Ontology and crow optimization-based deep belief network for privacy preservation of medical data

Rubin Thottupurathu Jose* and Sojan LalPoulse**

**Associate Professor, Amal Jyothi College of Engineering, Kottayam, Kerala, India*
***Principal and Professor, Mar-Baselious Institute of Technology and Science, Kerala, India*
**Corresponding Author: rubinthottupuram@amaljyothi.ac.in*

Submitted: 15/11/2019
Revised: 27/01/2020
Accepted: 03/02/2020

ABSTRACT

Medical data classification is used to find the hidden patterns of data by training a large amount of patient data collected from the providers. As the medical data is very sensitive, it must be a safeguard from all the noncollaborative means. Thus, it is important to take steps to preserve the confidential medical data. Accordingly, this paper proposes a classification method termed as crow optimization-based deep belief neural network (CS-DBN) to preserve the privacy of confidential medical data automatically. This classifier works based on three phases, including generation of the privacy-preserved data, construction of ontology, and classification. The Deep convolutional kernel approach is used to provide data confidentiality using the optimal coefficients. The construction of ontology is done with the cardiac heart disease terms used in the medical field for classification. Finally, the classification is performed using the deep belief network (DBN), which is trained using the crow search algorithm (CSA). The performance is analyzed in terms of the metrics, namely, accuracy, fitness, sensitivity, and specificity. The proposed CS-DBN method produces higher fitness, accuracy, sensitivity, and specificity of 0.9007, 0.8842, 1, and 0.8408, respectively.

Keywords: Medical data, ontology, privacy preservation, optimal coefficients, optimization, deep learning.

INTRODUCTION

Preservation of privacy of data in ontology-oriented systems, like preventing the unauthorised access of system and ontological data, acts as a basic need, specifically with the access of more number of users. In particular, there are large numbers of users who are interested inaccessing and obtaining the data from a single ontology. In this case, various access rights are provided to users, and the privacy preservation implies that the users can only retrieve the information they are allowed to access either directly or indirectly by way of logical inference. The privacy of data in the systems of information is a high area of research, which is active particularly in the case of the databases (DBs) (Stouppa & Studer, 2007). Existing works on the privacy of data in databases aim mainly at the complete relational DBs (Biskup & Bonatti, 2004; VitoRacanelli et al., 2006). Ontologies are highly coordinated with the DBs that are incomplete (Biskup & Weibert, 2008; Levy, 1996), with the variation that the languages of ontology are normally very expressive than the schema languages of DB. The privacy of data in case of the context of incomplete or semi-structured DBs has been analyzed in recent years (Fan et al., 2004). In addition, these works do not make any consideration about the presence of dependencies that are complex, like the ones that are in OWL ontologies. It is needed that the privacy-oriented issues may act considerably important as the ontology-oriented methods are hybridized in the mainstream applications.

Ontologies that are expressed in Web Ontology Language (OWL) are used in areas, such as astronomy, biomedicine, and defence. OWL ontologies are utilized to define the data meaning formally, for instance, electronic records of the patient in medical application. Applications then show the ontologies to process the related information more effectively. For instance, a medical ontology showing the patient data record may comprise the data, like "every patient with a mental disorder needed to be treated by a psychiatrist," "schizophrenia is a kind of psychosis," and "psychosis is a kind of mental disorder"; if a person's record related to health shows that he is suffering from schizophrenia, then ontology is utilized to finalize that the person is affected with a disease and needed to be examined by a psychiatrist (Grau, 2010). In ontology-based methods, the patients and their relations are considered to represent the medical information. Recently, ontology plays an important role in the number of applications to improve classification, data privacy, retrieval and extraction of information, fitness, and so on (Pramod P Jadhav and SD Joshi, 2019). The conventional classification techniques of data contain various difficulties, such as infeasibility in the case of large-scale distributed systems to share the datasets of individuals for checking the similarity of data and the leakage of private data about an entity. Thus, there occurs a requirement for the method of classification of data for the preservation of the data that is confidential (Karlekar & Gomathi, 2018).

Object classification acts as an important research area and is used in practical applications in a variety of fields, such as statistics, medicine, pattern recognition, and artificial intelligence (Keller *et al.,* 1985; Dasarathy, 1980). These methods work based on data encryption for the security of data. For the encryption of the entire data, it is very costly in terms of time requirements and memory requirements. It is important to divide the sensitive data initially from public data and then encrypt only the required sensitive data. There are various machine learning methods for the process of classification (Arul.V.H. et al., 2019). However, in various cases of pattern recognition, the input pattern classification relies on the data, where the sample size of all the respective classes is small. In some cases, the sample may not be the representative of the actual probability distribution, even if it is known. In those cases, there are many methods that are based on similarity and distance in the feature space, like clustering and discriminate analysis (Duba & Hart, 1973). Mostly, the K-NN algorithm is used to classify the instances, as it is the simplest clustering method. This rule offers a simple nonparametric flow to assign the class label to the input depending on the class labels indicated by the K-NN of the vector. The K-NN is more capable for the problem characterised with data that is partially shown to the system before usage (Whitney & Dwyer, 1966; Zardari *et al.,* 2014).

The need for this research is to model a method for privacy preservation and classification of the medical data. Initially, the medical data related to patients are obtained from the data providers (hospitals). The proposed CS-DBN method of privacy protection works based on three phases, such as the generation of the privacy preserved data, construction of ontology, and classification. The deep convolutional kernel approach is used to obtain the confidentiality of data with the generation of optimal coefficients. Then, the construction of ontology is carried out with the analysis of cardiac heart disease terms used in the medical field for classification. The construction of ontology is done to produce the clinical decision related to cardiac heart disease, and the classification is performed using the DBN (Vojt, 2016), which is trained using the CSA (Askarzadeh, 2016).

The main contribution of the paper is as follows.

Privacy preservation and medical data classification using CS-DBN: The proposed CS-DBN is the method of privacy protection in medical data using CSA that involves training the DBN to improve the accuracy in classification.

The paper is organized as follows. The introduction to the need for preservation and classification of medical data is detailed in section 1, and section 2 details the literature survey of the conventional methods involving medical data classification and their limitations. In section 3, the proposed CS-DBN is presented, and section 4 details the results of the CS-DBN method. Finally, section 5 provides the conclusion of the paper.

MOTIVATION

In this section, the literature review of various techniques used for medical data classification is presented, and the limitations of the conventional methods are discussed.

Literature Survey

The eight literatures related to the medical data classification are discussed. Ming Tao *et al.* (2018) developed a multilayer cloud architectural model in order to create the smart home applications useful and to be capable of solving the problem of heterogeneity, but it was possible to send the fake request by the attackers from outside. In the proposed method, for each piece of data, a key is generated to improve the privacy of data. Li Zhou *et al.* (2018) modelled the skeletal ontology for eco-industrial parks that was a more efficient and reliable information management system. When a new application case arises, the framework of the ontology was needed to be efficient in dealing with those modifications, and thus, this framework cannot be considered as a final framework. In the proposed system, the ontology is applied for the selection of features, so it did not affect the framework. Nandkishor P. Karlekar and N Gomathi (2018) designed an ontology and whale optimization-based support vector machine (OW-SVM) method, which outperformed the conventional methods, but the databases that were used in this method contain many missing values. In the proposed method, the analysis is performed using the dataset taken from the heart disease dataset of UCI machine learning repository, and the nature of this database is that the attributes are grouped as categorical, real, multivariate, and integer. Peter Geibel *et al.* (2015) modelled the clinical research data warehouse (CRDW) that involved the identification of patient with the extraction of features from the unstructured data. The false positives were needed to be eliminated for the improvement of precision and specificity, which was very tedious and considered as the major drawback of this method. The proposed method works by following three main phases, such as the generation of privacy-preserved data, construction of ontology, and classification to improve the accuracy, precision, and specificity of the data. Jorge Bernal Bernabe *et al.* (2015) developed the Security Ontology for the InterCloud (SOFIC) that provided a trust model for quantifying the trust indexes efficiently based on historical assessments, but there was a need for the CSPs to be interoperable in this method. To overcome this problem, the proposed method uses synthetic data generators that produce synthetic data, which look similar to the original data. Munwar Ali Zardari *et al.* (2014) developed the K-NN data classification method, which was appropriate without knowing the security requirements of the data but produced less accuracy. In the proposed system, the fitness is evaluated by using gradient descent algorithm, which improves the accuracy. Nandkishor P. Karlekar and N. Gomathi (2017) modelled the privacy preservation of cloud data with the use of Kronecker product and Bat algorithm-based generation of the coefficient that outperformed the conventional methods in terms of DBDR. For enhancing the performance, the bat algorithm is needed to be replaced with some other hybrid algorithms, which was a tedious process and considered as the major limitation of this method. Kamran Farooq and Amir Hussain (2016) designed the ontology-driven clinical risk assessment and recommendation system (ODCRARS) and the machine learning-driven prognostic system (MLDPS) that helped the clinicians differentiate acute angina/cardiac chest pain patients form the patients suffering from other causes of chest pain in an effective way. The major drawback of this method was the need for high computational time for the completion of the identification process. In the proposed method, the DBN is trained by using the CSA algorithm, which uses a limited number of parameters. So, the processing speed of the proposed system is high.

Challenges

The various challenges of this research are detailed as follows:

- Crypto coprocessor is a machine that offers security as a service (Ram & Sreenivaasan, 2010) on-demand, which is controlled using a third party. It permits the users to choose the method of encryption for the encryption of data and partition the data as various fixed chunks, but the drawback of this method is that the end-user is not capable of choosing the powerful method for the encryption of data (Zardari *et al.,* 2014).
- The storage model of the inner-cloud is highly trustworthy and reliable when compared to storage model of single cloud (Cachin & Haas, 2010), but this kind of sharing the keys may result in a key challenge with the

unavailability of one cloud (Zardari *et al.,* 2014).

- There are various machine learning methods that were developed and analyzed for the process of classification. However, in some problems of pattern recognition, the classification depends on data with small sample sizes of each class (Keller et al., 1985; Zardari et al., 2014).
- A hybrid clinical decision support method (Farooq & Hussain, 2016) consists of two main parts, such as recommendation system and machine learning-driven prognostic system, and ontology-driven clinical risk assessment. This method could solve the inaccuracies that are present in the clinical risk assessment, but its validation is limited (Karlekar & Gomathi, 2018).
- A security-oriented multikeyword ranked search method (Xia et al., 2016) was encrypted with operations, such as deletion and insertion of documents in a simultaneous manner. Even though this method is used to offer security, there arise a number of issues related to security that need an index to be rebuilt (Karlekar & Gomathi, 2018).

PROPOSED CS-DBN METHOD OF PRIVACY PROTECTION IN MEDICAL DATA

When data related to a patient is transferred, it passes over a security mechanism, like data encryption, without considering the data features or directly stored on the servers without performing the process of encryption. All data possess various types of sensitivity levels, and so it is nontechnical to transfer data just by knowing its security needs. The main aim of this research is to design a method of privacy preservation of the medical data of patients and to classify the medical data. The medical data of the patients are initially collected from the data providers. This method works by following three main phases, that is, the generation of privacy-preserved data, construction of ontology, and classification. Deep convolutional kernel approach is used to obtain the confidentiality of data with the production of optimal coefficients. The construction of ontology is performed with the analysis of terms related to cardiac heart disease and is then used for classification, which is performed using DBN that is trained with the CSA. The block diagram of CS-DBN method of medical data classification is depicted in Figure 1.

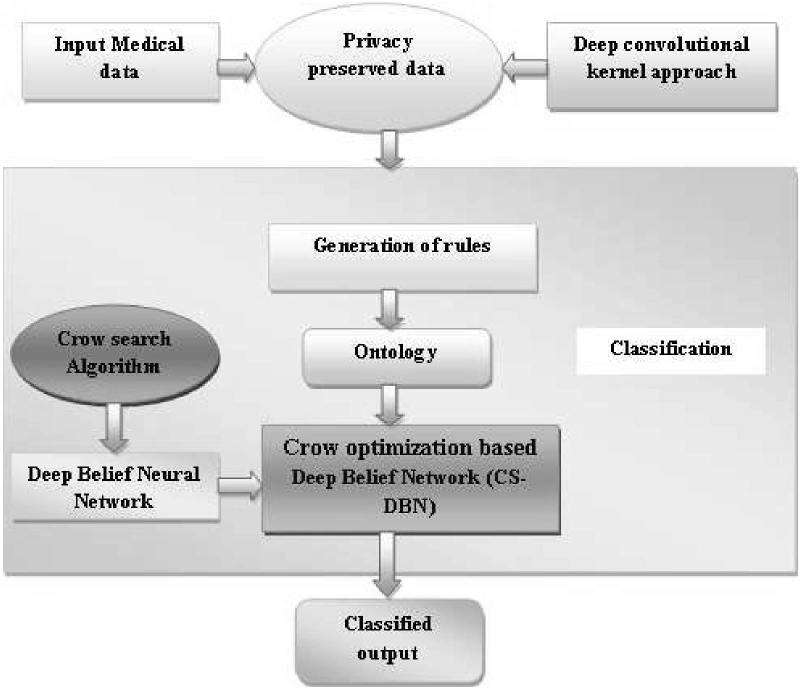


Figure1. Block diagram of the proposed technique of medical data preservation.

Deep convolutional kernel approach for privacy preservation of the medical data

There is a need for the preservation of privacy in case of medical data to safeguard the data related to a patient. If not, then the customers drop their trust and hesitate to use cloud computing. Thus, the protection of privacy is an important concern, which means that the patient's personal data, called as sensitive information, is preserved during the publication of data. There are three major types of privacy protection methods. The first one is the perturbative methods that create a certain type of change into each element of the original patient data. The second method is the generalization method that swaps the original values with less precise ones, and the third one is the synthetic data generators that produce the synthetic data, which look similar to the original data. The expression of the privacy-preserved data is given as

$$C^{*}_{s_{y\times z}} = \left(C_{y\times z} \otimes Y_{s_{z\times z}}\right) \times u_{s_{(1\times 1)}} \tag{1}$$

where C^{*}_{s} is the privacy-protected data, C is the original input data, Y_s is the kernel size of $z \times z$, $\otimes$ is the convolution, $\times$ is the element-wise multiplication, and u_s is the exponential kernel function. The term exponential kernel function is expressed as

$$u_s = \exp(\bullet) \tag{2}$$

In the eq. (1), for the generation of the privacy-protected data, the original data is multiplied with the kernel, followed by the element-wise multiplication with the keys. Each of the keys generated is applied for acquiring the privacy-protected data, and the output after applying the keys to the original data in the intent to ensure security to the data is expressed as

$$C_{OUT_{y\times z}} = C^{*}_{1_{y\times z}} + C^{*}_{2_{y\times z}} + \ldots + C^{*}_{y_{y\times z}} \tag{3}$$

where $C^{*}_{1_{y\times z}}, C^{*}_{2_{y\times z}}, \ldots, C^{*}_{y_{y\times z}}$ are the individual outputs generated after applying the keys, y is the number of deep layers, and + is the element-wise matrix addition.

Determination of keys: For each piece of data, a key is needed to be generated, and thus, for y number of data, y number of keys is generated. The generation of keys is performed to enable the privacy of the data in an effective way. The generated keys are in the dimension of $F = (z \times z) \times y$, where F represents the kernels, and z represents the generated keys. Each of the keys is applied to generate the privacy-preserved data.

Feature selection using ontology

The feature selection concept to develop the rules of ontology for medical data preservation is detailed in this section. Depending on the generated u_s, matching is performed to select the best feature so as to reduce the dimension. To create a clinical decision, ontology is constructed, and then it is used with the DBN to perform classification. Ontology is a formal knowledge that states the relations and concepts used in the classification of the medical data with the extracted features. Once the dataset is loaded, the most relevant features are extracted, in which the constructed ontology rules are applied to the selection of features. The commonly considered features of a patient include the patients' age, sex, diagnosis, symptoms, findings, and the therapies given to the patient. The feature selection using ontology is made based on the expression given as follows:

$$O(Feature_o) = \frac{1}{y}\sum_{s=1}^{y} u_{so}(mat)O_{ONT} \tag{4}$$

where $O(Feature_o)$ is the feature selection using ontology, y is the number of keys, u_s is the exponential kernel function, O_{ONT} is the ontology rule, and *mat* is the term denoted for matching with kernel functions.

$$u_{so}(mat)O_{ONT} = \begin{cases} 1, & if \ \ matched \\ 0 & otherwise \end{cases} \tag{5}$$

When the exponential kernel function matches the rules of ontology, the output is produced as ‘1,’ and if not, the output is produced as ‘0.’ Based on this output, the features are selected. While developing the ontology, there is a need to focus on observations, like symptoms, diagnoses, and findings, and the therapies with medications. Among the obtained features, select the top *Y* number of features to be fed as the input of the CSA algorithm that trains the DBN to classify the medical data.

Data classification using CS-DBN

The objective of the proposed CS-DBN is the classification of the medical data in an optimal and effective way that effectively supports decision-making. Here, the weights of the DBN classifier are tuned by the CSA, which is used in the optimization of constrained engineering design problems with different objectives, and decision variables that lead to estimate improved results. The CSA is a population-based metaheuristic method, it is very simple, and it uses limited numbers of parameters to find the optimal weights. Also, it solves complex engineering optimization problems. Hence, we use CSA to finds the weights of DBN classifier.

Computation of the fitness measure:

The fitness function decides the solutions of the problem, and the CSA tries to simulate the characteristics of the algorithm to select the best solution to the optimization problem, which is the optimal weight for training the classifier. To maximize the following objective, apply the gradient descent algorithm as follows:

$$U = \underset{F}{Arg} \ Max[(X(F))] \tag{6}$$

The fitness function $X(F)$ is expressed as a function of accuracy that is given as

$$X(F) = \frac{1}{3}[\tau + \sigma + X] \tag{7}$$

The term τ depends on the distance function and the normalization function and is expressed as

$$\tau = \frac{d(C, C^*)}{N_R} \tag{8}$$

where σ is the accuracy, $d(\bullet)$ represents the distance function, and N_R indicates the normalization function. The objective function is given as

$$X = \frac{1}{z \times y} \sum_{s=1}^{y} \sum_{o=1}^{z} u_{so}(mat)O_{ONT} \tag{9}$$

where O_{ONT} is the rules of ontology.

Algorithmic steps of CSA algorithm: The objective of the CSA algorithm is to provide the privacy protection for medical data in an effective way. Crows are the most brilliant birds, and the size of the brain of the crow is slightly less than that of the human brain, and the proof to state the ability of the crows is limitless. The crows possess self-awareness and the capability of tool-making and are able to remember the faces and may warn other crows under any unfavourable incident. The crows make use of the tools and communicate with other crows in a matured way and can remember the place of their hiding food even after several months. Crows usually watch other crows to know the hiding place of food and then take it if the other crow is not in its place. If a crow takes the food of other crows, it moves to a hiding place to avoid being caught by other crows. The crows make use of their own experience to know the safest place after taking the food to protect it.

While considering a flock of the crow, most of the characteristics of the crow match with the optimization process. Depending on these characteristics, the crows search for a hiding place to hide the food in excess and then make use of it whenever needed. The crows act selfishly as they follow each other to know about the better source of food. It is a tedious task for a crow to know the hidden food source of another crow, when some other crow follows it. If it occurs, then the crow fools the other by modifying its direction. In terms of optimization, the crows behave as searchers, the environment acts as the search space, and each position of the environment is responsible for a feasible solution. The food source quality is considered as the objective function, and the best food source, which is present in the environment, is considered as the global solution. Based on the similar factors, the CSA tries to make use of the brilliant characteristics of the crows to obtain the best solution to the problem. The standard relation of the CSA is expressed as

$$D_{i,j}^{t+1} = D_{i,j}^{t} + g_i \times H_{i,j}^{t} \times \left(l_p^t - D_{i,j}^{t} \right) \tag{10}$$

where $D_{i,j}^{t+1}$ is the position of the i^{th} crow in j^{th} dimension at the $(t+1)^{th}$ iteration, $D_{i,j}^{t}$ is the position of the i^{th} crow in j^{th} dimension at the t^{th} iteration, g_i is the random number that varies between '0' and '1,' l_p^t is the memory of p^{th} crow in the t^{th} iteration, and $H_{i,j}^{t}$ is the flight length of i^{th} crow at the t^{th} iteration.

Algorithmic steps of the CSA algorithm

Step 1: Initialization of problem and adjustable parameters: The constraints, optimization problem, and the decision variables $'j'$ that are utilized in the solution of the problem are initialized. The adjustable parameters that are used in CSA, such as the size of flock S, number of crows n, flight length $H_{i,j}^{t}$, the awareness probability P_N, and the maximum number of iterations $K_{\max}$, are also defined.

Step 2: Memory and position initialization of the crows: The n crows of flock S are positioned in a random manner in a search space of j dimension, where each of the crows represents a feasible solution. The position of the crows is given by

$$crows = \begin{bmatrix} D_1^1 & D_2^1 \; ... \; D_j^1 \\ D_1^2 & D_2^2 \; ... \; D_j^2 \\ . & . \;\; . \;\; . \\ . & . \;\; . \;\; . \\ . & . \;\; . \;\; . \\ D_1^S & D_2^S \; ... \; D_j^S \end{bmatrix} \tag{11}$$

The memory of each crow $'v'$ is initialized. During the initial iteration, the crow does not have any experiences and hence assumed to have hidden the foods at the initial positions.

$$memory = \begin{bmatrix} v_1^1 & v_2^1 \; ... v_j^1 \\ v_1^2 & v_2^2 \; ... \; v_j^2 \\ . & . \;\; . \;\; . \\ . & . \;\; . \;\; . \\ . & . \;\; . \;\; . \\ v_1^S & v_2^S \; ... \; v_j^S \end{bmatrix} \tag{12}$$

Step 3: Estimation of fitness function: For all the crows, the quality of position is estimated with the decision variable of the objective function. The fitness measure is calculated based on equation (7).

Step 4: Generation of new position: Each crow generates a new position within the search space. For instance, if the crow *j* wants to produce a new position, it selects one *d* among the flock in a random manner and keeps following it to obtain the food location that is hidden by the crow *d*. Equation (10) provides the new position of the crow, and this process is continued for all the crows that are left in the solution space.

Step 5: Feasibility checking for new positions: The feasibility of each new position for all the crows is checked. For feasible new positions, the crows update their position, and if not, then the crows do not change their position to the new position, and then they remain in the old position.

Step 6: Calculation of fitness for new positions: The fitness for each new position of the crow is determined based on equation (7).

Step 7: Update of memory:For each crow, the memory is updated using

$$v_{i,j}^{t+1} = \begin{cases} D_{i,j}^{t+1} & w\left(D_{i,j}^{t+1}\right) \textit{ is better than } v_{i,j}^{t} \\ v_{i,j}^{t} & \textit{otherwise} \end{cases} \tag{13}$$

where $w\left(D_{i,j}^{t+1}\right)$ represents the value of the objective function. If the fitness of the new position is found to be better compared to the fitness of the memorized position, then the crow memory is updated using the estimated new position.

Step 8: Stopping criteria: Repeat the above steps up to the maximum iteration. If the stopping condition is satisfied, the best memory position is considered as the optimized solution. The pseudocode of the CSA algorithm is depicted in algorithm 1.

Algorithm1. Pseudocode of CSAalgorithm.

```
CSA algorithm
Input: Solution vector → $D_{i,j}$
Output: Best solution → $D_{i,j}^{t+1}$
Start
Initialization of parameter
$n$: Maximum crows → total number of crows
decision variables → $j$
size of flock → $S$
flight length → $H_{i,j}^{t}$
Awareness probability → $P_N$
maximum number of iterations → $K_{\max}$
$rand$ → random number
Calculate the position of each crow
Initialize the memory of all the crows in the flock $S$
For $i = 1 : n$
```

```
    Select crow $i$ to follow crow $d$ in random manner
    If
        rand of $p^{th}$ crow ≥ $P_N$
        Estimate the position of crow using equation (10)
    Else
        Find a random position
    End if
End for
Perform feasibility check
Estimate the new position of crow
Update crow memory
Stop
```

Architecture of Deep Belief Network: CSA trains the DBN in order to find the weights of the DBN so as to classify the medical data of a patient. The basic structure of DBN comprises one multilayer perceptron (MLP) and the number of restricted Boltzmann machines (RBMs). The individual layers of RBM and MLP represent the structure of NN, and the individual layers are developed with the interconnection of neurons. In DBN, two RBMs are considered, and the input to the RBM1 is the feature vector corresponding to the medical data. The inputs are multiplied with the input neuron weights to obtain the output of the hidden layer that obtains the input of RBM2. The inputs in RBM2 are processed with the hidden weights of RBM2 to obtain the input to the MLP layer that processes the weights and produces the final output. The weights of the DBN are estimated using the CSA algorithm, and the structure of the DBN is shown in figure 3.

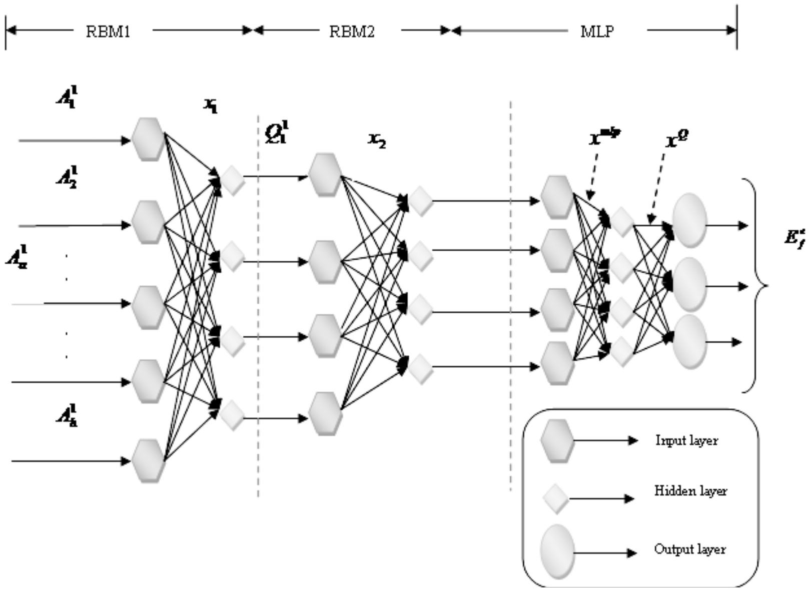


Figure 2. Architecture of incremental DBN model.

The mathematical model of the DBN is described as follows. Consider two RBMs, namely, RBM1 and RBM2, and the input to the RBM1 is the feature vector of the medical data. The input and hidden neurons in the input layer of RBM1 are expressed as

$$A^1 = \{A_1^1, A_2^1, A_3^1, \ldots, A_a^1, \ldots, A_h^1\}; \quad 1 \le a \le h \tag{14}$$

$$B^1 = \{B_1^1, B_2^1, \ldots, B_b^1, \ldots, B_k^1\}; 1 \le b \le k \tag{15}$$

where A_a^1 is the a^{th} input neuron present in RBM1, and the number of the input neurons in RBM1 equals the feature vector dimension. There are h neurons in the input layer of RBM1. Let us indicate that the total number of the hidden neurons in the RBM1 be k and let the j^{th} hidden neuron in RBM2 be indicated as B_b^1. Let the biases in the visible and the hidden neurons of RBM1 begiven as

$$J^1 = \{J_1^1, J_2^1, J_3^1, \ldots, J_a^1, \ldots, J_h^1\} \tag{16}$$

$$L^1 = \{L_1^1, L_2^1, \ldots, L_b^1, \ldots, L_k^1\} \tag{17}$$

The biases in the hidden and input layer of RBM1 equal the total neurons in both layers, and the weight of the RBM1 is expressed as

$$x^1 = \left\{x_{ab}^1\right\}; 1 \le a \le h;\ 1 \le b \le k \tag{18}$$

where x_{ab}^1 is the weight of the RBM1, and it is the weight between the a^{th} input neuron and b^{th} hidden neuron of RBM1. The dimension of weight is indicated as $(h \times k)$. Hence, the output from RBM1 is

$$Q_b^1 = \alpha\left[L_b^1 + \sum_a T_a^1\, x_{ab}^1\right] \tag{19}$$

where α is the activation function in RBM1 and T_a^1 is the feature vector as in equation (4). The output from RBM1 is expressed as

$$Q^1 = \{Q_b^1\}; 1 \le b \le k \tag{20}$$

The output from RBM1 is fed as the input to the RBM2, and the output of RBM2 is estimated depending on the equations given above. The output of the RBM2 is indicated as Q_b^2 and is fed as input to MLP layer. The input neurons in MLP are expressed as

$$G^e = \{G_1^e, G_2^e, \ldots, G_b^e, \ldots, G_k^e\} = \{M_b^2\}; 1 \le b \le k \tag{21}$$

where k represents the total number of input neurons in the MLP layer. The hidden neurons of MLP are expressed as

$$I^e = \{I_1^e, I_2^e, \ldots, I_c^e, \ldots, I_q^e\}; \quad 1 \le c \le q \tag{22}$$

where q is the total number of hidden neurons in the MLP. The bias of the hidden neurons is expressed as

$$E^e = \{E_1^e, E_2^e, \ldots, E_f^e, \ldots, E_V^e\}; 1 \le f \le V \tag{23}$$

where V represents the number of output neurons in the MLP layer. The weight among the input and the hidden layers is indicated as

$$x^{mlp} = \{x_{bc}^{mlp}\}; 1 \le b \le k; 1 \le c \le q \tag{24}$$

where x_{bc}^{mlp} represents the weight vector between b^{th} input neuron and c^{th} hidden neuron. The output of the hidden layer in MLP relies on the bias and weights, and the output is expressed as

$$R = \left[\sum_{b=1}^{k} x_{bc}^{mlp} \times W_b \right] x_c^e \ \ \forall W_b = Q_b^{\ 2} \tag{25}$$

where x_c^e represents the bias of the output layer. The weight vector among the hidden and the output layer is indicated as x^Q and is expressed by

$$x^Q = \{x_{cf}^Q\}; 1 \le c \le q; 1 \le f \le V \tag{26}$$

Thus, the output of MLP is estimated as

$$E_f = \sum_{c=1}^{q} x_{cf}^Q \times R \tag{27}$$

where x_{cf}^Q is the weight among the hidden and output neurons in MLP, and R is the output of the hidden layer.

a) Training phase of RBM layers: The training of the RBM1 and RBM2 is carried out depending on the CSA algorithm that determines the weights based on the minimum error.

b) Training of MLP layer:The steps involved in the training of the MLP layer are described as follows:

Step 1: Produce the weight vectors x^Q and x^{mlp} randomly as in equations (26) and (24), respectively.

Step 2: Read the input vector $Q_b^{\ 2}$ developed from the output layer of RBM2.

Step 3: Estimate R and E_f depending on equations (25) and (27), respectively.

Step 4: Evaluate the fitness function in terms of using equation (7).

Step 5: Update the weight using the CSA algorithm: The weights of the MLP layers are updated using equation (10), which is the weight derived using the CSA algorithm. The CSA algorithm derives the optimal weights for the classification of medical data.

Step 7: Estimate the average error function $\beta^1{}_{avg}$ with the weight vector that is updated with the CSA algorithm.

Step 8: Continue the steps from 2 to 7, until the best weight vector is obtained.

Thus, the classification is performed by the DBN with the construction of ontology using the terms related to cardiac heart disease. The fitness function is selected using the kernel space of DBN, and the parameters in the kernel are selected using the CSA.

RESULTS AND DISCUSSION

The results and discussion of CS-DBN classifier are detailed in this section. The results of CS-DBN classifier, when compared with the conventional classifiers of medical data based on accuracy, sensitivity, and specificity, are provided.

Experimental setup

The experimentation of the proposed CS-DBN method is performed in JAVA that runs in PC with Windows 8 OS. Table 1 discusses the simulation setup of the proposed system.

Table 1. Simulation setup.

Parameter	Value
Hidden Layers	5
Input layers	3
Number of neurons in Input Layer	50
Number of neuronsin Hidden Layer	20
Learning Rate	0.2

Dataset description

The analysis is performed using the dataset taken from the heart disease dataset of UCI machine learning repository https://archive.ics.uci.edu/ml/datasets/Heart+Disease.The heart disease database consists of databases like Switzerland, Hungary, and Cleveland, and VA long beach, out of which three databases, namely, Cleveland, Hungary, and Switzerland, are used for experimentation. The nature of the database and its attributes are grouped as categorical, real, multivariate, and integer with 303 instances and 75 attributes.

Evaluation metrics

The performance of CS-DBN algorithm is analyzed in terms of the evaluation metrics, namely, fitness, accuracy, sensitivity, and specificity.

Fitness: The fitness measure is estimated based on equation (7) as discussed in section 3.3.1.

Accuracy: The result, which verifies the level of exactness, is termed as accuracy and is given as

$$\text{Accuracy}=\frac{True\ positive+True\ negative}{True\ positive+True\ negative+false\ positive+false\ negative} \tag{28}$$

Sensitivity: The sensitivity is, otherwise, termed as true positive rate (TPR), and it is termed as the count of positives, which are identified correctly.

$$TPR=\frac{True\ positive}{True\ positive+False\ negative} \tag{29}$$

Specificity: The specificity is, otherwise, termed as true negative rate (TNR), and it is termed as the number of negatives, which are identified correctly.

$$TNR=\frac{True\ negative}{True\ negative+False\ positive} \tag{30}$$

Comparative methods of privacy protection in medical data

Various conventional techniques, namely, Decision Tree (Alabdulkarim et al., 2019) , Naive Bayes (Al-Aidaroo et al., 2012) , K-Nearest Neighbour (K-NN) (Zardari et al., 2014), and the Support Vector Machine (SVM) (Karlekar & Gomathi, 2018), are compared with CS-DBN classifier in terms of the evaluation metrics, namely, fitness, sensitivity, accuracy, and specificity.

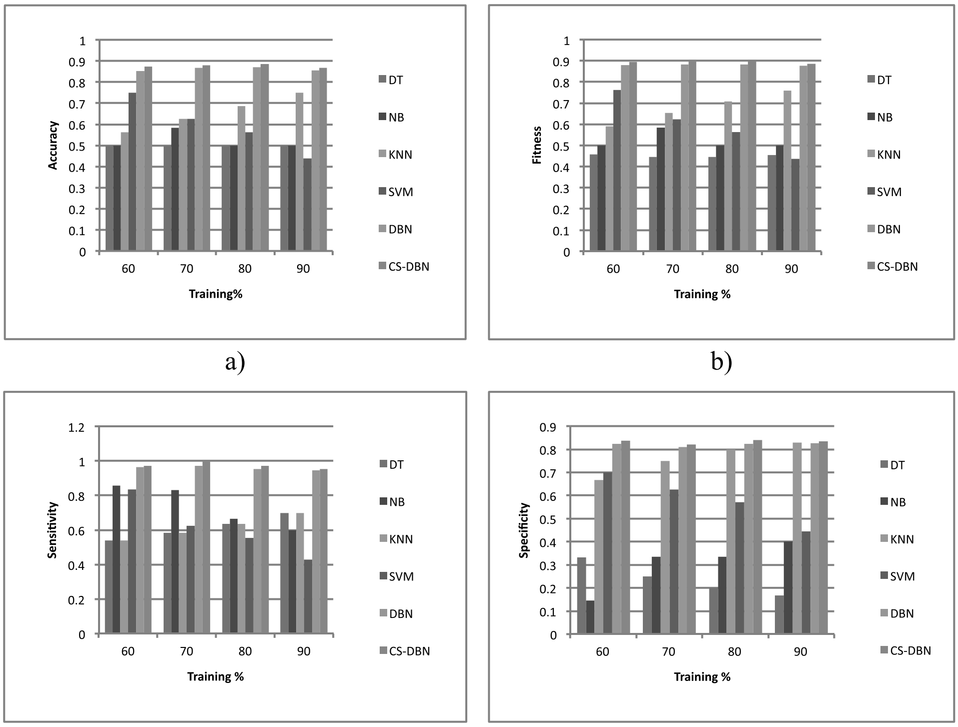
Comparative analysis of the methods of privacy protection in medical data

This section details the comparative study of CS-DBN technique based on fitness, sensitivity, accuracy, and specificity in terms of the datasets, namely, Cleveland, Hungary, and Switzerland.

Comparative analysis using Cleveland dataset:

The comparative analysis of the medical data classification methods using the Cleveland dataset is depicted in figure 4. Figure 4.a shows the accuracy of the methods for various training percentages based on the Cleveland dataset. With 70% training, the accuracy of DT, NB, K-NN, SVM, DBN, and the proposed CS-DBN is 0.5, 0.583, 0.625, 0.625, 0.868, and 0.8783, respectively. With 80% training, the accuracy of DT, NB, K-NN, SVM, DBN, and the proposed CS-DBN is 0.5, 0.5, 0.6875, 0.5625, 0.8708, and 0.8842, respectively. Figure 4.b depicts the fitness of the methods for various training percentages based on Cleveland dataset. When the training percentage is 70, the fitness of the methods, such as DT, NB, K-NN, SVM, DBN, and the proposed CS-DBN, is 0.4444, 0.583, 0.6528, 0.625, 0.8839, and 0.8995, respectively. When the training percentage is 80, the fitness of the methods, such as DT, NB, K-NN, SVM, DBN, and the proposed CS-DBN, is 0.4455, 0.5, 0.708, 0.5632, 0.8826, and 0.8984, respectively.

Figure 4.c depicts the sensitivity of the methods for various training percentages based on Cleveland dataset. When the training percentage is 70, the sensitivity of DT, NB, K-NN, SVM, DBN, and the proposed CS-DBN is 0.5833, 0.8322, 0.5833, 0.625,0.9728, and 1, respectively. When the training percentage is 80, the sensitivity of DT, NB, K-NN, SVM, DBN, and the proposed CS-DBN is 0.6364, 0.6662, 0.6364, 0.5556, 0.9529,and 0.9702, respectively. Figure 4.d depicts the specificity of the methods for various training percentages based on Cleveland dataset. When the training percentage is 70, the specificity of the methods, such as DT, NB, K-NN, SVM, DBN, and the proposed CS-DBN, is 0.25, 0.3339, 0.75, 0.625, 0.81, and 0.8203, respectively. When the training percentage is 80, the specificity of the methods, such as DT, NB, K-NN, SVM, DBN, and the proposed CS-DBN, is 0.2, 0.3338, 0.8, 0.5714,0.8243, and 0.8408, respectively.



a) b)

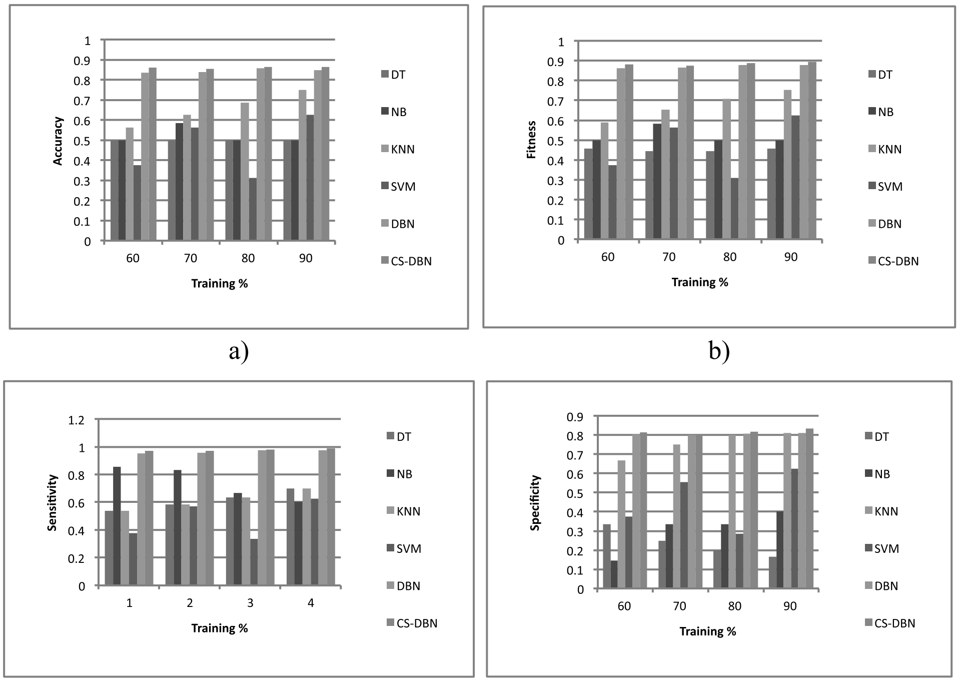
c) d)

Figure 3. Analysis using Cleveland dataset based on a) accuracy, b) fitness, c) sensitivity, d) specificity.

Comparative analysis using Hungarian dataset

The comparative analysis of the medical data classification methods using the Hungarian dataset is depicted in figure 5. Figure 5.a shows the accuracy of methods for various training percentages based on the Hungarian dataset. With 70% training, the accuracy of DT, NB, K-NN, SVM, DBN, and the proposed CS-DBN is 0.5, 0.583, 0.625, 0.5625,0.8383, and 0.8543, respectively. With 80% training, the accuracy of DT, NB, K-NN, SVM, DBN, and the proposed CS-DBN is 0.5, 0.5, 0.6875, 0.3125,0.8571, and 0.8638, respectively. Figure 5.b depicts the fitness of the methods for various training percentages based on the Hungarian dataset. When the training percentage is 70, the fitness of the methods, such as DT, NB, K-NN, SVM, DBN, and the proposed CS-DBN, is 0.4444, 0.583, 0.6528, 0.5632, 0.8656, and 0.8762, respectively. When the training percentage is 80, the fitness of the methods, such as DT, NB, K-NN, SVM, DBN, and the proposed CS-DBN, is 0.4455, 0.5, 0.708, 0.3105, 0.8799, and 0.8881, respectively.

Figure 5.c depicts the sensitivity of the methods for various training percentages based on the Hungarian dataset. When the training percentage is 70, the sensitivity of DT, NB, K-NN, SVM, DBN, and the proposed CS-DBN is 0.5833, 0.8322, 0.5833, 0.5714, 0.9583, and 0.9696, respectively. When the training percentage is 80, the sensitivity of DT, NB, K-NN, SVM, DBN, and the proposed CS-DBN is 0.6364, 0.6662, 0.6364, 0.3333, 0.9751, and 0.9819, respectively. Figure 5.d depicts the specificity of the methods for various training percentages based on the Hungarian dataset. When the training percentage is 70, the specificity of the methods, such as DT, NB, K-NN, SVM, DBN, and the proposed CS-DBN, is 0.25, 0.3339, 0.75, 0.5556,0.80, and 0.8046, respectively. When the training percentage is 80, the specificity of the methods, such as DT, NB, K-NN, SVM, DBN, and the proposed CS-DBN, is 0.2, 0.3338, 0.8, 0.2857,0.8074, and 0.8186, respectively.



a)

b)

c)

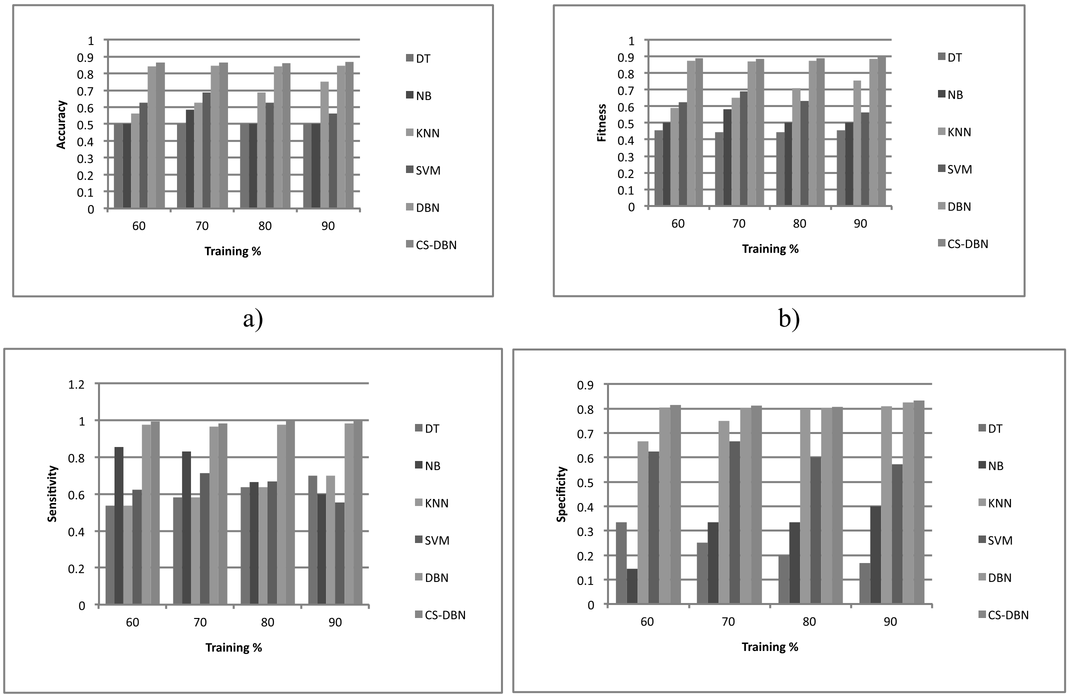
d)

Figure 4. Analysis using Hungarian dataset based on a) accuracy, b) fitness, c) sensitivity, d) specificity.

Comparative analysis using Switzerland dataset:

The comparative analysis of the medical data classification methods using the Switzerland dataset is depicted in figure 6. Figure 6.a shows the accuracy of methods for various training percentages based on the Switzerland dataset. With 70% training, the accuracy of DT, NB, K-NN, SVM, DBN, and the proposed CS-DBN is 0.5, 0.583, 0.625, 0.6875, 0.8441,and 0.863, respectively. With 80% training, the accuracy of DT, NB, K-NN, SVM, DBN, and the proposed CS-DBN is 0.5, 0.5, 0.6875, 0.625, 0.8412, and 0.8623, respectively. Figure 6.b depicts the fitness of the methods for various training percentages based on Switzerland dataset. When the training percentage is 70, the fitness of the methods, such as DT, NB, K-NN, SVM, DBN, and the proposed CS-DBN, is 0.4444, 0.583, 0.6528, 0.6895,0.8725, and 0.8858, respectively. When the training percentage is 80, the fitness of the methods, such as DT, NB, K-NN, SVM, DBN, and the proposed CS-DBN, is 0.4455, 0.5, 0.708, 0.6306, 0.8726, and 0.8882, respectively.

Figure 6.c depicts the sensitivity of the methods for various training percentages based on Switzerland dataset. When the training percentage is 70, the sensitivity of DT, NB, K-NN, SVM, DBN, and the proposed CS-DBN is 0.5833, 0.8322, 0.5833, 0.7143, 0.9653,and 0.9829, respectively. When the training percentage is 80, the sensitivity of DT, NB, K-NN, SVM, DBN, and the proposed CS-DBN is 0.6364, 0.6662, 0.6364, 0.6667, 0.976,and 0.9954, respectively. Figure 6.d depicts the specificity of the methods for various training percentages based on Switzerland dataset. When the training percentage is 70, the specificity of the methods, such as DT, NB, K-NN, SVM, DBN, and the proposed CS-DBN, is 0.25, 0.3339, 0.75, 0.6667, 0.8023, and 0.8115, respectively. When the training percentage is 80, the specificity of the methods, such as DT, NB, K-NN, SVM, DBN, and the proposed CS-DBN, is 0.2, 0.3338, 0.8, 0.6, 0.8005, and 0.8069, respectively.



a) b)

c) d)

Figure 5. Analysis using Switzerland dataset based on a) accuracy, b) fitness, c) sensitivity, d) specificity.

Analysis based on the computational time

The analysis of the proposed method with the existing methods in terms of computational time is provided in this section. Table 2 depicts the computational time of the proposed method and the existing methods, such as DT, NB, K-NN, SVM, and DBN, in which the proposed system has less computation time of 6 sec.

Table 2. Computational time.

Methods	DT	NB	K-NN	SVM	DBN	Proposed CS-DBN
Time (Sec)	13	11.5	10.4	8	7.5	6

CONCLUSION

Preserving the privacy of medical data in the ontology-based systems is a critical need, especially in the case when the system is used by more numbers of users with various privileges and is distributed over applications. Thus, it is necessary to take steps for the preservation of the medical data of the patients. This paper aims to preserve confidential medical data with the introduction of a medical data classification method. The proposed CS-DBN method works based on three main steps, namely, generation of privacy preserved data, construction of ontology, and classification. The deep convolutional kernel approach is utilized for the provision of data confidentiality with the generation of optimal coefficients. The ontology is developed with the terms related to cardiac heart disease for classification. The classification is carried out using deep belief network (DBN) that is trained by crow search algorithm (CSA). The analysis of CS-DBN is performed in terms of the metrics, namely, fitness, accuracy, sensitivity, and specificity, and it produces the higher fitness, accuracy, sensitivity, and specificity of 0.9007, 0.8842, 1, and 0.8408, respectively. In future, the data classification will be based on any hybrid optimizations, and the analysis will be done using more medical databases.

REFERENCES

Alabdulkarim, A., Al-Rodhaan, M., Ma, T & Tian, Y. (2019). PPSDT: A Novel Privacy-Preserving Single Decision Tree Algorithm for Clinical Decision-Support Systems Using IoT Devices, **3**(19): 1.

Al-Aidaroo, K.M., Bakar, A.A & Othman, Z. (2012). Medical Data Classification with Naive Bayes Approach, Information Technology Journal, **11**(9): 1166-1174.

Arul. V.H, V.G. Sivakumar, Ramalatha Marimuthu & Basabi Chakraborty. (2019). An Approach for Speech Enhancement Using Deep Convolutional Neural Network, Multimedia Research (MR), **2**(1): 37-44.

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

Bernabe, J.B., Perez, G.M., Antonio, F & Gomez, S. (2015). Intercloud Trust and Security Decision Support System: an Ontology-based Approach, Journals on Grid Computing.

Biskup, J & Bonatti, P.A. (2004). Controlled query evaluation for enforcing confidentiality in complete information systems, International Journal of Information Security, **3**(1)14-27.

Biskup, J & Weibert, T. (2008). Keeping secrets in incomplete databases, International Journal of Information Security, **7**(3): 199-217.

Cachin, C & Haas, R. (2010). Dependable Storage in the Intercloud, IBM Research Report RZ 3783.

Dasarathy, B.V. (1980). Nosing aroung the neighbourhood: A new system structure and classification rule for recognition in partially exposed environments, IEEE Transactions on Pattern Analysis and Machine Intelligence, **2**: 67-71.

Duba, R.O & Hart, P.E. (1973). Pattern Classification and Scene Analysis, New York: Wily.

Fan, W., Chan, C.Y. & Garofalakis, M.N. (2004). Secure xml querying with security views, In Proceedings of SIGMOD, 587-598.

Farooq, K & Hussain, A. (2016). A novel ontology and machine learning driven hybrid cardiovascular clinical prognosis as a complex adaptive clinical system,Complex adaptive syatems modelling, **4**(12).

Geibel, P., Trautwein, M., Erdur, H., Zimmermann, L., Jegzentis, K., Bengner, M., Nolte, C.H & Tolxdorff, T. (2015). Ontology-Based Information Extraction: Identifying Eligible Patients for Clinical Trials in Neurology, Journal on Data Semantics, **4**(2): 133-147.

Grau, B.C. (2010). Privacy in Ontology-based Information Systems: A Pending Matter, Semantic Web, **1**(1,2): 137-141,.

Heart disease dataset, "https://archive.ics.uci.edu/ml/datasets/Heart+Disease" accessed on January 2019.

Karlekar, N.P & Gomathi, N. (2017). Kronecker product and bat algorithm-based coefficient generation for privacy protection on cloud, International Journal of Modeling, Simulation and Scientific Computing, **8**(3).

Karlekar, N.P & Gomathi, N. (2018). OW-SVM: Ontology and whale optimization-based support vector machine for privacy-preserved medical data classification in cloud, **31**(12).

Keller, J.M., Gray, M.R & Givens, J.A. (1985). A Fussy-K-Nearest Neighbor Algorithm, IEEE Transactions on System, Man, and Cybernetics, **15**(4): 580-585.

Levy, A.Y. (1996). Obtaining complete answers from incomplete databases, In Proceedings of Very Large Data Bases, 402-412.

Multi-Keyword Ranked Search Scheme over Encrypted Cloud Data, IEEE Transactions on Parallel Distribution Systems, **27**(2): 340-352.

Pramod P Jadhav & SD Joshi, ACADF(2019): Ant Colony Unified with Adaptive Dragonfly Algorithm Enabled with Fitness Function for Model Transformation, Springer, Singapore:101-109.

Ram, C.P & Sreenivaasan, G. (2010). Security as a Service (SasS): Securing user data by coprocessor and distributing the data, Trendz in Information Sciences & Computing (TISC2010), 152-155.

Stouppa, P & Studer, T. (2007). A formal model of data privacy, In Proceedings of Perspectives of System Informatics 06, 4378.

Tao, M., Zuo, J., Liu, Z., Castiglione, A & Palmieri, F. (2018). Multi-layer cloud architectural model and ontology-based security service framework for IoT-based smart homes, Future Generation Computer Systems, **78**(3): 1040-1051.

Vito Racanelli, Claudia Brunetti, Valli De Re, Laura Caggiari, Mariangela De Zorzi, Patrizia Leone, Federico Perosa, Angelo Vacca & Franco Dammacco, (2011). Antibody Vh repertoire differences between resolving and chronically evolving hepatitis C virus infections, Public Library of Science, **6**(9).

Vojt, B.J. (2016). Deep neural networks and their implementation, Department of Theoretical Computer Science and Mathematical Logic, Prague.

Whitney, A & Dwyer, S.J. (1966). Performance and implementation of K-nearest neighbour decision rule with incorrectly identified training samples, In Proceedings of 4th Annual Allerton of Conference On Circuits Band System Theory.

Zardari, M.A., Jung, L.T. & Zakaria, N. (2014). K-NN Classifier for Data Confidentiality in Cloud Computing, International Conference on Computer and Information Sciences (ICCOINS).

Zhou, L., Zhang, C., Karimi, I.A & Kraft, M. (2018). An ontology framework towards decentralized information management for eco-industrial parks, Computers & Chemical Engineering, **118**: 49-63.