Biomedical Research 2016; 27 (3): 718-724

ISSN 0970-938X
www.biomedres.info

Software aided diagnosing of diseases using RBF based neural networks [RADD].

V Rajalakshmi[1*], GS Anandha Mala[2]

[1]Research Scholar, Faculty of Computing, Sathyabama University, Chennai, India

[2]Professor, Department of CSE, Easwari Engineering College, Chennai, India

Abstract:

With the advancement in technology and change in lifestyle, Indians are more prone to various disorders like diabetes, stroke, hypertension, etc. Presence of these diseases is identified only by regular monitoring and invasive blood tests. Though there is much successful management of communicable diseases and genetic disorders after they occur, there are only limited procedures to identify before they occur. A procedure, which detects the occurrence of such lifetime disorders would save the livelihood, before reaching the critical stage, is essential. The occurrence of such diseases can be detected by inherent watch of symptoms caused in existing patients. The similarity between the existing patients is monitored by using a FUNAP [Function Approximation] System. Existing patients' data are collected and the system is trained to detect the diseases of monitoring new patients. The method is efficient as it uses artificial neural network for matching and it can be tuned according to the requirements. The system is explained with its architecture and its performance is compared with existing methods.

Keywords: Radial basis function, Function approximation, Data anonymization, Neural Network.

Accepted on February 25, 2016

Introduction

The usage of internet and social networking web sites has increased the amount of globally available data. Because of this, we are drowning in data but starving for knowledge and privacy. These data provide us the information by data mining to retrieve non-trivial knowledge for future decision making. As techniques for revealing non-trivial patterns using various data mining algorithms are explored, the amounts of such data also increase in an uncontrolled manner. When such data are provided for mining in their original form, it forms a threat for the privacy of an individual. Typical example includes disease of a patient, credit card balance of a customer, purchase details from a departmental store, government weapon details in military, etc., Anonymization issues also occur in surveying, statistical databases, cryptographic computing, access control, social networking and so on. Hence data need to be modified before they are provided for mining or to any third party for processing. Hence the anonymized data should be used to identify the characteristics of a record similar to an original one. When a normal back propagation network or Hopfield network used for disease recognition, they cannot handle anonymized data and hence they are inefficient. The system built for identifying the diseases should be more sensitive to the input data and map with the correct disease even if the input is anonymized. There is an inverse relationship between privacy and utility of the data.

For mining, as the exact data is not required, a perfect approximation is sufficient, the modification is accepted. A neural network is capable of recognizing the exact pattern for an input. This characteristic is utilized to build a system and the system performance can be tuned with its parameters. Thus the efficiency of the output is tunable and hence the system can be utilized to identify the disease with the symptoms of tested patients. Attributes in a database are divided into three types – unique identifying attributes, sensitive attributes, quasi identifying attributes. When data are given for mining unique identifying attributes like patient ID, credit card number, Employee ID, etc., are removed completely from the database. Sensitive attributes like disease, credit card balance, salary, etc., are the primary concerns for mining and they are not altered. Quasi identifying attributes like age, zipcode, height, married, gender, symptoms, etc., are also available in a public database like voter's list. These are the values which are altered so that the exact individual of the record is not identified. Hence the symptoms used are modified slightly before they are given for mining.

The information disclosure is categorized into two types, Identity disclosure that specifies which record is associated with which individual in a released table and in Attribute disclosure, new information about some individuals is revealed by the released table. In this work, Identity disclosure is handled i.e., the association between an individual and a particular record is tried to be hidden. The features of the record with its sensitive attribute is associated using Radial Basis Function [RBF] based neural network.

Presently, diseases are detected by trial and error method, or by various tests for every disease. When existing methods are used, they are invasive and can be detected only after they occur. There is a need for a method that detects diseases before they occur. Symptoms of various patients are identified and a training data is formed. An RBF network is designed to identify the related disease. There may be more than one symptom for a disease and more than one disease with similar symptoms. Hence neural network are the best method to implement this procedure.

Related works

In [1], diseases are identified using genomics, ie., DNA sequence of the patient and their relatives are analyzed to predict the diseases. Since the DNA sequences of relatives are readily available, this method is not effective [2]. Is a goggle tool which identifies the outbreak of epidemics by number of searches [WBT]. The tool provides only the recently occurring diseases and do not consider a general disorders. In [3], an effective method to handle the symptoms of diabetes is considered. The method only discusses for existing patients and not for new ones [4]. Proved that Coronary artery disease can develop prematurely and is the leading cause of death among diabetics, making noninvasive risk stratification desirable. This work gave way for relating one disease as a serious symptom for the other to occur [5] explains the prediction of mortality and subsequent myocardial infarction in patients with unstable coronary artery disease [6]. Supplies the prediction of 2-day stroke risk to inform emergency management. It specifies various symptoms of patients to detect stroke just before two days. It also can detect stroke only after the occurrence of transient ischemic attack (TIA). In [7], a detailed examination of factors associated with variation in the risk for type 2 diabetes in women with prior gestational diabetes mellitus is given [8]. Provides a mathematical model to estimate the risk of a first stroke using data from 4549 newly diagnosed type 2 diabetic patients enrolled in UK. It relates diabetes as a symptom for 60 other diseases also.

In [9], a non-invasive procedure, the electron beam CT calcium score appeared to be an effective predictor of coronary artery disease [10] examines various measures that may be associated with diseases. This is utilized to specify the list of diseases and their symptoms [11]. Explains the recent advances in TB diagnostic technologies, their symptoms and precautions [12]. describes that depression and its medications causes various other disorders including myocardial infarction.

Neural networks are efficient systems which can identify similar items naturally and then they can be anonymized easily. In [13], back propagation network is used to implement classification of data and in [14], neural network is used to implement clustering. In [15] radial basis function [RBF] network is used for privacy preservation. The effectiveness of radial basis function is retrieved from this literature [16]. Specifies the procedure for using health data in the procedure of detecting diseases and various methods used in medical environments. In [17], a diagnosis of Alzheimer's type dementia is proposed using support vector machines [SVM], which specifies mathematical equations for specifying the limits of every category. In [18] a health service system specifying the probability of disease diagnosed.

Problem definition

In order to reduce the execution time of disease detection system, to enhance the efficiency in terms of both sensitivity and accuracy of the results and also to set a tunable parameter, RBF network is used as a function approximation network which can group similar data along with approximating them. The main objectives of this work are,

- To efficiently recognize the anonymized data and classify into one of maximum number of individual diseases.
- To provide proper tuning parameter between accuracy and sensitivity.
- The system should provide more accuracy with less time.
- The system should train itself when error increases.

Description of RADD

Radial basis network is used for grouping the symptoms efficiently to detect the right set of possible diseases. Radial basis network are very sensitive and clearly classifies the output according to input. When the network is trained the number of neurons, spread, gain and error value of the network are fixed a given set of inputs. After the network is trained it can be used to detect the right disease when the symptoms are given. Sensitivity is a value given by the user to choose possible number of outcomes. When required the network can be retrained with new set of inputs and the network will have a new set of parameters.

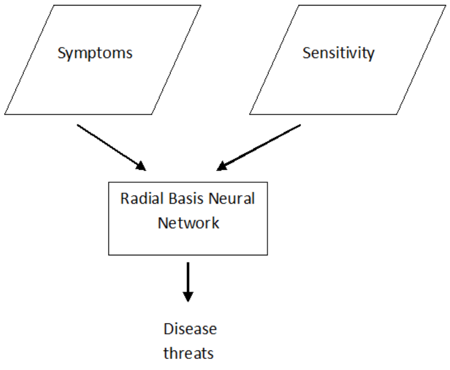

Figure 1. Architecture of the system.

The value can range between 0 and 100. 0% sensitivity provides all possible diseases for the specified symptoms and 100% sensitivity provides only very particular diseases matching the symptoms. Since the system is a non-invasive procedure for acquiring knowledge 100% sensitivity may not be required always.

Radial Basis Function Network

Radial basis function networks (RBFNs), are special type of neural networks which are being applied for problems such as function approximation, pattern recognition and time series prediction, etc., RBF networks when used directly as a universal approximators of desired accuracy, provide a solution for PPDM but with a less efficiency . The standard RBF network consists of three layers, i.e., the input, hidden and output layers.

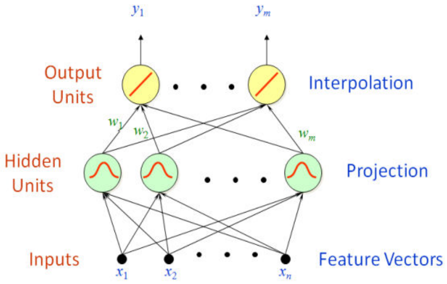

Figure 2. *Radial basis Network.*

The hidden layer of an RBF network can be viewed as a function that maps the input patterns from a nonlinear separable space to a linear separable space. In the new space, the responses of the hidden-layer neurons then form a new feature vectors for pattern representation. Each output vector can be assumed as a representation of a group of input patterns. The architecture of a radial basis network is specified in figure1 and 2. The figure has n inputs and m outputs.

Given a data set X with size N x m and the output vector Y of same size is shown below:

$$X = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,m} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,m} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ X_{N,1} & X_{N,2} & \cdots & X_{N,m} \end{pmatrix} \rightarrow (1)$$

$$Y = (y1\ \ y2\ \ \dots\ \ ym) \rightarrow (2)$$

Now each row of

$$x_i = x_{i,1}\ x_{i,2}\ \cdot\ x_{i,m}$$

targets a row of

$$y_i = y_{i,1}\ y_{i,2}\ \cdots\ y_{i,m}$$

. We want to find a target function f(x_i), that produces the lowest error when predicting the unknown related values Yj. This is equivalent to determining the weight vector W for finding Y with minimum error.$w = w_1\ \ w_2\ \ \dots\ \ w_p \rightarrow (3)$

$$Y\ =\ f\ (x) \rightarrow (4)$$

Using Radial Basis Network the function is chosen as a radial basis function as follows

$$f(x) = \sum_{k\,=\,1}^{p} w_k \varphi_k(x) \rightarrow (5)$$

The radial function can be specified as

$$\varphi_k(x) = \quad \varphi(||x - x_k||) \rightarrow (6)$$

where x_k is the center of the activated neuron.

The three main parameters of a radial basis function are

- Centre Xk

Distance Measure

$$r\ =\ ||x\ -\ x_k||$$

- Shape of the radial basis function

Figure 2 specifies the how an RBF network is given input for anonymization and gives a sample output of function approximators. It shows that the output of the network follows the input and they are not similar values as input.

For training the network, a training set of data is chosen and the network is built.

$$T = \quad \{(x^k, y^k)\}_{k\,-\,1}^{p} \rightarrow (7)$$

The goal is set as

$$y^{(k)} \approx f(x^{(k)}) \rightarrow (8)$$

Experimental Setup

Adult data set from UCI repository database is designed, for implementing the disease detection system. The data set contains 48842 records, from which NULL, outliers and obsolete valued records are removed to make 30,162 records after preprocessing. Attributes which are selected for the processing and their properties are shown in table 2.

Matlab is used to implement radial basis network and verify the results. Design of the network is done for different spread, goal and number of neurons and the best output is selected based on the performance metrics. Matlab is chosen because of its flexibility in altering the RBF network parameters for tuning and also provides graphical representation of outputs.

A set of 500 records are selected by stratified sampling and the network is trained. Stratified sampling is a probability sampling technique wherein the researcher divides the entire population into different subgroups or strata, then randomly selects the final subjects proportionally from the different strata.

A set of 2500 records are chosen to test the performance of the system with respect to the actual results. The remaining records are given directly and the results are compared with the standard web based technology [WBT] method. Creation of the content forms the core capability of this paper and is created by consulting physicians and various other sources like articles on similar subjects, web sites, etc (Table 1).

Table 1. *Sample of training data.*

Symptoms	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7	Case 8	Case 9	Case 10	Case 11	Case 12	Case 13	Case 14	Case 15	Case 16
Abdominal Pain	0	0	1	0	1	1	1	0	0	0	0	1	0	0	0	0
Blood In Stool	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0
Chest Pain	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
Cough	0	0	1	0	0	1	1	0	0	1	0	0	1	0	1	1
Dark Urine	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
Frequent watery motion	1	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0
Vomiting	1	0	1	1	0	0	1	1	0	1	0	1	0	0	0	0
Dizziness	0	1	0	0	0	1	0	0	0	0	0	1	1	0	0	0
Fatigue	0	0	1	0	1	1	0	1	1	0	1	1	1	1	0	1
Fever	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1
Frequent Urination	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Gas	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Headache	0	1	0	0	1	0	1	1	0	0	0	1	1	1	1	0
Joint Pain	0	0	1	0	1	1	1	1	0	0	0	1	1	1	1	0
Loss of Appetite	0	1	0	0	0	0	0	1	1	0	0	1	0	1	1	1
Nausea	1	1	0	1	1	0	0	1	0	1	1	1	1	0	0	0
Rash	0	0	0	0	1	1	0	0	0	0	0	1	0	0	1	0
Dehydration	1	1	1	1	0	1	1	0	1	0	0	1	1	0	1	1
Swollen Feet	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Sneezing	1	0	1	1	0	1	0	0	0	0	0	0	1	0	0	0
Stuffy nose	0	1	1	0	0	1	1	0	0	1	0	0	1	0	0	0
Cold sores	1	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0
Weight gain	0	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0
Stomach pain	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1
Back pain	1	0	1	0	1	1	0	0	0	0	0	1	0	1	0	1
Diseases	Diarrhea	Cholera	Malaria	Typhoid	Dengue	Swine Flu	Ebola	Liver Cancer	Diabetes	HIV	Hepatitis	Jaundice	Influenza	Leukaemia	Plague	Tuberculosis

Experimental outputs

A Radial basis network is set using training data for a spread of .85 and error =10 %. Network parameters control the efficiency of the system. Anyone parameter can be used to control sensitivity of the system.

As sensitivity is a run time input and it can be varied for every case, error is chosen to control the sensitivity of the system. When error value of the network is increased its sensitivity decreases and a broad classification is done by the system.

Hence error and sensitivity are inversely proportional parameters. Spread and number of neurons is static parameters and hence they cannot be used. Tables 3,4 and 5 show the outputs for the same sample of 5 cases with sensitivity values 30%, 50% and 80% respectively. The tables show that as the sensitivity increases the number of diseases predicted reduces.

Table 2. Adult data set Attributes

Attribute name	Attribute data type	Attribute relevance
Age	Continuous	Quasi Identifier
Gender	Categorical	Quasi Identifier
Hours per week	Continuous	Quasi Identifier
Occupation	Categorical	Quasi Identifier
Height	Continuous	Quasi Identifier
Weight	Continuous	Quasi Identifier
BMI	Continuous	Quasi Identifier
Symptom1	Boolean	Quasi Identifier
Symptom2	Boolean	Quasi Identifier
…		
Symptom n	Boolean	Quasi Identifier
Sensitivity	Continuous	-
Risk for Disease	Categorical	Sensitive

Table 3. Sensitivity = 30%.

Case	Symptoms	Disease detected
1	Frequent watery motion, Vomiting, Fatigue, Frequent Urination, Joint Pain	Diarrhea, Cholera, Malaria, Dengue, Diabetes
2	Loss of Appetite, Nausea, Rash, Dehydration, Swollen feet	Liver Cancer, Swine Flu, Typhoid, Ebola, Dengue
3	Blood in stool, Chest pain, Abdominal pain, Dark urine, Dizziness	Liver Cancer, Plague, Malaria, Jaundice, HIV, Tuberculosis
4	Cough, Fever, Gas, Headache, Sneezing	Malaria, Swine Flu, Influenza, Plague, Tuberculosis
5	Stuffy nose, Cold sores, Weight gain, Stomach pain, Back pain	Influenza, Dengue, Diabetes, Diarrhea, Leukemia

Figure 3 shows the variation of sensitivity and their corresponding number of diseases as output. As the sensitivity is increased, which is implemented by varying the value of error parameter is reduced in the radial basis network, thus classification becomes more specific giving less number of exactly matching diseases.

Performance evaluation

The validity of the Performance based value was thoroughly studied and due to the content sharing it is mentioned as follows. Performance of RADD is measured against the standard WBT based on two parameters.

- Disease detection time
- Accuracy

Table 4. Sensitivity = 50%.

Case	Symptoms	Disease detected
1	Abdominal pain, Fatigue, Fever, Joint pain, Stomach pain	Malaria, Dengue, Swine flu, Leukaemia
2	Frequent watery motion, Vomiting, Dehydration, Swollen feet	Diarrhea, Typhoid, Ebola, Jaundice
3	Blood in stool, Chest pain, Dark urine, Dizziness	Liver cancer, Tuberculosis, Jaundice, Swine Flu
4	Cough, Fever, Gas, Headache	HIV, Influenza, Plague, Tuberculosis
5	Cold sores, Weight gain, Stomach pain, Back pain	Dengue, Diabetes, Cholera, Leukemia

Figure 6 shows the capability of the method in terms of failure detection and correction done automatically. The network adapts itself when a failure occurs. The number of failure on voluminous data is also less when compared to existing methods.

Table 5. Sensitivity = 80%

Case	Symptoms	Disease detected
1	Fatigue, Fever, Joint pain	Leukemia, Malaria
2	Frequent watery motion, Vomiting, Nausea	Typhoid, Cholera
3	Stomach pain, Back pain, Fever	Diarrhea, Dengue
4	Fever, Joint pain, Blood in stool	Swine flu, Liver cancer
5	Cough, Stuffy nose, Nausea	HIV, Hepatitis

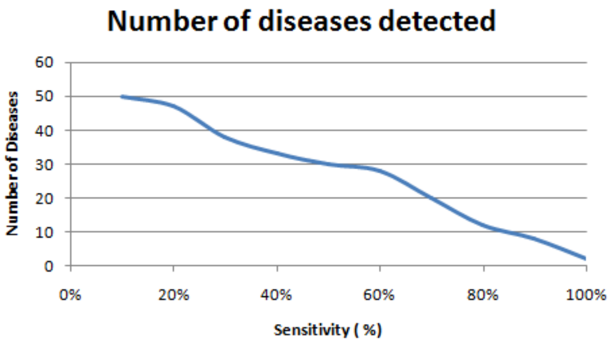

Figure 3. Sensitivity control.

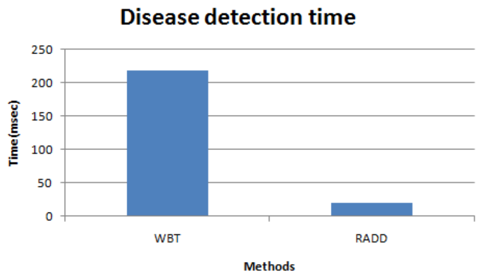

Figure 4. *Performance based on Disease detection time.*

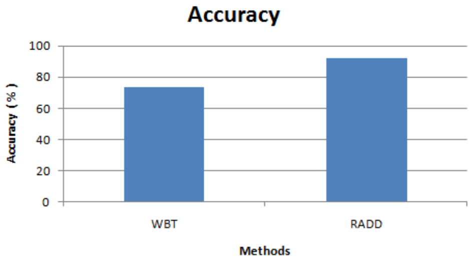

Figure 5. *Performance based on accuracy.*

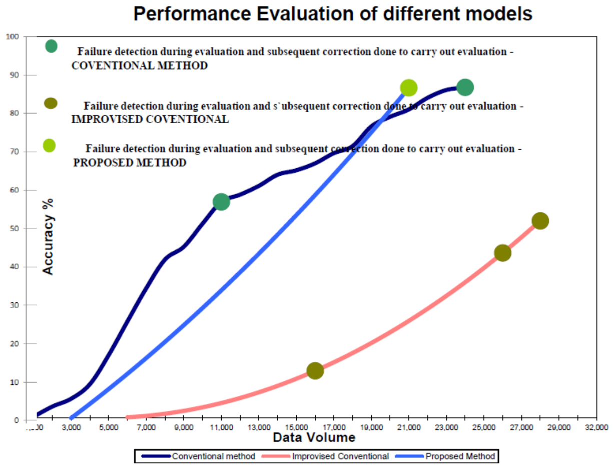

Figure 6. *Failure Detection and Subsequent Correction Analysis.*

Figure 4 shows the comparison based on disease detection time. WBT uses the existing symptoms, compares them based on the query. As there is no network built and for every query comparison with every symptom in the database is checked it takes more amount of time. RADD has an efficient radial basis neural network which is already trained to handle similar data inputs. Hence RADD produces the result in less time compared to WBT.

Figure 5 shows the performance comparison based on accuracy of the output. In WBT, the queries are evaluated based on the patient's experiences. The real symptoms may be hidden or not have occurred directly to the patients. The relationship between various symptoms is not handled and hence accuracy of WBT is less than RADD. RADD uses the actual patient's database that are anonymized and supplied by medical institutions. Hence the network which is trained to handle such data provides more accurate results than a web based searching system.

Conclusion and future work

Thus a system used for detecting diseases using radial basis neural network is built and its performance is hence compared with existing system. The reduction in the execution time of disease detection system, the efficiency in terms of both sensitivity and accuracy of the results were also compared along with usage of a tunable parameter, RBF network as a function approximation network which was used to group similar data. The proposed method has the following advantages - less time, reduction in number of tests and early detection.

The system recognized the anonymized data efficiently and classified maximum number of individual diseases. Excellent tuning parameter between accuracy and sensitivity was established and with more accuracy associated with less time. Also, the system self- assisted when error increases.

As continuation of this work, RADD will be put in use to effectively analyze the failures occurring in automotive vehicles. In present scenario, On-Board Diagnostics (OBD) is in place to assist in identifying the type of failure with certain limitations on the fault error codes. This proposed user-friendly system will surpass the existing methods and reduce lot of time and efforts invested into such analysis.

References

1. John B. Predicting disease using genomics. Nature 2004; 429: 453-456.
2. Carneiro HA, Mylonakis E. Google trends: a web-based tool for real-time surveillance of disease outbreaks. Clinical infectious diseases 2009; 49: 1557-1564.
3. Ciechanowski PS, Katon WJ, Russo JE, Hirsch IB. The relationship of depressive symptoms to symptom reporting, self-care and glucose control in diabetes. General hospital psychiatry 2003; 25: 246-252.
4. Giri S, Shaw LJ, Murthy DR, Travin MI, Miller DD, Hachamovitch R, Borges-Neto S, Berman DS, Waters DD, Heller GV. Impact of diabetes on the risk stratification using stress single-photon emission computed tomography myocardial perfusion imaging in patients with symptoms suggestive of coronary artery disease. Circulation 2002; 105: 32-40.
5. James SK, Lindahl B, Siegbahn A, Stridsberg M, Venge P, Armstrong P, Barnathan ES, Califf R, Topol EJ, Simoons ML, Wallentin L. N-Terminal Pro–Brain Natriuretic Peptide and Other Risk Markers for the Separate Prediction of Mortality and Subsequent Myocardial Infarction in Patients With Unstable Coronary Artery Disease. Circulation 2003; 108: 275-281.
6. Johnston SC, Rothwell PM, Nguyen-Huynh MN, Giles MF, Elkins JS, Bernstein AL, Sidney S. Validation and

refinement of scores to predict very early stroke risk after transient ischaemic attack. The Lancet 2007; 369: 283-292.

7. Kim C, Newton KM, Knopp RH. Gestational Diabetes and the Incidence of Type 2 Diabetes A systematic review. Diabetes care 2002; 25: 1862-1868.
8. Kothari V, Stevens RJ, Adler AI, Stratton IM, Manley SE, Neil HA, Holman R. UKPDS 60 risk of stroke in type 2 diabetes estimated by the UK Prospective Diabetes Study risk engine. Stroke 2002; 33: 1776-1781.
9. Mautner SL, Mautner GC, Froehlich J, Feuerstein IM, Proschan MA, Roberts WC, Doppman JL. Coronary artery disease: prediction with in vitro electron beam CT. Radiology 1994; 192: 625-630.
10. Paulsen JS, Hayden M, Stout JC, Langbehn DR, Aylward E, Ross CA, Guttman M, Nance M, Kieburtz K, Oakes D, Shoulson I, Kayson E, Johnson S, Penziner E. Preparing for preventive clinical trials: the Predict-HD study. Archives of neurology 2006; 63: 883-890.
11. Perkins MD. New diagnostic tools for tuberculosis [The Eddie O'Brien Lecture]. The International Journal of Tuberculosis and Lung Disease 2000; 4: S182-S188.
12. Pratt LA, Ford DE, Crum RM, Armenian HK, Gallo JJ, Eaton WW. Depression, psychotropic medication, and risk of myocardial infarction prospective data from the Baltimore ECA follow-up. Circulation 1996; 94: 3123-3129.
13. Samet S, Miri A. Privacy-preserving back-propagation and extreme learning machine algorithms. Data & Knowledge Engineering 2012; 79: 40-61.
14. Tsiafoulis S, Zorkadis VC, Karras DA. A Neural Network Clustering Based Algorithm for Privacy Preserving Data Mining. InComputational Intelligence and Security (CIS), International Conference 2010; pp. 401-405.
15. Lin CJ, Chen CH, Lee CY. A self-adaptive quantum radial basis function network for classification applications. In Neural Networks, Proceedings. IEEE International Joint Conference2004; 4: pp. 3263-3268.
16. Lambeau B, Damas C, van Lamsweerde A. Process Execution and Enactment in Medical Environments. Springer-Verlag Berlin Heidelberg 2011.
17. Ramíreza J, Górriza JM, Salas-Gonzaleza D, Romeroa A , Lópeza M, Álvareza I, Gómez-Ríob M. Computer-aided diagnosis of Alzheimer's type dementia combining support vector machines and discriminant set of features." Information Sciences 2013; 237: 59-72.
18. Byung WM, Cho HJ, Jeong HY. Implementation of mobile U-health service system based on personalized computer aided diagnosis probability. Multimedia Tools and Applications 2015: 1-24.

*Correspondence to:

Rajalakshmi V

Department of Computing

Sathyabama University

India