowon Hong et al [16] published work emphasizing privacy preserving on horizontally partitioning the data into two sets only. The experiment finds covariance matrices without disclosing mean statistics. The authors also proposes Expectation – Maximization based method for multi-party distributed data.

## III. METHODOLOGY

**Step1: Data Collection**
The present work considers health data from UCI repository [16]. It is a secondary data source. The data sets are openly available for download at free of cost. Machine learning community donate data sets to UCI, for promoting research & development activities. The structure of data sets is suitable for machine learning applications. As of now it has 373 data sets, belonging to various domains. Under heart domain, there are four databases: Cleveland, Hungarian, Switzerland and the VA Long Beach. Donors have their own policy for sharing their data. Donors decide the type of data and granularity of the data [17].
Step2: Data Description
The measurements of the data sets are shown in table 1.

**Table 1. Summary Measurements**

| S.No | Measurements | Value |
|---|---|---|
| 1 | Data Set Characteristics | Multivariate |
| 2 | Number of instances | 244/257/123 |
| 3 | Area | Life |
| 4 | Attribute Characteristics | Categorical, Integer, Real |
| 5 | Number of Attributes | 75 |
| 6 | Date (donated) | 1988-07-01 |
| 7 | Associated tasks | Classification |
| 8 | Missing values | Yes |

**Table 2 Machine Learning Databases – Heart Disease**

| S.No. | Database Name | No. of Attributes | No. of observations | Last Modified Date |
|---|---|---|---|---|
| 1. | Cleveland Data | 14 | 303 | 06/03/1990 |
| 2. | Hungarian Data | 14 | 294 | 14/08/1991 |
| 3. | Switzerland Data | 14 | 123 | 14/08/1991 |

The data sets shown in table 2 do not have name and SSN as they reveal anonymity. Priorly, donors replaced missing values their values with normalized values. Each of the data sets listed in the table 2 are studies and the details of the study are:
i) Cleveland data: The current experiment consider 14 out of the 76 attributes present in the Cleveland heart disease database. Often, researchers employ the Cleveland data sets. The survival of heart disease in the patient is specified in the ''goal'' field using an integer that can take any value from 0 (no presence) to 4. Disease non-existence is represented with value 0 (zero). Class distributions are 54% heart disease absent, 46% heart disease present [18].
iii) Switzerland data: It contains 123 data instances ad 14 features. Class distributions are 6.5% heart disease not present and 93.5% heart disease present [19].
iii) Hungarian data: Three of the features have been removed due to a vast percentage of missing values however the format of the data is precisely the similar as that of the Cleveland data. On account of missing values and 261 examples were present thirty-four examples of the database were removed. Class distributions are 62.5% heart disease not present and 37.5% heart disease present [20].
Step3: Data Privacy.
Each of the data set possess privacy features. With the help of domain experts, the researcher identify the privacy information in the data sets. The details of the features privacy are shown in the table 3.

**Table 3: Attribute Specifications.**

| S.No. | Attribute Name | Attribute Description | Type of Privacy |
|---|---|---|---|
| 1 | Age | Age | Physical |
| 2 | Sex | Sex | Physical |
| 3 | CP | Chest Pain | Physical |
| 4 | Trestbps | Resting BP | Physical |
| 5 | Chol | Serum Cholestoral | Physical |
| 6 | Fbs | Fasting BP | Physical |
| 7 | RestECG | Resting ECG | Physical |

| | | | |
|---|---|---|---|
| 8 | Thalach | Heart Rate | Physical |
| 9 | Exang | Exercise Induced Angina | Physical |
| 10 | Oldpeak | ST depression induced by Exercise | Physical |
| 11 | Slope | The slope of the peak exercise ST Segment | Physical |
| 12 | CA | Number of major vessels colored by Flourosopy | Physical |
| 13 | Thal | Thal | Associational |
| 14 | Num | Angiographic disease status | Associational Privacy |

## IV. IMPLEMENTATION: PSO OPTIMIZATION

To give pre-idea for beginners, this section describe the general PSO algorithm adopted [21, pgno. 06] Followed by the technical architecture of the model under consideration and self-designed swam_Kmeans algorithm pseudo code.

General PSO optimization:

1) Set Parameters N, $\varphi_1$, $\varphi_2$, $V_{max}$, $\omega_{max}$, $\omega_{min}$.
2) //Initialization
3) Initialize a population of N Particles with random position ($X^t$) and velocity ($V^t$).
4) For each Particle i $\epsilon$ 1 → N do
a. $P^t_{(0)} = X^t$
b. Compute the fitness value $f$.
5) end for.
6) Initialize the inertia weight $\omega$.
7) Select the best particle in the swarm $P^t_{(0)}$.
8) // Iteration process
9) for t = 1 to maximum number of iterations do
a. for each particle i $\epsilon$ 1 → N do
b. Update velocity $V^i_t$ and position $X^i_t$ for the particles.
c. Employ the inertia weight update rule.
d. Compute particles fitness value $f(X^i_t)$.
e. Update best particles: $p^i_t$ and $g_t$.
10) end for
11) if convergence criteria are met then
a. Exit from iteration process;
12) end if

The following figure shows the technical architecture of the model under consideration followed by description.

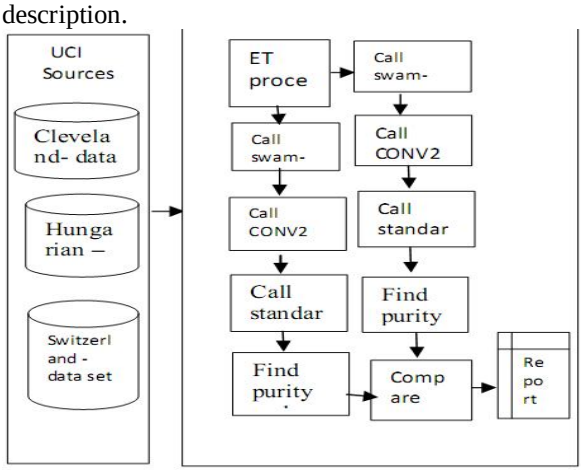



**Figure 1: Technical Architecture**

Represents two layers of implementation. The objective of the whole process is to check the accuracy levels when PSO optimization employed with standard FCM method and standard Kmeans method. Particle Swarm Optimization (PSO) is a population based stochastic optimization technique, developed by Dr. Eberhart and Dr. Kennedy in 1995.

The original idea of PSO was to graphically simulate the movement of bird flock during search of food. Later, it was found that PSO model can be used as an optimizer in computing. In the figure 1, the left layer is called data source layer. It shows the external data sources namely Cleveland data set, Hungarian data set, and Switerzerland data set. As it is mentioned in the above section, these data are donated data sets from various machine learning researchers. The layer on right, indicates process followed to design, implement and measure the privacy accuracy. The figure.1 shows the data movement through arrow point directions. According to the researcher, meaningful description of the overall process is given below:

The implementation of the process starts with data extraction and preprocess. All the input files are csv files. The program accesses the data from each of the csv files and performs preprocess operations. Generally, pre-process converts the data to suitable format and sets the data ready for execution. The author process as Extract –Transform (ET).

The ET process: It identify missing vales and replaces them with mean value, normalize the input data, and samples the input into: train data, class data, test data, test class data. Train data is 80% of the input data.

```
train_feature_data=randomized_feature_data
(1:round (size(feature_data,1)*0.8),:);
train_class_data=randomized_class_data
(1:round(size(feature_data, 1)*0.8),:);
test_feature_data=randomized_feature_data(round(siz
e(feature_data, 1)*0.8)+1:end,:);
test_class_data=randomized_class_data(round(size(fe
ature_data,1)*0.8)+1:end,:);
```

Swam-Kmeans: It performs PSO optimization on the input medical data. The method performs initializations: number of particles, number of dimensions, minimum fitness and maximum fitness, initial fitness, initialization of particles at random, finding Pbest, Gbest, finding filter coefficient for encryption, finding fitness of each particle, and purity.

CONV2 (build-in MATLAB Convolution method): It is a 2-D convolution. It is filtering method useful in data encryption. The current work encrypts feature data (i.e input data without class data) using filter based coefficients. For example, con=conv2 (D, E) computes the two-dimensional convolution of matrices D and E and stores the result into con variable. One of these matrices describes a two-dimensional finite impulse response (FIR) filter, the other matrix is filtered in two dimensions. In this

work, we produce a $5 \times 5$ filter coefficient matrix arbitrarily to apply over the randomized training data for encryption purpose and we obtain an encrypted medical data (or encrypted_featured_data) as output, which is specified as input for clustering.

Pass the encrypted_featured_data to the standard Kmeans method. The following pseudo code shows steps of the overall process. In this pseudo code, comments are indicated with symbol '%'.

Algorithm: Swam_Kmeans optimization

Purpose: The algorithm perform PSO based optimization of the featured data.

Input: population size, feature_data, feature_class_data, numberof_clusters.

Output: PSO best solution

Begin

Initialize input data

Start

For each data set

a) Preprocess the data

b) Normalize the data

c) Randomize the data

d) From Randomized data:

a. Select train _feature data

b. Select test_feature_data

c. Select train_class_data

d. Select train_class_data

e) Perform PSO optimization

i) Define PSO particle structure with

Elements: Next, Velocity, Current, bests far, best.

ii) Initialize PSO elements with random numbers.

%normalize solution u

f) u=pso.particles(p).next;

g) u=u./pso.normalization_divisor;

%obtain encrypted feature data

h) h=reshape(u, [5,5]);

i) h=(1/numel(h))*h;

%apply coefficient matrix on the input data (feature_data))

j) encrypted_feature_data=conv2(feature_data, h, 'same');

k) [idx,C] = means (encrypted_feature_data, num_clusters).

% Idx contains the cluster indices for each of the input data. In other way it contains the predicted class of the input data. 'C 'contains the mean centroid of for each cluster. Num_cluster starts with 2 and continue with 3, 4, and 5,6,7,8,9,10.

l) Find purity (accuracy) between predicted_class and actual class.

End

End

The above stated procedure repeats for number of cluster: 3, 4, 5, 6, 7, and 8,9,10. Initially, Swam-Kmeans starts with 2 clusters. Each time, the PSO accuracy is measured. For example, figure 2 shows, PSO-Kmeans accuracy starting from 2 to 10 clusters. The steps (a) to (e) are repeated for FCM (Fuzzy C Means) method also. Finally, we compare the results of the both the methods and store the results in the internal file. The following section shows the accuracy plots.

## 4.2 Result Evaluation

The idea of this work is to design and develop a technique to measure the privacy accuracy. We achieve optimized values shown in the following figures:

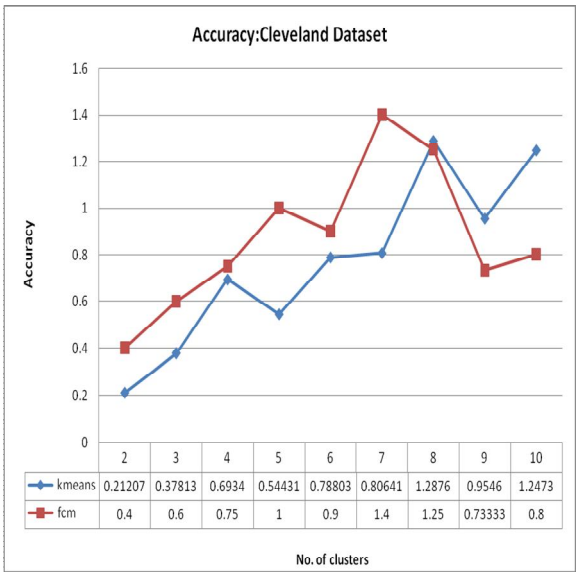



**Figure 2: Performance evaluation of accuracy plot for Cleveland dataset**

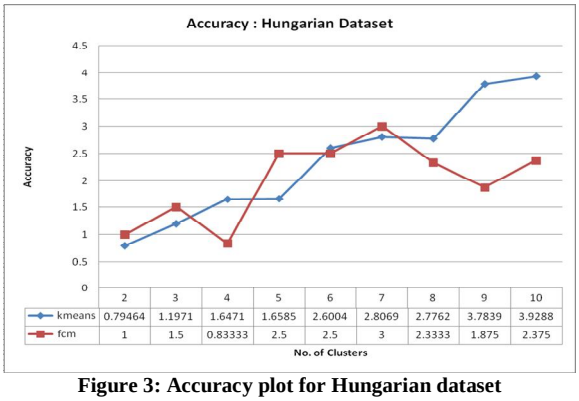



**Figure 3: Accuracy plot for Hungarian dataset**

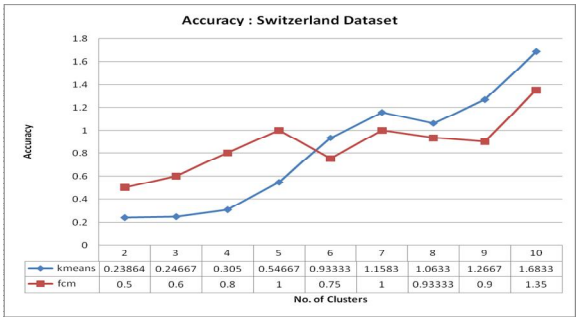



**Figure 4: Accuracy plot for Switzer land data set**

## ACKNOWLEDGEMENTS

Thanks to the Research Center, Abacus Institute of Computer Applications, Savitri Phule Pune

University for their encouragement for pursuing research in the privacy preserving topic. Thanks to, Moshe Lichman-UCI center for Machine Learning Repository, for his timely response to queries regarding data sets.

## CONCLUSION

The current work focused on secondary data sets. Real-time testing may give better insights. We have not focused on encryption time complexity and data transmission as they focus on encryption complexity. The current work lays the foundation for carrying the research further with additional data sets with varying sizes.

## REFERENCES


[1] [1] [1] Michael Minelli, Michele chambers, Ambiga Dhiraj, Big Data Big Analytics, Wiley CIO series, ISBN: 978-81-265-4469-1, 2016.

[2] [2] Clarkson, Topics in computer Ethics, available at http://people.clarkson.edu/~bjukic/BE IT/Topics%20in%20Computer%20Ethics.htm

[3] [3] Google Privacy policy, available at https://www.google.com/policies/privacy, last updated April 17 2017.

[4] [4] Bacal & Associates, Customer service zone,availableat,http://customerservicezone.com/db/Data_Mining, last updated:

[5] [5] Arthur Asuncion and David Newman, UCI machine learning repository, available, at http://archive.ics.uci.edu/ml/about.html.

[6] 2007.

[7] [6] Asha kiran, Manimala Puri, Srinivasa Suresh, PSO Enabled Privacy preservation, Indian Journal of Science and Technology, Vol 10(11), DOI: 10.17485/ijst/2017/v10i11/89318, March 2017, ISSN:0974-5645(online)

[8] [7] W. K. Wong, David W. Cheung ,Edward Hung, Ben Kao, Nikos Mamoulis, "Security in outsourcing of association rule mining", in proceedings of the 33rd international conference on Very large data bases, pp: 111- 122, Vienna, Austria, 2007.

[9] [8] Ling Qiu , Yingjiu Li, Xintao Wu, "Preserving privacy in association rule mining with bloom filters", Journal on Intelligent Information Systems, Vol. 29, No. 3, pp.: 253-278, 2007.

[10] [9] Jun-Lin Lin, Yung-Wei Cheng, "Privacy preserving item set mining through noisy items", in proc. of International Journal on Expert Systems with Applications, Vol. 36, No. 3, pp.: 5711- 5717, 2009.

[11] [10] In 2010, Maneesh Upmanyu, Anoop M. Namboodiri, Kannan Srinathan, and C.V. Jawahar published "Efficient Privacy preserving K -means clustering." _c Springer-Verlag Berlin Heidelberg 2010.

[12] [11] D.Jaya Kumari, Prof. Nistala.V.E.S.Murthy and Prof. S.Srinivasa Suresh, "An Approach on Privacy Preserving Data Mining Based on Fuzzy C-Means", IFRSA International Journal of Data Warehousing & Mining, Vol. 2, pp. 136-143, 2012.

[13] [12] D.Jaya Kumari, Prof. Nistala.V.E.S.Murthy, Prof.S.Srinivasa Suresh, "An Optimization based Modified Maximum Sensitive Item-Sets Conflict First Algorithm (MSICF) for Hiding Sensitive Item-Sets", International Journal of Computer Applications, Vol. 72, No.4, pp. 0975 – 8887, 2013.

[14] [13] Maheswari and K. Duraiswamy, "CLUST-SVD: Privacy preserving clustering in singular value decomposition", World Journal of Modelling and Simulation, Vol. 4, No. 4, pp. 250-256, 2008.

[15] [14] Mithun Karmakar and Dhruba K Bhattacharyya, "Privacy Preserving Data Mining Using Matrix Algebraic Approach", Journal of Convergence Information Technology, Vol. 4, No. 3, pp. 38-44, 2009.

[16] [15] Stanley R..M. Oliveira and Osmar R. Zaiane, "Revisiting Privacy Preserving Clustering by Data Transformation", Journal of Information and Data Management, Vol.1, No.1, pp.53-56, February 2010.

[17] [16] Luong The Dung and Ho Tu Bao, "Enhancing Privacy in Distributed Data Clustering", Journal of Computer Science and Cybernetics, Vol. 26, No. 2, pp. 1-15, 2010)

[18] [17] David W. Aha, UCI machine learning repository, Heart decease datasets[http://archive.ics.uci.edu/ml/citation_policy.html],Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D, University Hospital, Zurich, Switzerland: William Steinbrunn, M.D. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.

[19] [18]Robert Detrano, M.D., Ph.D. Machine learning repository, Heart decease data sets[http://archive.ics.uci.edu/ml/citation_policy.html], Cleveland Clinic Foundation.

[20] [19] William Steinbrunn, M.D. UCI repository, Heart deceasesets[http://archive.ics.uci.edu/ml/citation_policy.html] ,University Hospital, Zurich, Switzerland.

[21] [20] Andras Janosi, M.D. UCI repository, Heartdeceasesets[http://archive.ics.uci.edu/ml/citation_policy.html], Hungarian Institute of Cardiology.

[22] [21] Sanjay Saini et al "Review on Particle Swarm Optimization Algorithm and Its Variants to Human Motion Tracking". Hindawi Publishing Corporation Mathematical Problems in Engineering Volume 2014, Article ID 704861. Available on line at http://dx.doi.org/10.1155/2014/704861.


★★★