PSV-GWO: Particle Swarm Velocity Aided GWO for Privacy Preservation of Data

Jyothi Mandala[1,2,*] and Dr. M. V. P. Chandra Sekhara Rao[3]

[1] *Research Scholar, ANU, Guntur, Andhra Pradesh 522019, India*
[2] *Assistant Professor, GMRIT, Rajam, Andhra Pradesh 532127, India*
[3] *Professor, RVR & JC College of Engineering, Guntur, Andhra Pradesh 522019, India*
E-mail: jyothirajb4u@gmail.com
[*] *Corresponding Author*

Received 09 June 2018; Accepted 21 September 2018;
Publication 20 June 2019

Abstract

Due to the maximum usage of Social Networking Sites (SNS) the number of individuals that are posting their health information online is increasing. The health information of the user'sis disclosed on these sites, where the organization or various individuals can mine that for numerous research and commercial purposes. Because of this sensitive nature of the medical information, the privacy protection is said to be a main focus for the researchers. On analyzing many of the conventional methods, there is an improvement in the sanitization process but still lacks on the restoration of data. Thus, this paper focused on the privacy preservation over the healthcare records. The proposed model is about the enhancement in the sanitization technique that hides the raw information presented by the users. The sanitization process involves the generation of key that created optimally by introducing a new Particle Swarm Velocity aided GWO (PSV-GWO) algorithm. Additionally, the authorized user can restore these sanitized medical data securely. Finally, the traditional algorithms are compared with the proposed model in terms of Particle Swarm Optimization (PSO), Genetic Algorithm (GA), Differential Evolution (DE), Crow Search Optimization (CSA) and Adaptive Awareness Probability-based CSA (AAP-CSA) and the outcome is analyzed.

Journal of Cyber Security and Mobility, Vol. 8_4, 439–466.
doi: 10.13052/jcsm2245-1439.843

Keywords: Healthcare Data Preservation, Data Restoration, Sanitized Data, Key generation, Modified Optimization.

1 Introduction

The large quantities of data that are extracted from the unknown previously interesting patterns by using the automatic and semi-automatic analysis are said to be the task in data mining. The data mining [9, 11] includes a collection of unusual records (anomaly detection); dependencies (association rule mining) and similar data records (cluster analysis). Generally, the task in data mining is divided into two phases: descriptive and predictive. The general properties of data are characterized by the Descriptive mining in the database. The task of Predictive mining was to execute the current data inference thereby predictions are made.

The user data in social network sites of entire kind, i.e., the search engines and shopping sites can be further utilized and analyzed by data mining [15, 17] in organizations and individuals. The raw data is unavoidably in revealing and privacy leakage can occur at the time of this process, because of the use of private and sensitive information. Diversely, in publishing applications of many data that presented the data directly to the users in database, the data protection has to be made by the data publishers, or it will lead to the leakage in sensitivity data. Hence, the privacy has to be provided without any compromise in significant accuracy of data mining [19] by using privacy protection technique [20]; this is the major challenge in the data mining. For science and business purposes the ATA mining is used widely. The data that are collected by the individuals or information providers are the major one for pattern recognition or decision making.

Data encryption, data distortion, and limited data publishing, etc. are involved in the existed privacy preservation [12, 13] strategy. The encryption technique is accepted in the mining data procedure by data encryption for hiding the data that is sensitive that is utilized often in the dispersed environment. The data is published provisionally on definite conditions by the limited data publishing thereby the path of publishing definite values of data, anonymizing or generalizing the data, and so on. In privacy protection [14], the sensitive data is distorted by using the Data distortion strategy when maintaining the data attributes or some data is integrated by the addition of noise, blocking, making exchange and randomization and so on. The processed data can be making sure to protect the definite statistics properties in mining data and for additional process.

To build the algorithmic scale and to attain a larger accuracy, when managing the guaranteed privacy is the major challenge in Privacy preservation in data mining (PPDM) [10, 30]. The existed methods and definitions of privacy are not suitable for the PPDM [16, 18, 29] techniques. Additional statements will lead to the low computational cost and help to acquire better accuracy. Big data is taken into consideration because of its risk process. This involves the lifecycle of information, collection process, and data creation, and also the requirement in security process. The objective of this big data security is the same as that of the previous methods. They are to protect the availability, confidentiality, and integrity.

The proposed privacy preservation model implements the improvement of sanitized method for hiding the raw data that are offered by the users. The key generation is involved in this sanitized process. This paper introduces a new algorithm namely PSV-GWO to find the optimal key. Further, the sanitized medical data is restored effectively by the authorized user. At last, the proposed method is compared over the traditional algorithms like PSO, GA, DE, CSA and AAP-CSA and the resultant outcome is analyzed. The organization of this paper is as follows: Section 2 explains the Literature review. Section 3 analyses the modeling of medical data privacy preservation. The optimal key extraction strategy is described in Section 4. Section 5 explains the result and discussion work of this paper, and finally, Section 6 concludes the paper.

2 Literature Review

2.1 Related Works

In 2016, Li et al. [1] have implemented two dispersed privacy-preserving protocols that have been based on the distributed ensemble techniques. The main impact of the implemented technique was for the purpose of learning the data distribution, to outline an elegant approach more accurately. Further to transmit the achieved healthcare knowledge not by showing and sharing the sensitive data of client or patient, and thereby privacy of the patient has been protected. They have verified the implemented model has an effective performance in accuracy as well as in the time robust prediction techniques. The implemented model performance has estimated with the help of type-2 diabetes 'electronic health records EHRs', which was gathered from numerous sources. Further, with the assist of implemented model, they can found the fundamental biomarkers (both universal as well as region-specific), and also have certified the chosen biomarkers by biomedical literature.

In 2018, Ni et al. [2] have implemented a Differential Privacy Preservation Multiple Cores DBSCAN Clustering (DP-MCDBSCAN) schema on the basis of the differential privacy that was powerful as well as by the algorithm named DBSCAN for effectively influence the privacy leakage problems for the user data network within the procedure of mining the data, and to improve the efficiency in the data clustering thereby in addition of Laplace noise. The wide theoretical review and simulations were performed to estimate that the resultant structure has shown enhanced accuracy, efficiency, and privacy preservation result when compared with the conventional structures.

In 2017, Gao et al. [3] have developed an arrative reversible data hiding (RDH) algorithm mainly for medical images. The major aim of this implemented algorithm was to achieve a divergence enhancement in 'region of interest (ROI)' without any distortion and thereby accomplished interfere localization above assault in ROI. At first, the background and the ROI of consequent image were segmented with the assist of a threshold method 'Otsu's'. Moreover, a better technique was applied for pre-processing the diminishing intention of visual distortion. The implemented method was comparing with other conventional models, which show that Experiment was estimated the dominance over implemented algorithm corresponding to an enhancement of contrast in ROI, also to safeguard the quality of visual and location of tamper.

In 2017, Zhang et al. [4] have developed a new privacy-preserving decision tree classification construction model on the basis of various privacy-protection methods, to review the issues in the privacy disclosure at the time of data mining. The feedback that was used in efficient classifier was divided into two various noises through exponential mechanisms and Laplace, thereby the resultant calculation was disturbed and was presented to a construction algorithm in which a secure data assessment interface was provided to the users. The continuous and discrete values was provided with various split solution and was utilized for the optimization of search structure to minimize the rate of error in classifier. The lower sensitivity quality function was available, thereby chosen to make decisions and to enhance the allocation budget in the privacy method. The personal information that was obtained by the unknown sensitive nodes in tree data type in the potential problem was resolved accordingly. The simulation experiment showed the better accuracy and the privacy protection in the implemented model.

In 2017, Kim et al. [5] have intended the efficiency evaluation and also the proficiency of data cubes (preservation of privacy) in 'electronic medical records (EMRs)'. EMR statistics became difficult because of these data cubes that were summarized by the entire feasible blends of attributes. The big data was analyzed effectively by these extensive data cubes, which was a large probability in analyzing such as EMR analysis. They have developed a privacy-preserving structure for EMR data cube, mainly; the privacy of data has to be attained by using the anonymization models. Moreover, they have paid attention on alterations that happened in privacy preservation by the process of anonymization. Hence virtually evaluated the several types of 'privacy-preserving EMR data cubes' with the utilization of definite metrics as well as argued about the ability of every anonymization method.

In 2012, Fong and Jahnke [6] had introduced the privacy-preserving method, which was used in decision tree learning with no associated accuracy loss. Here the privacy preservation in the gathered data samples was described at times when the sample database information was lost partially. In this method, the sample original datasets were converted into anunreal dataset groups; thereby the sample original datasets was not recreated by not utilizing the total unreal datasets group. At the same time, these unreal datasets directly build the decision tree with clear accuracy. The data storage was applied directly by this approach by how fast the first sample was gathered. This method was suitable for the other conventional methods, like cryptography, thereby promotes additional protection.

In 2017, Poulis et al. [7] have presented a narrative method, which enforced the exact requirement. The efficacy constriction concept was developed for both codes (demographics and diagnosis codes). The generation number can be limited by these efficacy constrictions that in turn defined by data owners. The algorithm was developed in order to appreciate the developed model in which it enforce (k,km)-anonymity on a dataset that includes both stated codes; thereby it would convince the exact efficacy constrictions with fewer information loss. The experiment along with a large dataset that involves more than 200,000 'electronic health records' was examined for the effectiveness and proficiency in implemented algorithm.

In 2016, Xu et al. [8] have stated that the researchers required a path for organizing the various ongoing works, to guard the sensitive information in the data mining. The Rampart framework categorizes protection method was implemented for encouraging the inter-disciplinary clarification in growth variation of privacy issues connected along the data with knowledge discovery.

2.2 Review

The features and challenges of privacy protection in data mining are summarized in Table 1. The methodology along with the features and challenges are

Table 1 Features and challenges of various privacy preservation models for anonymous data

Author	Method	Features	Challenges
Li et al. [1]	Ensemble learning	• Identification is on ease for the region definite biomarkers. • Guide a new innovative clinical area	• Less accuracy • Further enhancement is required to attain the idea of anonymity.
Ni et al. [2]	DP-MCDBSCAN	• Noise added amount is independent of the dataset scale • Huge dataset needs only a small amount of noise	• Influence of input parameter have to be reduced • Less accurate
Gao et al. [3]	Reversible data hiding	• ROI contrast is largely improved • Redundant shifting process is evaded	• Unable to find the tamper • Cannot execute contrast enhancement
Zhang et al. [4]	Differential privacy decision tree construction model	• Entirely independent of any background knowledge • Non sensitive to any data modification in records.	• No accurate access to potential loss in privacy • Offer protection against a single knowledge attack model
Kim et al. [5]	Generalization method	• Can access and build the anonymized EMR data cubes • It measured the features of EMR analysis.	• Difficulties in getting the optimal result • Calculated only the count measure.
Fong and Jahnke	Decision Tree Generation	• Improves privacy security • Utility of the sample data setsare preserved	• Optimization needed for storage size of the unrealized samples • The processing time is low.
Pouliset al. [7]	Generalization method	• Protect large data utility • More efficient	• Only thinks the unordered sets • Cannot appropriate when a setting gets varied.
Xu et al. [8]	Rampart framework	• Avoided direct use of sensitive raw data • Security in delivering mining results was provided	• Control in use of personal information has to be enhanced • Risk in covering user's personal information

as follows: Ensemble learning [1] makes ease in identification of the region definite biomarkers and guides a new innovative clinical area records. But it still possesses some challenges that are it has less accuracy and moreover, the enhancement is needed to attain the idea of anonymity. DP-MCDBSCAN [2] the amount of noise added is independent of the dataset scale, and a small amount of noise is needed for the huge datasets. The drawback of this method is less accuracy and influence of the input parameters have to be reduced. Reversible data hiding [3] has an improved ROI contrast, and the redundant shifting process has to be evaded. Unable to find the temper and the contrast enhancement cannot be executed; these are the major challenges in this model. Differential privacy decision tree construction model [4] is entirely independent of any background knowledge, and it is nonsensitive to any data modification in records. The drawbacks are no access to accuracy of potential loss in privacy, and it offers protection only against the single knowledge attack model. Generalization method [5] can access and build the anonymized EMR data cubes, and the features of EMR analysis are measured. The major disadvantage is, there is a difficulty in getting the optimal result, and only the count measures are calculated. Decision tree generation [6] improves the privacy security, and the utility of the sample data is preserved. The challenges that are to be rectified for future use are: the storage size in the unrealized samples has to be optimized and the processing time is low. Generalization method [7] protects large data utility, and it is more efficient. It only thinks about the unordered sets which are the major drawback of this model and cannot be corrected when a setting gets varied. Rampart framework [8] avoids the direct use of sensitive raw data and provides the security in delivering mining results. The limitations it posses are, enhancement needed to control the use of personal information, and there is a risk in covering user's personal information.

3 Modeling of Medical Data Privacy Preservation

3.1 General Architecture

The data sanitization and data restoration are the two processes that are involved in the proposed medical data preservation model. At first, sanitization process carried out under sensitive data, and for hiding the sensitive data, a key is created. As the optimal key is the major issue, this paper uses a new PSV-GWO for generating optimal key. Hence the secure transmission of the sanitized data in database is done via transmission line, after that it puts to

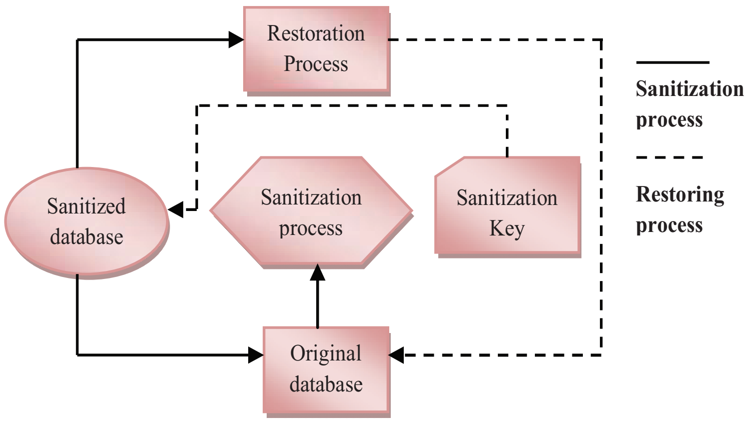


Figure 1 Privacy preservation model's overall architecture.

Table 2 Transactions in the database

Transactions	Data		
I_1	1	2	
I_2	1	3	
I_3	2	3	4
I_4	1	3	4
I_5	3	4	

the restoration process from where the medical data that is sanitized was recovered effectively by the authorized user. Here the Figure 1 illustrates the overall architecture model of privacy preservation.

Let assume the database be I from Table 2. Here the first transaction is given as I_1 and the second transaction is given as I_2 and so on. The maximum length is defined as $I_{\max}$, V_I is the number of transactions. The closest higher perfect square of V_I is referred by V_I^v. The $I_{\max}$ value as per the table is 3. $V_I = 5$, $V_I^v = 9$ (perfect square next to 5 is 9).

3.2 Data Sanitization

The Figure 2 shows the sanitized data, here I' is achieved by the sanitized key generated by the key generation process in the original database. The outcome key matrix M_2 and I is binarized for carrying the XOR operation. Subsequently, the unit step input is summed up and I' is achieved as per the Equation (1).

$$I' = (M_2 \oplus I) + 1 \tag{1}$$

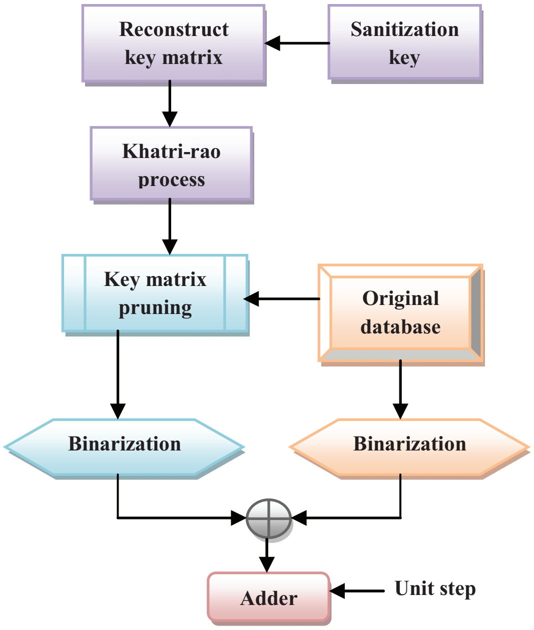


Figure 2 Sanitization process framework.

3.3 Key Generation

The key generation process for doing sanitization is resembled in the Figure 3. The key is generated by the PSV-GWO model by randomly initializing the population of various keys. This is followed after the sanitization process, from where the sanitized database is achieved. At the same time, the sanitization process that attains the sanitized database along with the original database attain the association rule and evaluates the objective functions as o_1, o_2, and o_3, respectively. Finally, the key value is constantly updated until attaining the extreme termination measure and gains the best needed solution. The key is optimally produced by the proposed PSV-GWO method for data sanitization process. The length of the chromosome is allocated on the basis of $\sqrt{V_I^v}$ value. The parameters are defined by $\left[0, \sqrt{\max(V)}\right]$, here V refers to the original database. As per the Table 2, $\max(V) = 4$ that is the database's largest item set.

3.4 Data Restoration

The decoding process is illustrated in Figure 4. The sanitization process offers I' and M_2 from key generating criteria are needed to be binarized in this decoding process. The sanitized database from binarization block is minimized

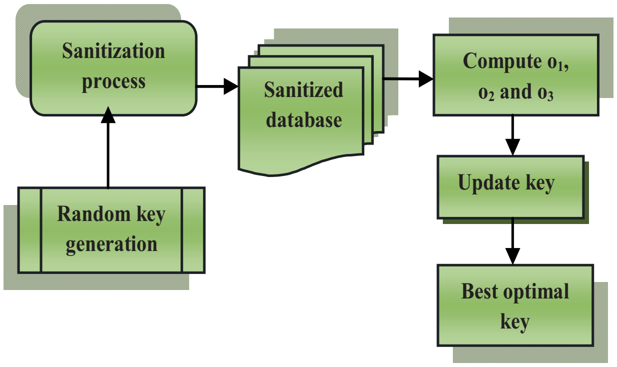


Figure 3 Key generation process framework.

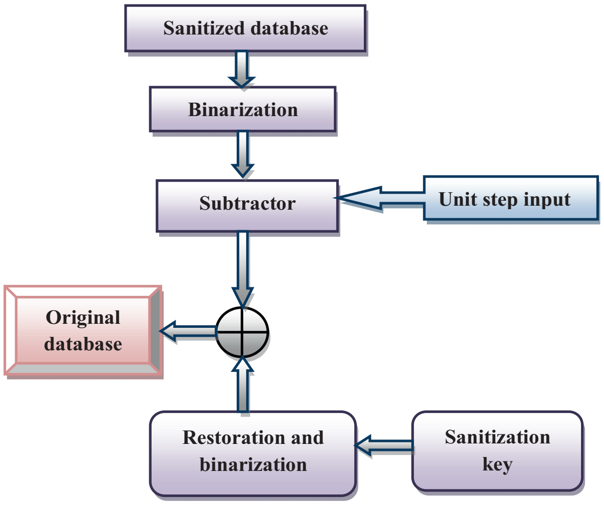


Figure 4 Decoding process framework.

from unit input step. Meanwhile, the XOR operation is carried out in the key matrix that is binarized and the database after minimization, and the restored database is recovered. Further, it is expressed as; the key generation process produces sanitized key, which is exploited to do the restoration of database I. It is utilized to create the sanitized database I', from which the loss-less restoration cloud takes place with respect to Equation (2). Here $\hat{I}$ refers to the restored data and the sanitizing key matrix is defined by M_2 that are recreated by M. The algorithm for the restoration process is given in Algorithm 1.

$$\hat{I} = (I' - 1) \oplus M_2 \tag{2}$$

```
Algorithm 1 Restoration process
Input: $I$ refers to sanitized database, $M_2$ refers to the sanitizing key
Output: $I^{'}$ denotes the restored database
                Restoration process
Step 1:   Solution transformation of $M_2$
Step 2:   Binarization of $I^{'}$ and $M_2$
Step 3:   Subtraction by unit step input
Step 4:   Carry out XOR function
Step 5:   Return $\hat{I}$
```

4 Optimal Key Extraction Strategy: Proposed Sanitization and Restoration Model

4.1 Key Encoding

The chromosomes (key) M that are utilized for the process of sanitization are given to the proposed PSV-GWO algorithm. The number of key ranges between M_1 and M_2 is optimized by deploying the proposed PSV-GWO method; thereby the optimal key is attained. The Figure 5 illustrates the solution encoding process in which the chromosome or key length is given as $\sqrt{V_I^v}$.

4.2 Key Transformation

The transformation of the chromosome M in the solution transformation process is done by Khatri-rao product. At first, M is reconstructed as M_1 with matrix dimension $\left[\sqrt{V_I^v} \times I_{\max}\right]$. For example, the reconstruction process of $M = 0,2,1$ takes place the row-wise duplication that generates the key matrix, the Equation (3) shows the dimension of M_1 as $\left[\sqrt{V_I^v} \times I_{\max}\right]$, here the row matrix is allocated on the basis of $\sqrt{V_I^v}$ and the column matrix is on the basis of $I_{\max}$.

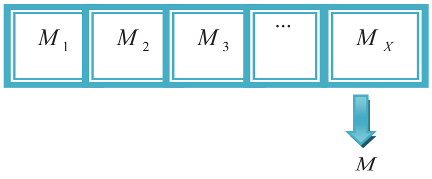


Figure 5 Key Encoding.

Rebuild $M$with size $\left[\sqrt{V_I^v} \times I_{\max}\right]$

$$M_1 = \begin{bmatrix} 0 & 0 & 0 \\ 2 & 2 & 2 \\ 1 & 1 & 1 \end{bmatrix} \tag{3}$$

Consequently, M_2 with key dimension $[V_I \times I_{\max}]$ is accomplished by Khatri-rao product as $M_1 \otimes M_1$, here $\otimes$ refers to the Kronecker product and the sizes are trimmed with respect to the original data base dimension is given in Equation (4).

$$M_2 = \begin{bmatrix} 0 & 0 & 0 \\ 2 & 2 & 2 \\ 1 & 1 & 1 \end{bmatrix} \otimes \begin{bmatrix} 0 & 0 & 0 \\ 2 & 2 & 2 \\ 1 & 1 & 1 \end{bmatrix} \tag{4}$$

On the basis of the Khatri-rao function, the key generation is performed by the M_1, thereby it creates the similar matrix that is equivalent to original database $M_2[V_I \times I_{\max}]$. Finally, the rule hiding process is involved to attain a sanitization data base, I' by hiding the sensitive data. Additionally, the binarization of key matrix and the original database takes place. Because of this, the rule hiding operation is process with the binarized key matrix pruning, in which the binarized original database carries out the XOR function by similar matrix sizes and added up by the one that created the sanitized database, given in Equation (10), here M_2 defines the pruned key matrix. Added to this, I' withdraw the association rules and the sensitive rules that achieved from the sanitization process which is prior to the M sanitization. Hence Equation (1) is evaluated and the sanitized database I' is achieved on the basis of Khatri-rao process.

A. Fitness Evaluation

Subsequent to the generation of the sensitive and association rules of sanitized and original database, the three objective functions o_1, o_2, and o_3 are calculated with respect to the Equations (5), (6) and (7). Here in Equation (5), Q_K denotes the sensitive item set frequency in sanitized information and Q_G denotes the sensitive item set frequency in original information. Consequently, Q_N in Equation (6) denotes the non-sensitive item set frequency in sanitized information. The Euclidean distance among the original information I and sanitized information I' is given in Equation (7). The distance among every item set/element in sanitized and original information is denoted by o_4 and is shown in Equation (8). Moreover, the proposed structure's fitness function is denoted as Q.

$$o_1 = \frac{Q_K}{Q_G} \tag{5}$$

$$o_2 = \frac{Q_N}{Q_G} \tag{6}$$

$$o_3 = D(I, I') \rightarrow Euclidean\ distance \tag{7}$$

$$\begin{aligned} Q = {} & \frac{z_1 o_1}{\max[o_1, o_2]} + z_2 \left[1 - \frac{o_2}{\max[o_1, o_2]}\right] \\ & + z_3 \left[\frac{o_3}{\max(o_4)}\right] \end{aligned} \tag{8}$$

The implemented privacy prevention's objective function in medical data is shown in Equation (9).

$$H = Min(Q) \tag{9}$$

B. Conventional GWO Algorithm

To model the social hierarchy of wolves mathematically while making GWO [22], the fittest solution is defined as alpha (α). The better solution of second and third is given as beta (β) and delta (δ), correspondingly. The remaining solution of the candidate is referred to as omega (ω). The hunting or optimization in GWO algorithm is aided with the help of α, β and δ. The wolves ω chase these wolves.

Encircling Prey: At the time of hunting process the grey wolves encircle the prey, and this encircling behavior is proposed mathematically by this following Equations (10) and (11), respectively.

$$\vec{B} = |\vec{E}.\vec{M}_q(u) - \vec{M}(u)| \tag{10}$$

$$\vec{M}(u+1) = \vec{M}_q(u) - \vec{H}.\vec{B} \tag{11}$$

Here, the current iteration is given by u. $\vec{H}$ and $\vec{E}$ are referred as the coefficient vector. The position vector of the prey is given as $\vec{M}_q$ and the grey wolf position vector is given as $\vec{M}$. The vector calculation of $\vec{H}$ and $\vec{E}$ are given below in Equations (12) and (13), respectively.

$$\bar{H} = 2\vec{e}.\vec{s}_1 - \vec{e} \tag{12}$$

$$\bar{E} = 2.\vec{s}_2 \tag{13}$$

Here, the $\vec{e}$ component is reduced linearly between 2 to 0 for a course of iteration and $\vec{s}_1$ and $\vec{s}_2$ are random vectors within [0,1].

Hunting: Grey wolves have a special capability of identifying the prey location and surround them. Generally, the alpha aided the hunt, and occasionally beta and gamma joined with them. Though within the certain search space, the location of the prey is not known. The hunting behavior of grey wolves is mathematically simulated, by using alpha, beta and delta wolves' better knowledge on the possible location of the prey. The first three best solutions are taken into account, whether the rest is compelled. The Equations (14), (15) and (16) is given as follows.

$$\begin{aligned}&\vec{B}_\alpha = \left|\vec{E}_1.\vec{M}_\alpha - \vec{M}\right|, \vec{B}_\beta = \left|\vec{E}_2.\vec{M}_\beta - \vec{M}\right|,\\&\vec{B}_\delta = \left|\vec{E}_3.\vec{M}_\delta - \vec{M}\right|\end{aligned} \tag{14}$$

$$\begin{aligned}&\vec{M}_1 = \vec{M}_\alpha - \bar{H}_1.(\vec{B}_\alpha), \vec{M}_2 = \vec{M}_\beta - \bar{H}_2.(\vec{B}_\beta),\\&\vec{M}_3 = \vec{M}_\delta - \bar{H}_3.(\vec{B}_\delta)\end{aligned} \tag{15}$$

$$\vec{M}(u+1) = \frac{\vec{M}_1 + \vec{M}_2 + \vec{M}_3}{3} \tag{16}$$

Attacking Prey: the model was mathematically implemented by reduce the value of $\vec{e}$, while approaching the prey. Hence $\vec{H}$ fluctuation range also reduces with the $\vec{e}$. Further, it says that $\vec{H}$ is a random value with interval $[-2e, e]$, whether $\vec{e}$ component is reduced linearly between 2 to 0 for a course of iteration.

C. Conventional PSO Algorithm

The PSO [21] model is proposed originally to stimulate the bird flock's social behavior, yet the simplification is made in this algorithm and has understood that the individuals are termed as particles, which are performing the optimization.

In this PSO model, initially the particles are placed randomly within the search space, i.e., randomly moved in defined directions. The particle's direction can changed gradually, and hence it started to go along the direction of the previous best position by itself. Then it searches the neighborhood and discovered the best positions regarding some fitness function. $fit = S^m - S$.

Here the particle‘s position is given as $\vec{M} \in S^m$ and velocity be $\vec{w}$. At first, these two variables are randomly chosen, after that in accordance with two formulas it is iteratively updated and is given in Equation (17)

$$\vec{w} = \omega\vec{w} + c_1 r_1(\vec{q} - \vec{M}) + c_2 r_2(\vec{f} - \vec{M}) \tag{17}$$

Here, user-defined behavioral parameter ω is referred as an inertia weight that controlled the recurrence amount in the velocity of particles. The preceding best position (personal best) of particle is $\vec{q}$ and $\vec{f}$ is the preceding best position in swarm (global best); thereby which the particles implicitly communicate with one another. This is weighted by using the stochastic variable $r_1, r_2 \sim U(0, 1)$ and c_1, c_2 is the acceleration constant. The velocity is added to the current position of the particle to move to the next position in search space, in spite of any fitness improvement.

$$\vec{M} \leftarrow \vec{M} + \vec{w} \tag{18}$$

D. Proposed PSV-GWO

Though the conventional algorithms have better performance in optimization problems, still they pose some limitations that have to be rectified. The conventional GWO algorithm also poses some limitations such as slow convergence, bad local searching ability, and low solving precision. The PSO algorithm yet poses some drawbacks such as it still needs an improvement over the wide range of field. More work is needed further for improving the convergence and the robustness. To overcome these problems, this paper aims to introduce a new hybrid algorithm. The proposed PSV-GWO is explained as follows: Here, the PSO characteristic is incorporated in GWO algorithm. In the proposed model, the encircling of prey mathematical model is given in Equations (10) and (11). The hunting process mathematical model is illustrated by the Equations (14), (15) and (16). The major modification of proposed model is goes with the position update. The new position update of the PSV-GWO algorithm is given in Equation (19). Here $\vec{M}$ is the velocity of the position update of PSO and it is given in Equations (17) and (18).

$$M(u + 1) = \frac{\vec{M}_1 + \vec{M}_2 + \vec{M}_3 + \vec{M}}{4} \tag{19}$$

In the conventional PSO algorithm, both c_1, c_2 is said to be acceleration constant. Here in the proposed algorithm, c_1, c_2 is varied in accordance with the values 0.1, 0.3, 0.5, 0.7 and 1. The PSV-GWO based optimal key selection

Algorithm 2 PSV-GWO based Optimal Key Selection

```
Initialize the grey wolf population $M_i (i = 1, 2, ...., n)$
Initialize $e, H\, and\, E$
Estimate the fitness value for every search agent
$M_\alpha = the\, best\, search\, agent$
$M_\beta = the\ second\, best\, search\, agent$
$M_\delta = the\ third\, best\, search\, agent$
while$(u < \max\ iteration)$
for every search agent
Update the current position of search agent by Equation (19)
end for
Update $e, H\, and\, E$
Estimate the fitness value of entire search agent
Update $M_\alpha, M_\beta\, and\, M_\delta$
$u = u + 1$
end while
Return $M_\alpha$
```

is given by the Algorithm 2, and the pseudo code for the proposed model is illustrated in Figure 6.

5 Results and Discussions

5.1 Experimental Setup

The PSV-GWO algorithm for the preservation of medical data has been implemented in JAVA. Four medical datasets are used for the simulation process that includes Autism-Adolescent dataset, Autism-Child dataset, Cryotherapy dataset and Immunotherapy dataset. Further, the performance of the proposed method is compared over the conventional methods like PSO [21, 26, 28], GA [23], DE [24], CSA [25] and AAP-CSA [27] algorithms on the basis of recovered data. Additionally, on the basis of the different attacks, the simulation was done namely, Known Plaintext Attack (KPA), Known Cipher Attack (KCA), Chosen Plaintext Attack (CPA), and Chosen Cipher Attack (CCA). Moreover, the analysis has been made by the variation in c_1 and c_2 regarding the cost function, and the outcome was thus demonstrated.

E. Analysis on Attack

The various types of attack such as KCA and KPA were analyzed and compared with the existed algorithms, and it is shown in Figure 7. The KCA attack with

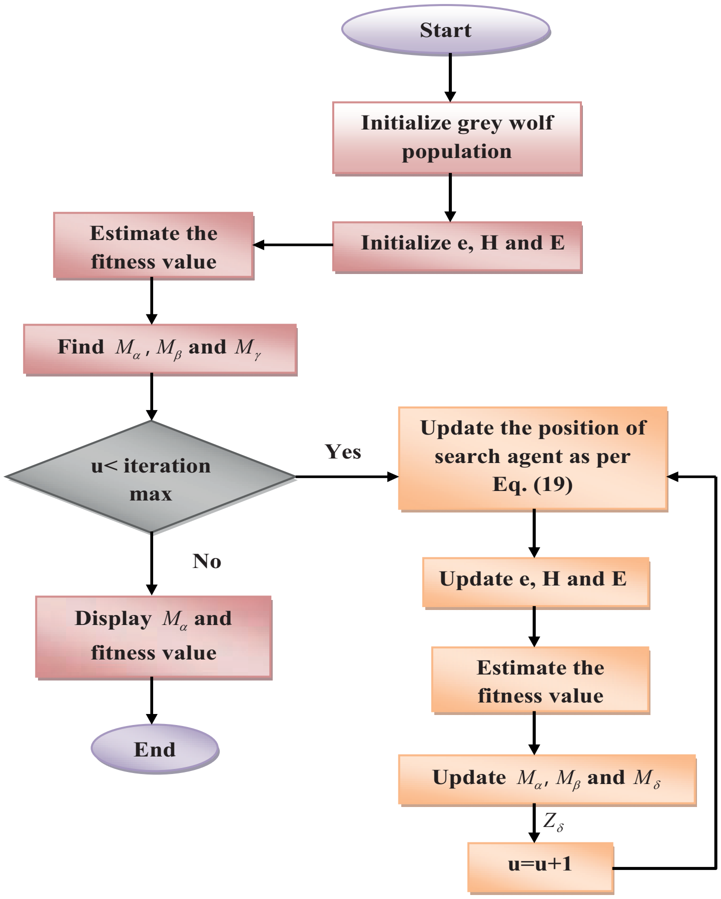
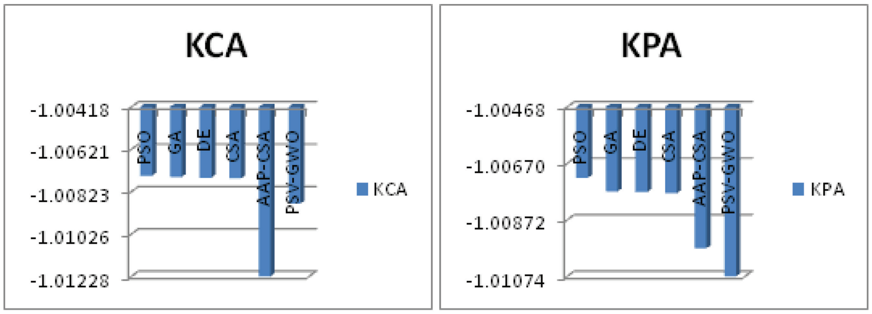

Start
Initialize grey wolf population
Initialize e, H and E
Estimate the fitness value
Find M_α , M_β and M_γ
u< iteration max
Yes
No
Update the position of search agent as per Eq. (19)
Update e, H and E
Estimate the fitness value
Update M_α, M_β and M_δ
Z_δ
u=u+1
Display M_α and fitness value
End

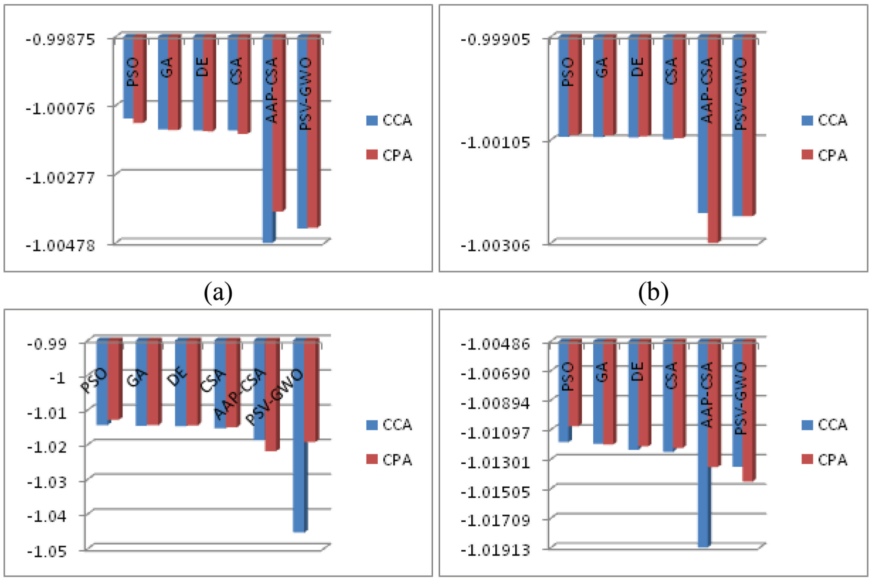
Figure 6 Flowchart for Proposed PSV-GWO Algorithm.

respect to the proposed scheme in Figure 7(a) is 0.13% better than the PSO and GA, also 0.12% better from DE and CSA. From Figure 7(b), the reduction of KPA attack over the proposed model is 0.35% and 0.01% better than PSO and AAP-CSA. It is also 0.30% better than the remaining GA, DE and CSA algorithms. The CCA and CPA attacks on the four datasets are illustrated in Figure 8. In Figure 8(a), the dataset with autism adult with respect to CCA is 0.32%, and 0.30% better from PSO and GA and also 0.29% superior to DE and CSA, respectively. The CPA attack scheme is 0.31%, 0.29%, 0.28%, 0.27% and 0.05% better from PSO, GA, DE, CSA and AAP-CSA, respectively. In autism child dataset with regards to the CCA scheme is 0.15% better



(a) (b)

Figure 7 Analysis on cost function based on attacks (a) KCA attack (b) KPA attack.



(a) (b)

(c) (d)

Figure 8 Analysis on cost function based on attacks (a) CCA attack (b) CPA attack.

than all other conventional methods. The CPA attack under the cryotherapy dataset is 0.63% better than PSO, 0.47% superior to GA and DE and 0.42% better from CSA algorithms. The CCA attack under Immunotherapy dataset in Figure 8(d) is 0.17%, 0.16%, 0.12% and 0.10% better from PSO, GA, DE and CSA, respectively. Hence the analysis shows that the proposed model has an improvement over the existed ones in terms of attacks.

F. Restoration Analysis

The PSV-GWO model restoration process for the four datasetis shown in Tables 3, 4, 5 and 6. In Table 3, the result of the proposed model with respect to the autism adult dataset for C_1 is 99.10%, 99.18%, 99.46%, 95.49%, and 93.68% better than PSO, GA, DE, CSA, and AAP-CSA, respectively. For C_3, the implanted method is 75.81% better from GA and also for F, the implemented model is 38.99% superior to AAP-CSA. Table 4 shows the performance of proposed model over other methods in terms of Autism child dataset. It is observed that for C_1, the proposed method is 99.80%, 99.55%, 99.39%, 99.17% and 99.64% better from PSO, GA, DE, CSA, and AAP-CSA, respectively. For C_3, the proposed method is 41.17% better than GA. From Table 5, the implemented model for autism Cryotherapy dataset in terms of

Table 3 Analysis on Recovery for Autism-Adolescent-Dataset

Functions	PSO [21]	GA [23]	DE [24]	CSA [25]	AAP-CSA [26]	PSV-GWO
C1	5.84210526	6.444444	9.777778	1.166667	0.833333	0.052631579
C2	0.95754499	0.954755	0.927054	0.998615	1.001385	1.008306414
C3	427.050348	2209.028	464.5105	375.2453	488.4015	534.4541889
F	21.1718268	2.703697	23.48654	0.612257	58.01838	35.39705858

Table 4 Analysis on Recovery for Autism-Child-Dataset

Functions	PSO [21]	GA [23]	DE [24]	CSA [25]	AAP-CSA [26]	PSV-GWO
C1	3.34	1.47651	1.087248	0.805369	1.85906	0.006666667
C2	0.94134358	0.988133	0.997827	1.004847	0.978606	1.024899733
C3	1125.3675	2475.185	1405.386	1636.437	1055.604	1456.172067
F	28.2255105	23.32239	61.46219	5.610919	51.53938	74.25923573

Table 5 Analysis on Recovery for Autism-Cryotherapy-Dataset

Functions	PSO [21]	GA [23]	DE [24]	CSA [25]	AAP-CSA [26]	PSV-GWO
C1	0.75	0.090909	0.5	0.090909	1.363636	0.083333333
C2	1.00483871	1.016155	1.017771	1.016155	0.993538	1.017741935
C3	6493.45831	8937.301	7920.009	6019.631	4537.144	8743.042174
F	143.191788	16.53284	343.6985	38.07991	194.4128	119.3211976

Table 6 Analysis on Recovery forAutism-Immunotherapy-Dataset

Functions	PSO [21]	GA [23]	DE [24]	CSA [25]	AAP-CSA [26]	PSV-GWO
C1	0.09090909	0.5	0.1	0.2	0.3	0.090909091
C2	1.0140647	1.007042	1.012676	1.011268	1.009859	1.014064698
C3	12423.538	3930.424	9396.339	8369.21	9603.537	9332.909191
F	144.945618	18.44217	311.7101	9.779016	337.2987	132.8116969

C_1 is 88.89%, 8.33%, 83.33%, 8.33% and 93.89% superior to PSO, GA, DE, CSA, and AAP-CSA, respectively. For C_3, the introduced scheme is 2.17% better than GA. Also for F, the proposed model is 16.67%, 65.28%, and 38.62% better from PSO, DE and AAP-CSE, respectively. Finally, the autism immunotherapy dataset that is illustrated in Table 6 observed that in C_1 the recovery function is 81.81%, 9.09%, 54.54% and 69.70 superior to DA, GE, CSA, and AAP-CSA. And for C_3, the implemented scheme is 24.88%, 0.68%, and 2.82% better from PSO, DE and AAP-CSA, respectively. For F, the proposed method is 9.14%, 57.39%, and 60.62% better than PSO, DE and AAP-CSA, respectively. Hence the result shown that the recovery process in the proposed model reveals an improvement over other conventional methods.

G. Effect on Varying c_1 and c_2

The cost function value by varying the c_1 and c_2 in Equation (17) of PSV-GWO model is illustrated in Figure 9–13 for the four datasets. Figure 9(a) represent

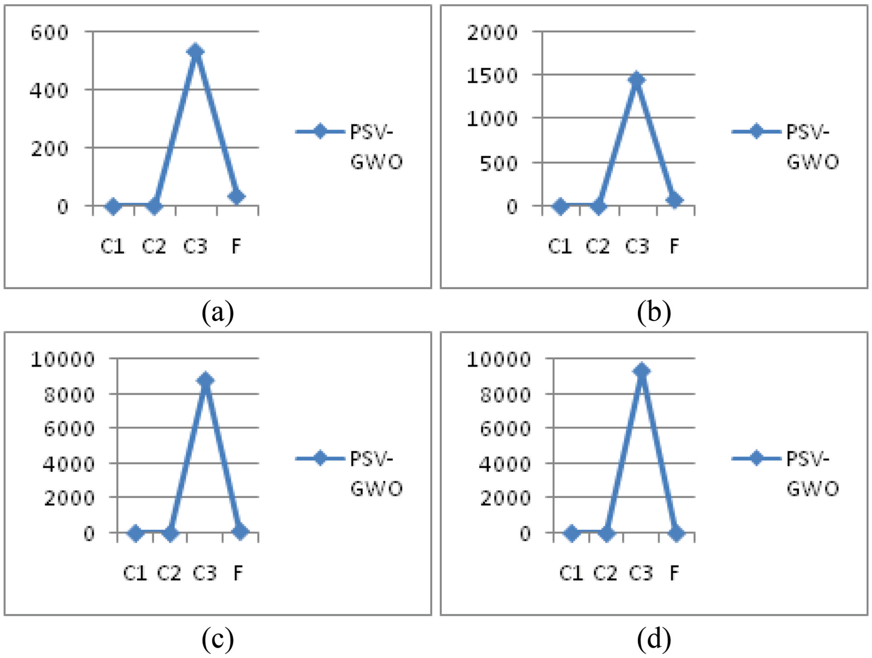


Figure 9 Analysis on cost function on varying $c_1 = 0.1, c_2 = 0.1$ for four datasets (a) autism-adolescent database (b) autism-child database (c) autism-cryotherpy database (d) autism-immunotherapy database.

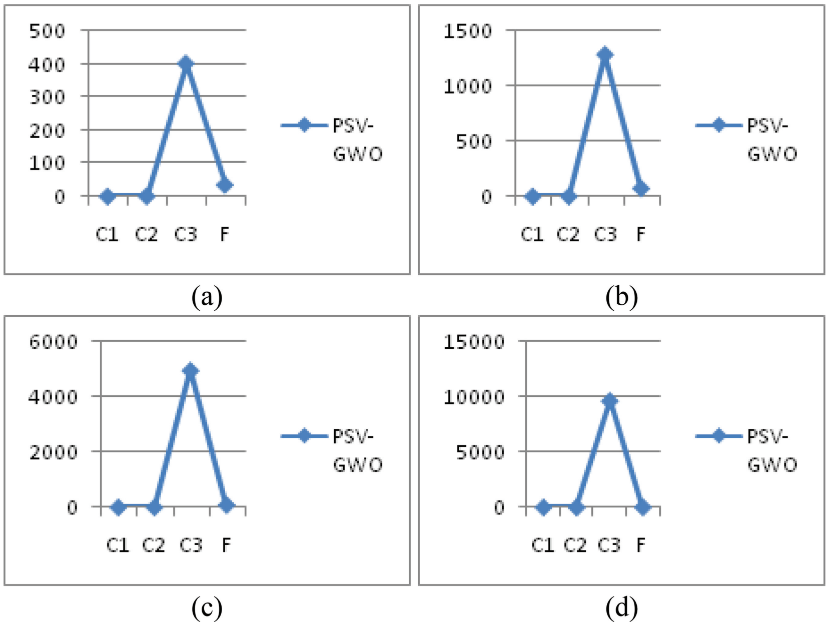
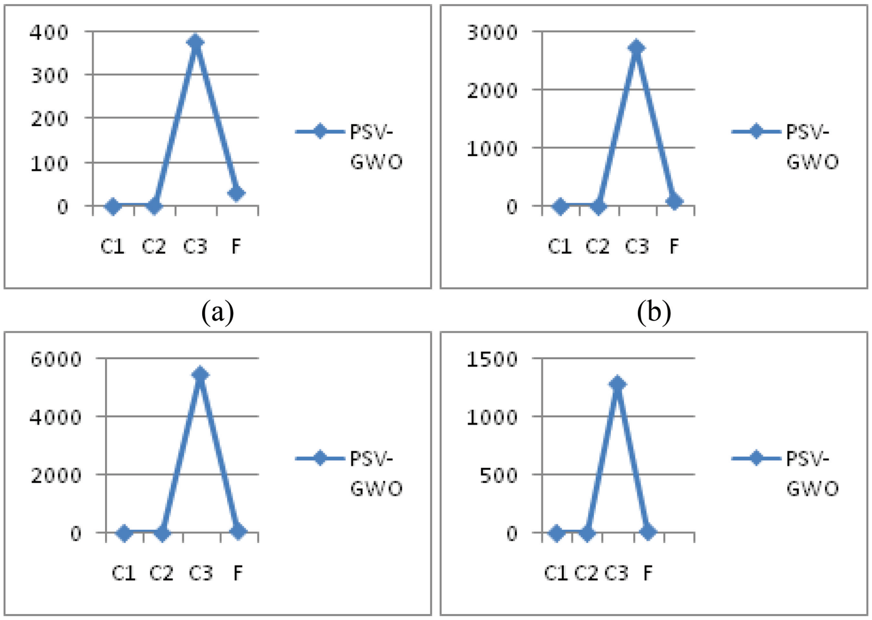


Figure 10 Analysis on cost function on varying $c_1 = 0.3, c_2 = 0.3$ for four datasets (a) autism-adolescent database (b) autism-child database (c) autism-cryotherpy database (d) autism-immunotherapy database.

the proposed model over the autism-adult dataset with respect to $c_1 = 0.1$ and $c_2 = 0.1$, the obtained value of C_1, C_2, C_3 and F is 0.052, 1.008, 534.45 and 35.39, respectively. For the autism-child database with regards to $c_1 = 0.1$ and $c_2 = 0.1$ in Figure 9(b), the attained value of C_1, C_2, C_3 and F is 0.006, 1.02, 1456.17 and 74.25. From Figure 9(c) the Cryotherapy dataset regarding the $c_1 = 0.1$ and $c_2 = 0.1$, the achieved value of C_1, C_2, C_3 and F is 0.083, 1.017, 8743.04 and 119.32, respectively. For the autism-immunotherapy database with regards to $c_1 = 0.1$ and $c_2 = 0.1$ in Figure 9(d), the accomplished value of C_1, C_2, C_3 and F is 0.09, 1.014, 9332.90 and 13.28, respectively.

Additionally, the Figure 10(a) illustrates the implemented model under the autism-adult dataset with respect to$c_1 = 0.3$ and $c_2 = 0.3$, the attained value of C_1, C_2, C_3 and F is 0.052, 1.008, 400.05 and 35.12, respectively. For the autism-child database regarding $c_1 = 0.3$ and $c_2 = 0.3$ in Figure 10(b), the attained value of C_1, C_2, C_3 and F is 0.006, 1.02, 1278.73 and 73.32. From Figure 10(c) the Cryotherapy dataset with regards to the $c_1 = 0.3$ and $c_2 = 0.3$, the obtained value of C_1, C_2, C_3 and F is 0.083, 1.017, 4973.08 and 86.96, respectively. For the immunotherapy database with regards to $c_1 = 0.3$



(a) (b)

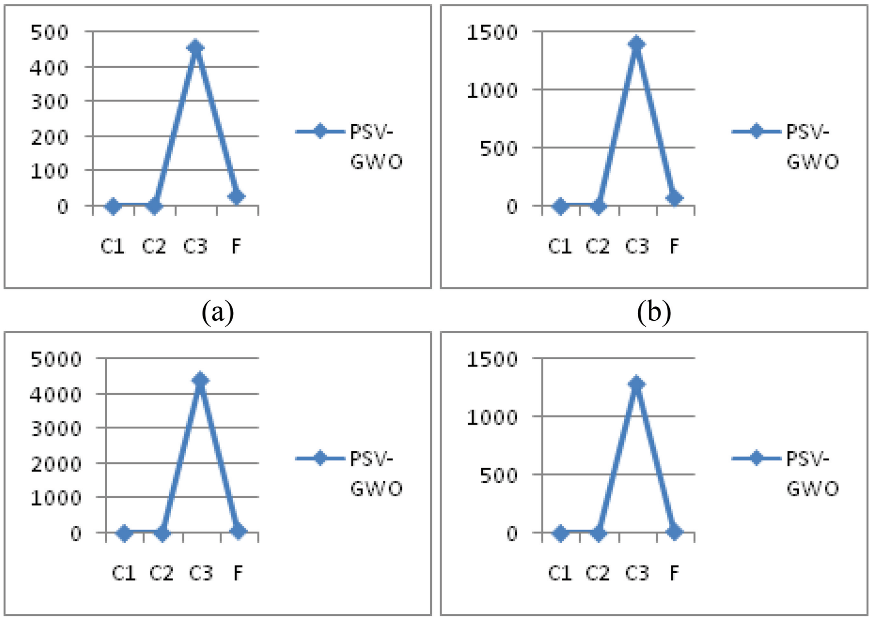
(c) (d)

Figure 11 Analysis on cost function on varying $c_1 = 0.5, c_2 = 0.5$ for four datasets (a) autism-adolescent database (b) autism-child database (c) autism-cryotherpy database (d) autism-immunotherapy database.

and c_2 = 0.3 in Figure 10(d), the gained value of C_1, C_2, C_3 and F is 0.09, 1.014, 9654.89 and 14.83, respectively.

Similarly, the Figure 11(a) symbolize the autism-adult dataset over the proposed model with respect to c_1 = 0.5 and c_2 = 0.5, the gained value of C_1, C_2, C_3 and F is 0.052, 1.008, 377.62 and 31.43, respectively. Further, the autism-child database with respect to c_1 = 0.5 and c_2 = 0.5 in Figure 11(b), the attained value of C_1, C_2, C_3 and F is 0.006, 1.02, 2715.42 and 91.16. From Figure 11(c) the Cryotherapy dataset regarding the c_1 = 0.5 and c_2 = 0.5, the obtained value of C_1, C_2, C_3 and F is 0.083, 1.017, 5437.35 and 76.32, respectively. Moreover, for the autism-immunotherapy database with respect to c_1 = 0.5 and c_2 = 0.5 in Figure 11(d), the attained value of C_1, C_2, C_3 and F is 0.09, 1.014, 1282.73 and 14.99, respectively

On considering the autism-adult dataset in Figure 12(a) for c_1 = 0.7 and c_2 = 0.7, the obtained values of C_1, C_2, C_3 and F are 0.105, 1.007, 456.52 and 29.26, respectively. The Child dataset over the proposed method



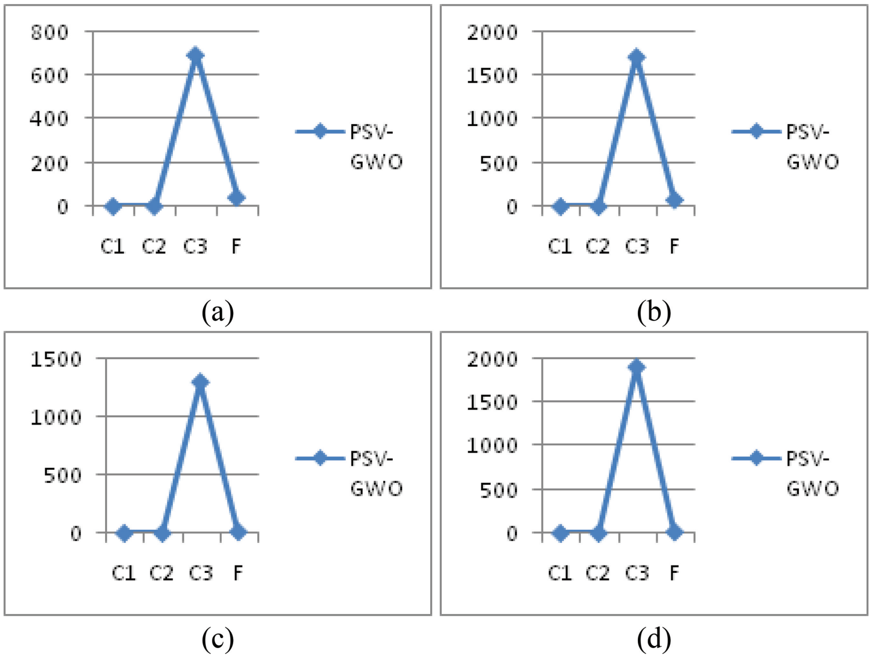
(a) (b)

(c) (d)

Figure 12 Analysis on cost function on varying $c_1 = 0.7, c_2 = 0.7$ for four datasets (a) autism-adolescent database (b) autism-child database (c) autism-cryotherpy database (d) autism-immunotherapy database.

in Figure 12(b) for c_1 = 0.7 and c_2 = 0.7, the gained value is 0.006, 1.024, 1396.38 and 72.57, respectively. The implemented model under the cryotherpy dataset scheme in Figure 12(c) with c_1 = 0.7 and c_2 = 0.7, the accomplished value of C_1, C_2, C_3 and F is 0.08, 1.017, 4394.39 and 72.48, correspondingly. For the immunotherapy dataset in Figure 12(d) with respect to c_1 = 0.7 and c_2 = 0.7, the value that attained is 0.09, 1.014, 1291.94 and 14.92, respectively.

Finally, the Figure 13 explains the four dataset with respect to the CCA and CPA attacks. Here, the autism-adult dataset with regard to c_1 = 1 and c_2 = 1, the value that gained for C_1, C_2, C_3 and F is 0.052, 1.008, 695.18 and 41.18, correspondingly. The proposed model under the child dataset for the optimization function c_1 = 1 and c_2 = 1 is analyzed, the achieved value of C_1, C_2 ,C_3 and F is 0.293, 1.017, 1720.37 and 77.12, respectively. The cryotherpy dataset regarding the c_1 = 1 and c_2 = 1, the attained value is 0.083, 1.017, 1299.17 and 13.51. The immunotherapy over the proposed model is analyzed with c_1 = 1 and c_2 = 1, and the result that obtained is 0.09, 1.014, 1902.36 and 15.76 for C_1, C_2, C_3 and F, respectively. Thus the implemented method was



(a) (b) (c) (d)

Figure 13 Analysis on cost function on varying $c_1 = 1, c_2 = 1$ for four datasets (a) autism-adolescent database (b) autism-child database (c) autism-cryotherpy database (d) autism-immunotherapy database.

effectively solved over various attacks by varying the optimization function c_1 and c_2.

6 Conclusion

In this paper, the privacy preservation method in medical data was implemented. The main focus of this recommended technique was to introduce an effective sanitization process to hide the client's sensitive rules. A key was created for hiding the private healthcare data, which can be chosen optimally with the help of the PSV-GWO algorithm. Further, the resultant sanitized data was recovered securely by the authorized user. Additionally, the result was obtained by comparing the proposed model over traditional algorithms. From the experimental analysis, it is shown that the proposed scheme was evaluated with regards to the various attacks and the result were attained. The analysis over the attacks on the proposed model under the cryotherapy dataset is 0.63% better than PSO, 0.47% superior to GA and DE and 0.42% better from CSA algorithms. The CCA attack under Immunotherapy dataset is 0.17%, 0.16%, 0.12% and 0.10% better from PSO, GA, DE and CSA, respectively. The dataset with autism adult with respect to CCA is 0.32%, and 0.30% better from PSO and GA and also 0.29% superior to DE and CSA, respectively. Thus, the simulation result shows that the implemented method has an efficient performance over the traditional algorithms.

References

[1] Li, Y., Bai, C and Chandan Reddy, K. (2016). A distributed ensemble approach for mining healthcare data under privacy constraints, *Information Sciences*, 330, 245–259.

[2] Ni, L., Li, C., Wang, X., Jiang, H and Yu, J. (2018). DP-MCDBSCAN: Differential Privacy Preserving Multi-Core DBSCAN Clustering for Network User Data, *IEEE Access*, 6, 21053–21063,

[3] Gao, G., Wan, X., Yao, S., Cui, Z., Zhou C and Sun, X. (2017). Reversible data hiding with contrast enhancement and tamper localization for medical images, *Information Sciences*, 385–386, 250–265.

[4] Zhang, L., Liu, Y., Wang, R., Fu, X and Lin, Q. (2017). Efficient privacy-preserving classification construction model with differential privacy technology, *Journal of Systems Engineering and Electronics*, 28(1), 170–178.

[5] Kim, S., Lee, H and DohnChung, Y. (2017). Privacy-preserving data cube for electronic medical records: An experimental evaluation, *International Journal of Medical Informatics*, 97, 33–42.

[6] Fong, P. K and Weber-Jahnke, J. H. (2012). Privacy Preserving Decision Tree Learning Using Unrealized Data Sets, *IEEE Transactions on Knowledge and Data Engineering*, 24(2), 353–364.

[7] Poulis, G., Loukides, G., Skiadopoulos, S and Gkoulalas-Divanis, A. (2017). Anonymizing datasets with demographics and diagnosis codes in the presence of utility constraints, *Journal of Biomedical Informatics*, 65, 76–96.

[8] Xu, L., Jiang, C., Chen, Y., Wang, J and Ren, Y. (2016). A Framework for Categorizing and Applying Privacy-Preservation Techniques in Big Data Mining, *Computer*, 49(2), 54–62.

[9] Bhaduri, K., Stefanski, M. D. and Srivastava, A. N. (2011). Privacy-Preserving Outlier Detection Through Random Nonlinear Data

Distortion, *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 41(1), 260–272.

[10] Zhang, N and Zhao, W. (2007). Privacy-Preserving Data Mining Systems, *Computer*, 40(4), 52–58.

[11] Wang, J., Deng, C and Li, X. (2018). Two Privacy-Preserving Approaches for Publishing Transactional Data Streams, *IEEE Access,* 6, 23648–23658.

[12] Terrovitis, M., Poulis, G., Mamoulis, N and Skiadopoulos, S. (2017). Local Suppression and Splitting Techniques for Privacy Preserving Publication of Trajectories, *IEEE Transactions on Knowledge and Data Engineering*, 29(7),1466–1479.

[13] Ahluwalia, M. V., Gangopadhyay, A., Chen, Z and Yesha, Y. (2017). Target-Based, Privacy Preserving, and Incremental Association Rule Mining, *IEEE Transactions on Services Computing*, 10(4), 633–645.

[14] Lin, K. P and Chen, M. S. (2011). On the Design and Analysis of the Privacy-Preserving SVM Classifier, *IEEE Transactions on Knowledge and Data Engineering,* 23(11), 1704–1717.

[15] Fung, B. C. M., Wang, K. and Yu, P. S. (2007). Anonymizing Classification Data for Privacy Preservation, *IEEE Transactions on Knowledge and Data Engineering,* 19(5), 711–725.

[16] Upadhyay, S., Sharma, C., Sharma, P., Bharadwaj, P., Seeja, K. R. (2016). *Privacy preserving data mining with 3-D rotation transformation, Journal of King Saud University – Computer and Information Sciences,* Available online 28 November 2016.

[17] BALOGLU, U. B., DEMİR, Y. (2018). Lightweight Privacy-Preserving Data Aggregation Scheme for Smart Grid Metering Infrastructure Protection, *International Journal of Critical Infrastructure Protection,* Available online 10 May 2018.

[18] Lin, C. (2016). A reversible data transform algorithm using integer transform for privacy-preserving data mining, *Journal of Systems and Software*, 117, 104–112.

[19] Dhasarathan, C., Thirumal, V., Ponnurangam, D. (2017). A secure data privacy preservation for on-demand cloud service, *Journal of King Saud University – Engineering Sciences*, 29(2), 144–150.

[20] Aldeen, Y. A. A. S., Salleh, M., Aljeroud, Y. (2016). An innovative privacy preserving technique for incremental datasets on cloud computing, *Journal of Biomedical Informatics*, 62, 107–116.

[21] Marini, F., Walczak, B. (2015). Particle swarm optimization (PSO). A tutorial, *Chemometrics and Intelligent Laboratory Systems*, 149, pp. 153–165.

[22] Mirjalili, S., Mirjalili, S. M., Lewis, A. (2014). Grey Wolf Optimizer, *Advances in Engineering Software*, 69, 46–61.

[23] McCal, J. (2005). Genetic algorithms for modelling and optimisation, *Journal of Computational and Applied Mathematics*, 184(1), 205–222.

[24] Zheng, L. M., Zhang, S. X., Tang, K.S., and Zheng, S. Y. (2017). Differential evolution powered by collective information, *Information Sciences*, 399, 13–29.

[25] Askarzadeh, A. (2016). A novel metaheuristic method for solving constrained engineering optimization problems: Crow search algorithm, *Computers & Structures*, 169, 1–12.

[26] Sridhar Mandapati, Dr. Raveendra Babu Bhogapathi and Dr. M. V. P. Chandra Sekhara Rao, (2013). Swarm Optimization Algorithm for Privacy Preserving in Data Mining, *IJCSI International Journal of Computer Science Issues*, Vol. 10(2), 1694–0784.

[27] Mandala, J., Dr. M. V. P. Chandra Sekhara Rao, (2018). Privacy Preservation of Data Using Crow Search with Adaptive Awareness Probability, *in communication*.

[28] G. Kalyani, Dr. M. V. P. Chandra Sekhara Rao, (2017). Particle Swarm Intelligence and Impact Factor-Based Privacy Preserving Association Rule Mining for Balancing Data Utility and Knowledge Privacy, *Arabian Journal for Science and Engineering*, 43(8), 4161–4178.

[29] De Giorgio A., Loscalzo R. M., Ponte M., Padovan A. M., Graceffa G. and Gulotta F. (2016). An innovative mindfulness and educational care approach in an adult patient affected by gastroesophageal reflux: the IARA model, 14(4).

[30] Satish Ramchandra Todmal, Suhas Haribhau Patil. (2015), Optimal Image Watermarking using Hybrid Optimization Algorithm, *International Journal of Computer Vision and Image Processing (IJCVIP)*, 5(1), 27–47.

Biographies

Jyothi Mandala received her B.Tech degree in Computer Science and Information Technology from AITAM College, Tekkali, India and M.Tech degree in Computer Science and Engineering from JNTUK, Kakinada, India. Present she is pursuing her Ph.D. in the area of Privacy Preserving Data. She has 12 years of teaching experience. Currently she is working as Assistant professor in Information Technology Department at GMRIT, Rajam. Her areas of interest are Information Security and Data Mining.

Dr. M. V. P. Chandra Sekhara Rao received his M.Tech degree in Computer Science and Engineering from JNTUK, Kakinada, India and Ph.D from JNTUH, Hyderabad in 2012. He has nearly 22 years of teaching experience. Currently he is working as professor in Computer Science Engineering Department at RVR & JC College of Engineering, Guntur. His areas of interest are Data Warehousing and Data Mining.