Cuckoo Search Optimized Reduction and Fuzzy Logic Classifier for Heart Disease and Diabetes Prediction

Thippa Reddy Gadekallu, VIT University, Vellore, India

Neelu Khare, VIT University, Vellore, India

ABSTRACT

Disease forecasting using soft computing techniques is major area of research in data mining in recent years. To classify heart and diabetes diseases, this paper proposes a diagnosis system using cuckoo search optimized rough sets based attribute reduction and fuzzy logic system. The disease prediction is done as per the following steps 1) feature reduction using cuckoo search with rough set theory 2) Disease prediction using fuzzy logic system. The first step reduces the computational burden and enhances performance of fuzzy logic system. Second step is based on the fuzzy rules and membership functions which classifies the disease datasets. The authors have tested this approach on Cleveland, Hungarian, Switzerland heart disease data sets and a real-time diabetes dataset. The experimentation result demonstrates that the proposed algorithm outperforms the existing approaches.

KEYWORDS

Attribute Reduction, Cuckoo Search (CS) Optimization Algorithm, Disease Prediction, Fuzzy Logic System (FLS), Rough Sets Theory (RS)

1. INTRODUCTION

Data mining is the subfield in knowledge management. Data mining helps in healthcare for effective treatment, healthcare management, customer relation management, fraud and mistreatment detection and decision making (Silwattananusarn & Tuamsuk, 2012). In health care, Heart disease has considerably amplified for the past ten years and has become the foremost reason of death for people in most countries around the world. The structure or the function of the heart gets distressed by many characteristics of these heart diseases (Chitra & Seenivasagam, 2013). Computer program acknowledged as Medical Decision-Support System was anticipated to support health professionals formulate medical decision (Shortliffe, 1987).

In disease forecast, feature extraction and selection are important steps. An optimum feature set must have resourceful and perceptive characteristics; and also reduce the redundancy of features to avoid "curse of dimensionality" issue (Osareh & Shadgar, 2011). The impact of unrelated features on the presentation of classifier systems can be scrutinized by feature selection strategies (Acir,

DOI: 10.4018/IJFSA.2017040102

Ozdamar, & Guzelis, 2006; Valentini, Muselli, & Ruffino, 2004)). In this phase an optimal subset of features that are necessary are selected. By lessening the dimensionality and ignoring unrelated features, feature selection develops the exactness of algorithms (Zhang, Guo Du, & Li, 2005; Karabak & Ince, 2009). Traditional Principal Component Analysis (PCA) is one of the most frequently used feature extraction methods. It depends on extracting the axes on which data exhibit the maximum randomness (Jollife, 1986). Cluster analysis is a normally applied data mining method to scrutinize the relationships among attributes, samples and the relationships among attributes and samples. Hierarchical clustering tree (HCT) (Eisen, Spellman, Brown, & Botstein, 1998) and k-means (Tavazoie, Hughes, Campbell, Cho, & Church, 1999) are the two most well-known clustering techniques used to eliminate the features from the medical data's. Alternatively, the rough sets provide a proficient method of managing uncertainties and can be utilized for tasks such as data dependency study, feature identification, dimensionality reduction, and pattern categorization. Rough set theory (Pawlak, 1991; Polkowski, 2003) is a reasonably fresh intelligent method for managing ambiguity that is employed to find out data dependencies, to review the implication of attributes, detecting patterns in data, and to decrease redundancies.

Medical mining uses computerized tools and methods that assist in giving the merits to health systems. Particularly artificial intelligence methods are most frequently employed for disease diagnosis (Overiu & Simon, 2010; Rajeswari, Vaithiyanathan, & Amirtharaj, 2011; Srimani & Koti, 2014). The neural network classifier assists in diagnosing the diseases by creating a model employing feed forward neural network, multi-layer perceptron neural network, and back propagation neural network. Various intelligent systems have been enhanced for the purpose of developing health-care and offer better health care facilities, reduce cost. Data mining itself is an AI method that can be employed efficiently to improve the health care process (Hussain, Ishak, & Siraj, 2002). Neural Networks (NN) is a set of acquaintances of many straightforward processors or units (Hayashi, Setiono, & Yoshida, 2000; Kononenko, 2001). NN has been employed in various medical applications like coronary artery, Myocardial Infarction, cancer (Zhou, Jiang, Yang, & Chen, 2002; Karkanis, Magoulas, Grigoriadou, & Schurr, 1999), pneumonia and brain disorders (Pranckeviciene, 1999). Numerous machine learning methods have been utilized to classify the tumor, along with Fisher Linear Discriminate analysis (Dudoit, Fridyand, & Speed, 2002), k-nearest neighbour (Li, Darden, Weinberg, Levine, & Pederson, 2001) decision tree, multilayer perceptron (Khan et al., 2001), and hold up vector machine (Jong, Mary, Cornuejols, Marchiori, & Sebag, 2004). Gene selection (Xu, Wang, Zhang, Wang, & Feng, 2008) and neural networks (Berrar, Downes, & Dubitzky, 2003) dependent categorization were additionally added in microarray data analysis. Soft computing has been successfully employed in bioinformatics thus providing low cost, low, better approximation and positively good and more accurate solutions.

In this paper, we propose an efficient disease diagnosis model to predict heart and diabetes diseases more accurately with reduced number of attributes. In the proposed model, the cuckoo search (Gandomi, Hossein, Yang, & Alavi, 2013) with rough sets is introduced to reduce the set of attributes. The optimized attributes are used as input for the fuzzy logic system. The basic organization of the paper is as follows: Section 2 presents the review of related works. Background of the proposed algorithm is explained in section 3. The proposed disease prediction algorithm is explained in section 4 and result and analysis are shown in section 5. The conclusion part is presented in section 6.

2. RELATED WORKS

To forecast the heart and diabetes diseases, several researchers have recommended many methods in medical diagnosis. Nguyen Cong Long et al. (2015) have described the heart disease diagnosis system employing rough sets dependent characteristic minimization and interval type-2 fuzzy logic system (IT2FLS). IT2FLS uses a hybrid learning procedure consisting fuzzy c-mean clustering algorithm and parameters tuning by chaos firefly and genetic hybrid algorithms. This learning process is very costly when merged with high-dimensional dataset.

Santhanam and Ephzibah (2015) have presented the Heart Disease Prediction Using Hybrid Genetic Fuzzy Model. The genetic algorithm was employed for a stochastic search that gives the optimal solution to the feature selection. The relevant characteristics chosen from the dataset help the diagnosing system to grow a classification model employing fuzzy inference techniques. The rules for the fuzzy system were created from the sample data. The significant and relevant subset of rules is chosen by employing genetic algorithm from the pool of rule set. The chosen parameters were sex, serum cholesterol (chol), maximum heart rate obtained (thalach), Exercise induced angina (exang), ST depression induced by exercise relative to rest (oldpeak), number of major vessels coloured (ca) and thal value. Fuzzification employing Fuzzy Gaussian membership function and de-fuzzification using centroid technique enhances the performance of the system. The effort has been estimated using the performance parameters such as accuracy, specificity, sensitivity, confusion matrix that help in provide evidence for the effectiveness of the work.

Srinivas et al. (2014) have described the heart disease forecast based on rough-fuzzy classifier that merges rough set theory with the fuzzy set. The procedure of the rough-fuzzy classifier is (1) rule generation using rough set theory, and (2) prediction using fuzzy classifier. The experimentation was done by employing the Cleveland, Hungarian and Switzerland datasets.

Manjeevan Seera and Chee Peng Lim (2014) have discussed the hybrid intelligent system for medical data categorization. The hybrid intelligent system includes the Fuzzy Min–Max neural network, the Classification and Regression Tree, and the Random Forest model.

Employing Gene Expression Data, the SVM based Tumour Classification with Symmetry Non-Negative Matrix Factorization has been explained by Yuvaraj and Vivekanandan (2012). They selected genes by means of Nonnegative Matrix Factorization (NMF). Symmetry NMF (Sym-NMF) was employed in this method in order to develop the presentation of classification. Subsequently, features were extracted from the selected genes by virtue Sym-NMF. As a final step, a competent machine learning approach was applied to categorize the tumor samples by the extracted features. To obtain the improved classification Support Vector Machine with Weighted Kernel Width (WSVM) was employed. Similarly, Vafaie et al. (2014) have explained the Heart diseases prediction based on ECG signals' classification using a genetic-fuzzy system and dynamical model of ECG signals. In Nguyen et al. (2015) have explained the A highly accurate firefly based algorithm for heart disease prediction", journal of Expert Systems with Applications.

The two Neural Network methods have been explained by Kharat et al. (2012) for the classification of the magnetic resonance human brain images. The Neural Network method comprises of three stages, namely, feature extraction, dimensionality reduction, and classification. In the earlier stage, they obtained the features correlated with MRI images by using discrete wavelet transformation (DWT). In the successive stage, the parameters of magnetic resonance images (MRI) were reduced by utilizing the principles component analysis (PCA) to the more required features. In the categorization stage, two classifiers depending on supervised machine learning were used. The earlier classifier depends on feed forward artificial neural network (FF-ANN) and the successive classifier depends on Back-Propagation Neural Network. The classification of MRI brain images as normal or abnormal by using classifiers. Henriques et al. (2014) have explained the Prediction of Heart Failure Decompensation Events by Trend Analysis of Telemonitoring Data.

Healthcare community and researchers have proposed various methods for predicting diabetes to reduce the time, cost of tests and for higher level of prediction accuracy. A monitoring system for type 2 diabetes mellitus described by Wang and Kang (2008). This paper uses the following three algorithms, namely, DT for classifying and generating rules, it is relatively very fast and efficient in generating the rules, ANN is the thinking process which processes nonlinear problems, Back propagation neural network (BPNN) is most commonly used diagnosis and prediction problems which belong to supervised learning network, and Time Series helps in prediction procedure which refers to three major models namely Autoregressive (AR), Integrated (I) and Moving Average (MA), combination of these three models produce 2 hybrid models namely Autoregressive Moving Average

(ARMA) and Autoregressive Integrated Moving Average (ARIMA). Once the data pre-processing and feature selection is done using, ANN takes over the data and predicts the forecast and generate suggestions on the implementation of clinical procedures and diabetes control strategies if possible.

Reddy and Neelukhare (2016) have introduced an algorithm, FFBAT- Optimized Rule Based Fuzzy Logic Classifier, to classify the diabetes disease. In this paper, Locality Preserving Projection (LPP) algorithm is used for feature reduction and secondly classification of diabetes is done by means of RBFL classifier. LPP algorithm has identified the related attributes and then the fuzzy rules are produced from RBFL then rules are optimized using FFBAT algorithm. Next, the fuzzy system is designed with the help of optimized fuzzy rules and membership functions that will classify the diabetes data. FFBAT is the optimization algorithm which combines the features of BAT and Firefly (FF) optimization techniques.

Gandomi et al. (2013) introduced Cuckoo Search algorithm in combination with Lévy flights which is verified using a benchmark nonlinear constrained optimization problem. For the validation against structural engineering optimization problems, Cuckoo Search is subsequently applied to 13 design problems.

3. BACKGROUND OF THE ALGORITHM

3.1. Rough Set Theory

Rough sets theory (RST) was introduced by Pawlak (1994). It is used in analyzing intelligent schemes classified by undecided or unclear details. Attribute reduction is the major part in the analysis of RST. In this research, common notion of rough sets theory is reviewed as follows.

Consider $I = \left(U, A \cup \{d\}\right)$ be an information scheme, where U is the universe among a non-empty group of limited objects, A is a non-empty limited group of state attributes, and d is the decision feature (such a table is also known as decision table), $\forall a \in A$ there is an equivalent task $f_a : U \rightarrow V_a$, where V_a is the group of value of a. If $P \subseteq A$, the P-indiscernibility association is symbolized by $IND(P)$, is distinct as:

$$IND(P) = \left\{(x, y) \in U \middle| \forall a \in P, f_a(x) = f_a(y)\right\} \tag{1}$$

The separation of U produced by $IND(P)$ is symbolized U/P. If $(x, y) \in IND(P)$, then x and y are indiscernible by feature from P. After that similarity classes of the P-indiscernibility associations are symbolized by $[x]_P$. Consider $X \subseteq U$, the P-power approximation $\underline{PX}$ and P-upper approximation $\overline{PX}$ of set X-can be distinct as:

$$\underline{PX} = \left\{x \in U \middle\| [x]_P \not\subset X\right\} \tag{2}$$

$$\overline{PX} = \left\{xU \middle\| [x]_P\right\} \cap X \neq \varphi \tag{3}$$

Let P, $Q \subseteq A$ be equivalence relations over U, then the positive, negative and boundary regions can be defined as:

$$POS_P\left(Q\right) = \bigcup_{X\in U|Q} \underline{PX} \tag{4}$$

$$NEG_P\left(Q\right) = U - \bigcup_{X\in U|Q} \overline{PX} \tag{5}$$

$$BND_P\left(Q\right) = \bigcup_{X\in U|Q} \overline{PX} - \bigcup_{X\in U|Q} \underline{PX} \tag{6}$$

The optimistic section of the separation $U\left|Q\right.$ with corresponding to P, $POS_P\left(Q\right)$, is the group of every objects of U that can be positively categorized to obstruct of the separation $U\left|Q\right.$ by means of P. Q Reliant on P in a level $k\left(0\le k\le 1\right)$ symbolized by $P\Rightarrow_k Q$:

$$\gamma_P\left(Q\right) = \frac{\left|POS_P\left(Q\right)\right|}{\left|U\right|} \tag{7}$$

where; P is a set of conditional attributes, Q is the decision and $\gamma_p\left(Q\right)$ is the quality of classification. If k=1, Q reliant entirely on P; if 0<k<1, Q oriented incompletely on P; and if k=0 the Q does not based on P. $\left|.\right|$ represent the cardinality of a set. The aim of feature diminution is to eradicate unnecessary features therefore that the condensed group offers the identical superiority of categorization as the innovative. The group of every reduct is distinct as:

$$Red\left(C\right) = \left\{R\subseteq C\left|\gamma_R\left(D\right) = \gamma_C\left(D\right), \forall B\subset R, \gamma_B\left(D\right)\neq\gamma_C\left(D\right)\right.\right\} \tag{8}$$

A set of minimal reductions is defined as:

$$Red\left(C\right)_{\min} = \left\{R\in Red\left|\forall R'\in Red, \left|R\right|\le\left|R'\right|\right.\right\} \tag{9}$$

Determining the smallest feature diminution for a specified data group is an NP-hard difficulty. Consequently, increasing a competent algorithm for this difficulty is a demanding assignment. Depending on those notions of rough sets a suitable fitness utility of the CS algorithm.

3.2. Cuckoo Search Algorithm

Cuckoo search is an optimization algorithm developed by Xin-she Yang and Suash Deb in 2009. It was motivated by the obligate brood parasitism of some cuckoo species by laying their eggs in the nests of other host birds. Some host birds can engage direct conflict with the intruding cuckoos. For example, if a host bird discovers the eggs are not their own, it will either throw these alien eggs away or simply abandon its nest and build a new nest elsewhere. Some cuckoo species have evolved in such a way that female parasitic cuckoos are often very specialized in the mimicry in colors and pattern of the eggs of a few chosen host species. Cuckoo search idealized such breeding behavior, and thus can be applied for various optimization problems.

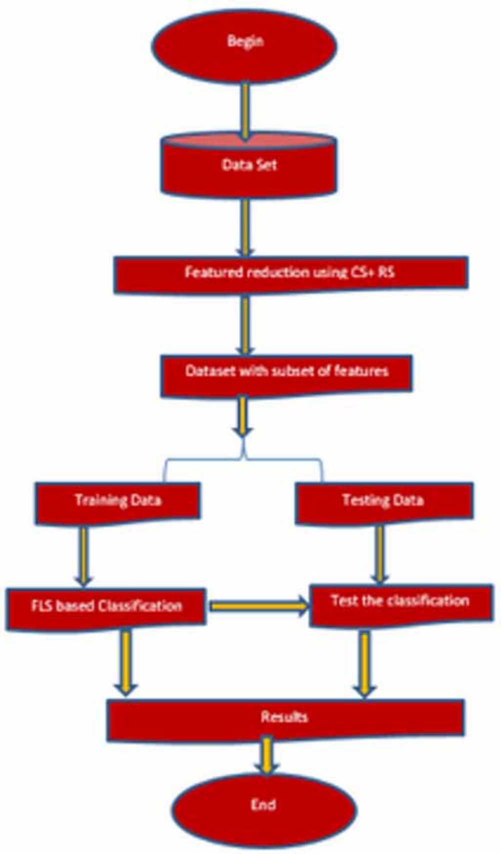
4. AN APPROACH FOR PROPOSED HEART AND DIABETES DISEASE PREDICTION

The prediction of disease has turned out to be an ever more demanding difficulty. Feature extraction or selection process is a major element of medical data categorization. Due to feature reduction process, the computation cost decreases and the classification performance can increase. In order to achieve promising results in disease prediction, in this paper, we have planned to utilize novel feature reduction method and FL classifier. The disease prediction system consists of three major steps: Normalization, Cuckoo Search algorithm and rough sets based attribute reduction and fuzzy logic system based classification. Initially the input dataset D^D is normalized to the range [0, 1]. The Cuckoo Search (CS) algorithm and rough set oriented attribute selection technique is applied on D^D, to pick the finest subset of attributes. The obtained reduced subset of attributes is divided into two subsets: training dataset T^R and testing dataset T^S. The training dataset T^R is fed to the fuzzy logic system while the testing dataset T^S is used to test the obtained fuzzy logic system. The overall process of proposed heart disease and diabetes prediction system is illustrated in Figure 1. The details of each stage are discussed in the following subsections.

4.1. Stage 1: Normalization

Consider the dataset D^D, which has M number of attributes and N number of enties. Every characteristic encompasses dissimilar arithmetical values which increases the computational

Figure 1. Process of proposed heart disease and diabetes prediction system



complexity. Therefore, we applied Normalization process to dataset D^D. This process converts the dataset into specific interval. In this paper, we used a popular method min-max normalization method for normalization process. Min-max normalization maps a value D of original dataset to D'' in range of $[\text{new\_min}, \text{ new\_max}]$ by calculating Equation (10):

$$D'' = \frac{D - D_{\min}}{D_{\max} - D_{\min}} \times [\text{new\_min} - \text{new\_max}] + \text{new\_min} \tag{10}$$

where new_min, new_max describe the range of transform dataset. Here, we employed $\text{new\_min} = 0$ and $\text{new\_max} = 1$. Subsequent to normalization, transformed datasets are utilized for attribute reduction method. The detailed algorithm for attribute reduction is introduced in next subsection.

4.2. Stage 2: Attribute Reduction Based on Cuckoo Search and Rough Sets

The basic idea of this section is to reduce the dimension of the features using Cuckoo Search with Rough sets (CS+RS) algorithm. Attribute dimension reduction method is mandatory to diminish the attributes' space without losing the precision of prediction. In addition, we reduce the amount of attributes and eliminate unrelated, unconnected, redundant or noisy details.

4.2.1. Step 1: Solution Representation

The probable solution of this scheme (attribute reduction) is symbolized as binary strings of length M. In those binary strings, each bit symbolize attributes, the value "1" demonstrate that the equivalent attribute is selected, the value "0" express that the equivalent attribute is not selected. For example, the dataset has 10 attributes (a_1, a_2, a_3… a_{10}) and a solution Y=1010110010, then the attributes selected are (a_1, a_3, a_5, a_6, a_9).

4.2.2. Step 2: Fitness Function

Before applying the CS algorithm, the fitness calculation of the preliminary solution is initially carried out. The fitness function generates a fitness value of each solution. Based on the fitness value we select the best solution. For this we used the fitness function as specified in Equation (11):

$$Fitness(F_F) = \frac{M - |X|}{M} + \frac{n|R|\gamma_X(D)}{M\Gamma} \tag{11}$$

where $M = |C|$, $n = |U|$ $_{,}$ $\gamma_X(D)$ is quality of classification, R is a reduct of condition attribute C, R is computed by rough set theory (Chitra & Seenivasagam, 2013). $\Gamma = |Y_1 + Y_2|$ if the decision table DT is consistent and $\Gamma = |Y_1|$ if DT is not consistent. The goodness of position of each solution is evaluated by this fitness function.

4.2.3. Step 3: Apply Cuckoo Search

After the fitness calculation, we update the solution based on Cuckoo Search algorithm. Here, at first Set number of iterations to process Cuckoo Search algorithm. The solution is shaped as illustrated in the following Equation 12:

$$X_i^{t+1} = X_i^t + \alpha \oplus V_i^t \tag{12}$$

where:

X_i^t → Characterizes the current solution i of iteration t

V_i^t → Indicates the velocity of iteration t

α → Signifies the step size related to problem search space

The updation will be done by the equation given below:

$$V_i^{t+1} = \omega V_i^t + c_1 r_1 [X_i^t - P_i^t] + c_2 r_2 [X_i^t - G_i^t] \tag{13}$$

where, ω indicates the inertia weight to avert the unrestrained development of the velocity of the particle, r_1 and r_2 represents the evenly disseminated arbitrary number in the range of [0, 1], c_1 and c_2 correspond to the learning factors.

4.2.4. Step 4: Termination Criteria

The algorithm stops its implementation only if maximum number of iterations is attained and the solution which is containing the best fitness value is chosen using CS and it has to be classified. As explained best attribute is utilized as an input of a fuzzy classifier, after training process, we can decide if the data under test is disease or not. The pseudo code of proposed feature reduction approach is given in Algorithm 1.

4.3. Stage 3 Prediction Based on Fuzzy Logic System

Once we reduce the features from the input dataset, the disease prediction is done by fuzzy logic classifier. Fundamentally, the fuzzy logic classifier scheme contains three main steps:

1. Fuzzification
2. Fuzzy inference engine
3. De-fuzzification

4.3.1. Fuzzy Inference System

A fuzzy inference system aids in recording the inputs to the equivalent output by predefined fuzzy rules obtainable in the knowledge support. The knowledge support includes if-then rules that denote the relationship among the input and output fuzzy groups. The inference system is enhanced by a sequence of actions like:

- Developing the fuzzy rules;
- Fuzzifying the input values depend on the membership function;
- Merge the fuzzy input and the fuzzy rules to produce the rule strength;
- The regulation power effect is once more mutual among the output attachment task to produce the output allotment;
- Finally, the output is de-fuzzified to give the output in crisp value.

4.3.2. Membership Function

A membership function (MF) is a transforms the input data to a membership value (or degree of membership) among 0 and 1. We have chosen the triangular membership task to modify the input data into the fuzzified value. The Triangular membership task includes three vertices i, j and k of

Algorithm 1. Pseudo code of Cuckoo Search with rough set based attribute reduction algorithm

```
Objective function:

   {\displaystyle f(\mathbf {x}),\quad \mathbf {x} =(x_{1},x_{2},\dots, x_{d});\,}

f(x), x= (x1, x2, x3, … $x_d$)
Generate an initial population of

   {\displaystyle n}

n host nests;
While (t<MaxGeneration) or (stop criterion)
  Get a cuckoo randomly (say, i) and replace its solution by performing Lévy flights;
  Evaluate its quality/fitness

   {\displaystyle F_{i}}

$F_i$
     [For maximization, $F_i$

   {\displaystyle F_{i}\propto f(\mathbf {x} _{i})}
 α f($x_i$)];
  Choose a nest among n (say, j) randomly;
  if ($F_i$

   {\displaystyle F_{i}>F_{j}}

> $F_j$),
     Replace j by the new solution;
  end if
  A fraction (

   {\displaystyle p_{a}}

$p_a$) of the worse nests are abandoned and new ones are built;
  Keep the best solutions/nests;
  Rank the solutions/nests and find the current best;
  Pass the current best solutions to the next generation;
End while
```

$f(x)$ in a fuzzy set A (i : minor edge, k : better edge where membership scale is zero and j : the centre where membership scale is 1). The principle engaged to analyze the membership values is illustrated below:

$$f\left(x\right)=\begin{cases} 0 & if\ x \le i \\ \dfrac{x-i}{j-i} & if\ i \le x \le j \\ \dfrac{k-x}{k-j} & if\ j \le x \le k \\ 0 & if\ x \ge k \end{cases} \tag{14}$$

4.3.3. Fuzzy Rule Generation

The fuzzy rule generation is a very important mission that assists in recording the input to its equivalent output. Rules can be enclosed by several techniques that supply a precursor and consequential outline as specified beneath: If A_1, A_2… A_N are the attributes and C_1, C_2 is the class labels then a fuzzy rule can be framed based on the linguistic values like high, medium, low. The values N and M are the number of attributes and number of classes respectively. Therefore, the fuzzy rule can be framed as follows:

- If A_1 is high and A_2 is low and A_3 is medium, then class is C_2;
- If A_1 is low and A_2 is medium and A_3 is medium, then class is C_1;
- If A_1 is high and A_2 is medium and A_3 is low, then class is C_2.

4.3.4. Rule Based Fuzzy Score Computation

The testing data $\left(D^{TE}\right)$ with reduced attribute $\left(N\right)$ is fed to the fuzzy logic system, where the test data is converted to the fuzzified value based on the fuzzy membership function. Then, the fuzzified input is matched with the fuzzy rules defined in the rule base. Here, the rule inference procedure is used to obtain the linguistic value that is then converted to the fuzzy score using the average weighted method. From the fuzzy score obtained, the classification decision is produced.

5. RESULTS AND DISCUSSION

In this section, we discuss the results obtained from the proposed technique. For implementing the proposed technique, we have used Mat lab version (7.12). This proposed technique is done in windows machine having Intel Core i5 processor with speed 1.6 GHz and 8 GB RAM. In this work, we have evaluated the accuracy of proposed method using three different data sets from UCI machine learning repository and one diabetes dataset.

5.1. Evaluation Metrics

The evaluation of proposed heart/diabetes disease prediction technique is carried out using the following metrics as suggested by below equations:

- **Sensitivity:** The sensitivity of disease prediction is determined by taking the ratio of number of true positives to the sum of true positive and false negative. This relation can be expressed as:

$$S_t = \frac{T_p}{T_p + F_n}$$

- **Specificity:** The specificity of the disease prediction can be evaluated by taking the relation of number of true negatives to the combined true negative and the false positive. The specificity can be expressed as:

$$S_p = \frac{T_n}{T_n + F_p}$$

- **Accuracy:** The accuracy of disease prediction can be calculated by taking the ratio of true values present in the population. The accuracy can be described by the following equation:

$$A = \frac{T_p + T_n}{T_p + F_p + F_n + T_n}$$

where:

T_p → True positive

T_n →True negative

F_p → False positive

F_n →False negative

5.2. Dataset Description

The proposed system is experimented with the four datasets: Cleveland, Hungarian, Switzerland heart disease datasets taken from UCI machine learning repository and Real time diabetes dataset. The real-time diabetes database was collected from the Sree Diabetic Care Center, Kurnool in Andhra Pradesh for the experimental analysis. The data was collected between September 2014 to January 2015 from the hospital database. The patients were between 18 and 77 years old for this study. Furthermore, information collected and recorded included patient's information age, sex, BMI, BP, cholesterol, LDL, HDL, Triglycerides, FBS, Smoking, alcohol, family history of diabetics, heart disease.

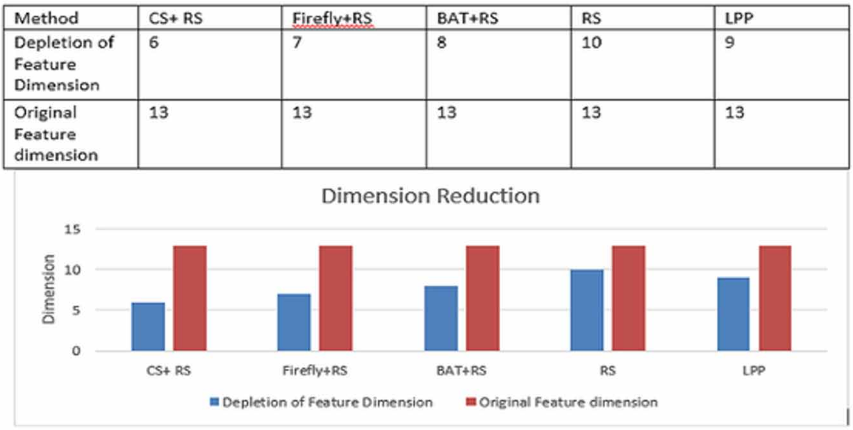
5.3. Performance Evaluation of Proposed Approach

5.3.1. Performance Based on Feature Reduction

Feature reduction is an important stage for disease prediction, because high numbers of features are the great obstacle of the classification. In this approach, we used Cuckoo Search with Rough set theory (CS+RS with other approaches such as, firefly+RS, BAT+RS, RS and LPP. Figure 3 shows the performance of feature reduction approach based on heart/diabetes disease prediction.

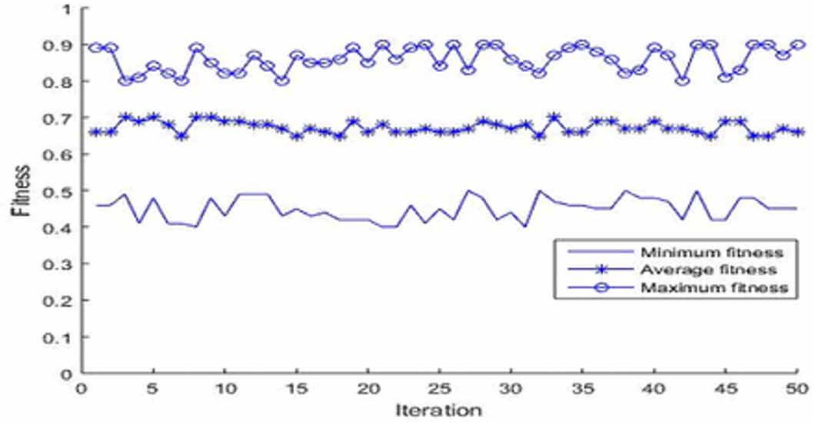
Figure 2 shows the comparative analysis of different dimension reduction approaches. In this proposed approach we used CS+RS algorithm for feature reduction. When analyzing Figure 2, using firefly with rough set theory for attributes reduction we obtained seven attributes out of 13. When we use BAT with rough set theory for feature reduction we obtained eight attributes out of 13. When we using rough set theory without any optimization approach we obtain the 10 attributes out of 13. Using LPP based feature reduction approach we obtained nine attributes. Compared to all the works our proposed feature reduction approach selects only five attributes out of 13. Figure 3 shows the fitness performance of the proposed approach.

Figure 2. Comparative analysis of different dimension reduction approaches



Method	CS+ RS	Firefly+RS	BAT+RS	RS	LPP
Depletion of Feature Dimension	6	7	8	10	9
Original Feature dimension	13	13	13	13	13

Figure 3. Fitness performance of the proposed approach



5.3.2. Performance Based on Classification

Here, we used different types of classifiers to perform heart/diabetes disease prediction. We have constantly set the feature extraction approach (CS+RS) and change the classifier to predict the heart diseases. The Figures 4-6 shows the performance of disease prediction.

Figure 4 shows the performance of proposed approach based on accuracy for different datasets. The effectiveness of the proposed technique is demonstrated by performing a comparison between the matching result of the proposed method of EISOS+WRC and the other methods. The methods used network (NN) and support vector machine (SVM) are the best known among existing schemes for heart disease classification. Furthermore, they characterize local details of the patient data, variation representation. Therefore, we have chosen to compare the performance of our proposed algorithm against these approaches. When analyzing Figure 4, using our proposed approach (CS+RS) +FL we obtain the maximum accuracy of 91% for Cleveland dataset, 91.5% for Hungarian dataset and 90% for Switzerland dataset and 89.5% for Real Time Data Set. Similarly, using (CS+RS) +NN we obtain the maximum accuracy of 85% for using Cleveland, 86% for using Hungarian, 89% for using Switzerland and 87.5% for using real time dataset. Likewise, using (CS+RS) +SVM we obtain the maximum accuracy of 84% for using Cleveland, 86.3% for using Hungarian and 77% for using Switzerland dataset and 69% for Real Time Dataset. Moreover, Figure 5 shows the performance

Figure 4. Performance of proposed approach based on accuracy for different datasets

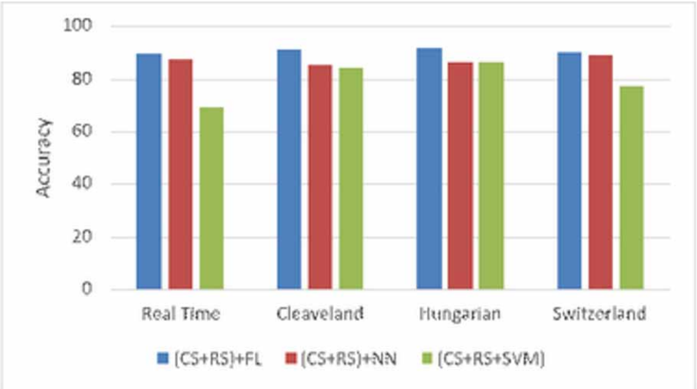


Figure 5. Performance of proposed approach based on sensitivity

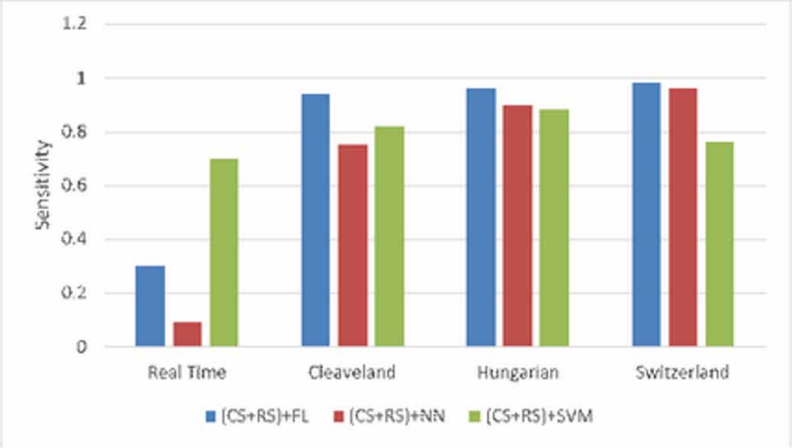
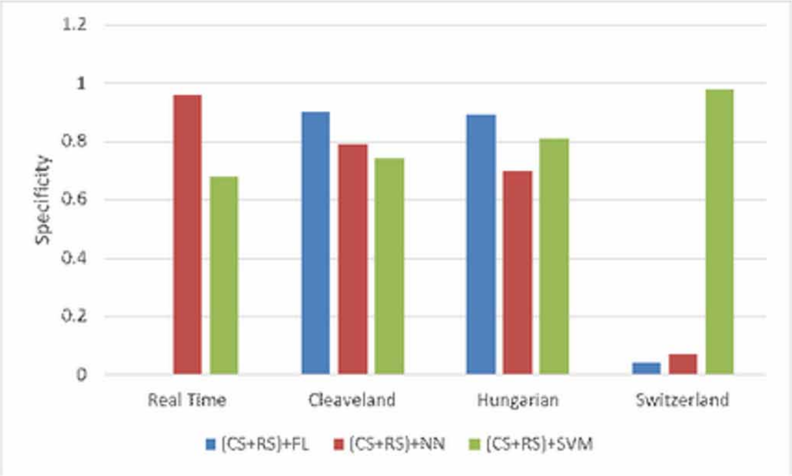


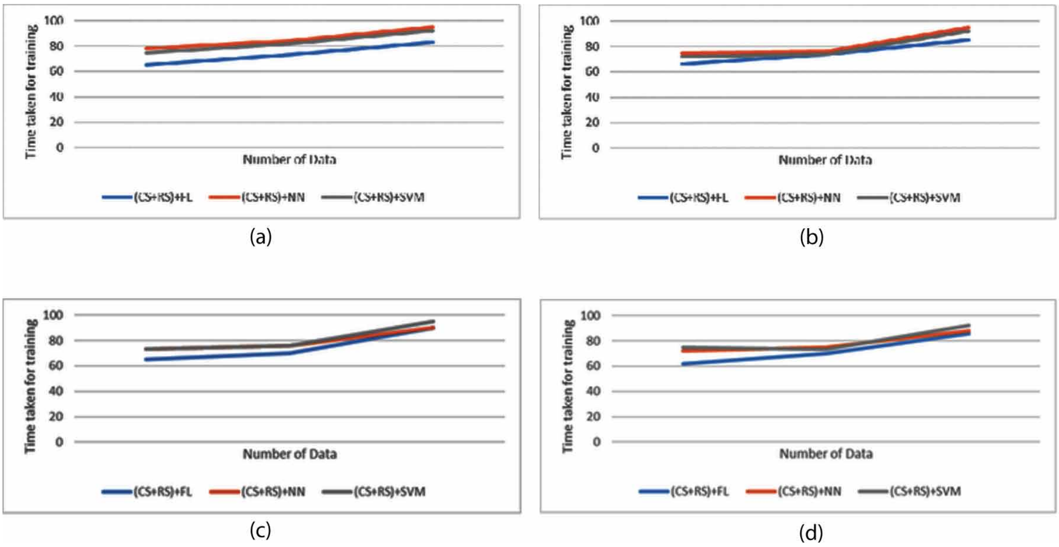
Figure 6. Performance of proposed approach based on specificity



of proposed approach based on sensitivity. Here, we obtain the maximum sensitivity of 98% for Switzerland dataset which is high compare to the other approaches. Similarly, Figure 6 shows the Performance of proposed approach based on specificity. Here, also our proposed approach obtains maximum result. From these figures, we clearly understand our proposed approach achieves the better performance compare to the other approaches.

Figures 7(a), 7(b), 7(c) and 7(d) represents the time complexity performance of the three algorithm ((CS+RS) +FL, (CS+RS) +NN and (CS+RS) +SVM)) for the dataset Cleveland, Hungarian, Switzerland and Real Time respectively with 50 iterations. The amount of time needed by the system to run to completion referred to as Time complexity, it is depending on the size of the input. From Figure 7 it could be inferred that as data sizes were increasing, the time taken for classification stage also increases. From the figure we understand our proposed work takes minimum time to complete the work compare to (CS+RS) +NN and (CS+RS) +SVM)). The approach (CS+RS) +NN takes highest time than the (CS+RS) +FL and (CS+RS) +SVM.

Figure 7. (a, b, c, d) represent the time complexity performance of the three algorithms



5.4. Comparison of the Other Approaches

Heart disease and diabetes prediction based on novel feature reduction and classification. In classification we used CS+RS algorithm is used and in classification we used fuzzy logic classifier. In previous work (Silwattananusarn & Tuamsuk, 2012), we have explained LPP algorithm for feature reduction and CS+RBFL for prediction. Table 1 shows the comparative analyzes of proposed approach based on accuracy sensitivity and specificity measures. The table proves that our approach outperforms the existing approaches.

6. CONCLUSION

A system that supports the physicians for perfect prediction of heart and diabetes disease in patients has been developed using computing techniques like optimized rough set theory and fuzzy logic. Among various classification and prediction models this model is evaluated to be better. An optimized rough set theory has been used to perform a stochastic search on the dataset to reduce the number

Table 1. Comparative analysis of proposed approach with other approaches based on accuracy, sensitivity and specificity

Methods	Cleveland			Hungarian			Switzerland			Real Time		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
Proposed (CS+RS)+RS	91	94	90	91.5	96	89	90	98	40	89.5	30	97
Previous (LPP+RBFL)	68	79	84	67	87	38	72	76	67	64	87	70
Existing (RS+FL)	72.6	100	0	69.7	86	35	63.4	67	72	63.7	86	35

of attributes. The fuzzy inference system predicts the test data with the help of fuzzy triangular membership function and de-fuzzification method. Finally, the experimentation is carried out using the Cleveland, Hungarian and Switzerland heart disease datasets and real time diabetes dataset. The performance was analyzed with sensitivity, specificity and accuracy. Our experimentation results show that our approach outperformed the existing approaches. In future works space complexity can be taken into consideration for overall performance of the proposed method.

REFERENCES

Acır, N., Ozdamar, O., & Guzelis, C. (2006). Automatic classification of auditory brainstem responses using SVM-based feature selection algorithm for threshold detection.Journal of Engineering Applications of Artificial Intelligence, 19(2), 209-218.

Berrar, D. P., Downes, C. S., & Dubitzky, W. (2003). Multiclass Cancer Classification Using Gene Expression Profiling and Probabilistic Neural Networks. *Journal of Pacific Symposium on Bio computing*, 8, 5-16.

Chitra, R., & Seenivasagam, V. (2013). Review of heart disease prediction system using data mining and hybrid intelligent techniques.Journal on soft computing, 3(4).

Dudoit, S., Fridlyand, J., & Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, *97*(457), 77–87. doi:10.1198/016214502753479248

Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(25), 14863–14868. doi:10.1073/pnas.95.25.14863 PMID:9843981

Gandomi, A. H., Yang, X.-S., & Alavi, A. H. (2013). Cuckoo search algorithm: A metaheuristic approach to solve structural optimization problems. *Engineering with Computers*, *29*(1), 17–35. doi:10.1007/s00366-011-0241-y

Han, J., Rodriguez, J. C., & Beheshti, M. (2008, December). Diabetes data analysis and prediction model discovery using RapidMiner. *Proc. 2nd IEEE Int. Conf. Future Generation Commun. and Networking* (Vol. 3, pp. 96-99).

Hayashi, Y., Setiono, R., & Yoshida, K. (2000). A comparison between two neural network rule extraction techniques for the diagnosis of hepatobiliary disorders. *Artificial Intelligence in Medicine*, *20*(3), 205–216. doi:10.1016/S0933-3657(00)00064-6 PMID:10998587

Henriques, J., Carvalho, P., Paredes, S., Rocha, T., Habetha, J., Antunes, M., & Morais, J. (2014). Prediction of Heart Failure Decompensation Events by Trend Analysis of Telemonitoring Data. *IEEE Journal of Biomedical and Health Informatics*, *19*(5), 1757–1769. doi:10.1109/JBHI.2014.2358715 PMID:25248206

Hussain, W., Ishak, W., & Siraj, F. (2002). Artificial Intelligence in Medical Application: An Exploration", journal of Health Informatics Europe, vol.16, .

Jollife, I. T. (1986). *Principal Component Analysis*. New York: Springer-Verlag. doi:10.1007/978-1-4757-1904-8

Jong, K., Mary, J., Cornuejols, A., Marchiori, E., & Sebag, M. (2004). Ensemble feature ranking.*Proceedings Eur. Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*.

Karabatak, M., & Ince, M. C. (2009). An expert system for detection of breast cancer based on association rules and neural network. Journal of Expert Systems with Applications, 36(2), 3465-3469.

Karkanis, S., Magoulas, G. D., Grigoriadou, M., & Schurr, M. (1999). Detecting Abnormalities in Colonoscopic Images by Textual Description and Neural Networks. Proceedings of Machine Learning and Applications: Machine Learning in Medical Applications, Chania, Greece (pp. 59-62).

Khan, J., Wei, J. S., & Ringner, M. et al.. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, *7*(673), 679. PMID:11385503

Kharat, K. D., Kulkarni, P. P., & Nagori, M. B. (2012). Brain Tumor Classification Using Neural Network Based Methods. *International Journal of Computer Science and Informatics, Vol*, *1*(4).

Kononenko, I. (2001). Machine learning for medical diagnosis: History, state of the art and perspective. *Artificial Intelligence in Medicine*, *23*(1), 89–109. doi:10.1016/S0933-3657(01)00077-X PMID:11470218

Li, L., Darden, T.A., Weinberg, C.R., Levine. A.J., & Pedersen, L.G. (2001). Gene assessment and sample classification for Gene expression data using a genetic algorithm and k-nearest neighbor method. *Combinational chemistry and high throughput screening*, 4(8), 727-739.

Long, N.C., Meesad, P., & Unger, H. (2015). A highly accurate firefly based algorithm for heart disease prediction. *Journal of Expert Systems with Applications*.

Long, N. C., Meesad, P., & Unger, H. (2015). A highly accurate firefly based algorithm for heart disease prediction.Journal of Expert Systems with Applications, 42(21), 8221–8231.

Osareh, A., & Shadgar, B. (2011). A Computer Aided Diagnosis System for Breast Cancer. *International Journal of Computer Science Issues*, *8*(2).

Overiu, M., & Simon, D. (2010). Biogeography-based optimization of neuro-fuzzy system parameters for diagnosis of cardiac disease. *Proceedings of the 12th annual conference on Genetic and evolutionary computation* (pp. 1235–1242). doi:10.1145/1830483.1830706

Pawlak, Z. (1991). *Rough Sets- Theoretical Aspects of Reasoning about Data*.

Pawlak, Z., & Slowinski, R. (1994). Rough set approach to multi-attribute decision analysis. *European Journal of Operational Research*, *72*(3), 443–459. doi:10.1016/0377-2217(94)90415-4

Polkowski, L. (2003). *Rough Sets- Mathematical Foundations*. Physica-Verlag Heidelberg.

Pranckeviciene. (1999). Finding Similarities Between An Activity of the Different EEG's by means of a Single layer Perceptron. *Proceedings of Machine Learning and Applications: Machine Learning in Medical Applications* (pp. 49-52).

Rajeswari, K., Vaithiyanathan, V., & Amirtharaj, P. (2011). Prediction of risk score for heart disease in India using machine intelligence. *Proceedings of theInternational Conference on Information and Network Technology* (pp. 19–22).

Reddy, G. T., & Khare, N. (2017). An Efficient System for Heart Disease Prediction Using Hybrid OFBAT with Rule-Based Fuzzy Logic Model. *Journal of Circuits, Systems, and Computers*, *26*(04), 1750061. doi:10.1142/S021812661750061X

Santhanam, T., & Ephzibah, E. P. (2015). Heart Disease Prediction Using Hybrid Genetic Fuzzy Model. *Journal of Indian Journal of Science and Technology*, *8*(9), 797–803. doi:10.17485/ijst/2015/v8i9/52930

Seera, M., & Lim, C. P. (2014). A hybrid intelligent system for medical data classification. *Expert Systems with Applications*, *41*(5), 2239–2249. doi:10.1016/j.eswa.2013.09.022

Shortliffe, E. H. (1987). Computer Programs to Support Clinical Decision Making. *Journal of the American Medical Association*, *258*(1), 61. doi:10.1001/jama.1987.03400010065029 PMID:3586293

Silwattananusarn, T., & Tuamsuk, K. (2012). Data Mining and Its Applications For Knowledge Management: A Literature Review from 2007 to 2012. *International Journal of Data Mining & Knowledge Management Process*, *2*(5), 13–24. doi:10.5121/ijdkp.2012.2502

Srimani, P. K., & Koti, M. S. (2014). Knowledge discovery in medical data by using rough set rule induction algorithms. *Indian Journal of Science and Technology*, *7*(7), 905–915.

Srinivas, K., Raghavendra Rao, G., & Govardhan, A. (2014). *Rough-Fuzzy Classifier: A System to Predict the Heart Disease by Blending Two Different Set Theories* (Vol. 39). Arabian Journal for Science and Engineering.

Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J., & Church, G.M. (1999). Systematic determination of genetic network architecture. *Journal of Nature genetics*, 22(3), 281-285.

Thippa Reddy, G., & Khare, N. (2016). FFBAT- Optimized Rule Based Fuzzy Logic Classifier for Diabetes. *International Journal of Engineering Research in Africa*, *24*, 137–151. doi:10.4028/www.scientific.net/JERA.24.137

Vafaie, M. H., Ataei, M., & Koofigar, H. R. (2014). Heart diseases prediction based on ECG signals' classification using a genetic-fuzzy system and dynamical model of ECG signals. *Biomedical Signal Processing and Control*, *14*(291–296).

Valentini, G., Muselli, M., & Ruffino, F. (2004). Cancer recognition with bagged ensembles of support vector machines. *Journal of Neuro computing*, 56, 461-466.

Xu, W., Wang, M., Zhang, X., Wang, L., & Feng, H. (2008). SDED: A novel filter method for cancer-related gene selection.Journal of Bio information, 2(7), 301-303.

Yang, X.-S., & Deb, S. (2009). Cuckoo search via Lévy flights. Proceedings of the World Congress on Nature & Biologically Inspired Computing NaBIC ‘09. IEEE.

Yuvaraj, N., & Vivekanandan, P. (2013). An Efficient SVM based Tumor Classification with Symmetry Non-Negative Matrix Factorization Using Gene Expression Data. In Information Communication and Embedded Systems (pp. 761–768). doi:10.1109/ICICES.2013.6508193

Zhang, Y. L., Guo, N., Du, H., & Li, W. H. (2005). Automated defect recognition of C- SAM images in IC packaging using Support Vector Machines. *International Journal of Advanced Manufacturing Technology*, *25*(11-12), 1191–1196. doi:10.1007/s00170-003-1942-1

Zhou, Z. H., Jiang, Y., Yang, Y. B., & Chen, S. F. (2002). Lung cancer cell identification based on artificial neural network ensembles. Journal of Artificial Intelligence in Medicine, 24(1), 25-36.

Thippa Reddy Gadekallu has received his BTech in Computer Science and Engineering from Nagarjuna University, India in the year 2003, MTech in Computer Science and Engineering from Anna University, India in the Year 2010, and currently perusing his PhD from VIT University, India. He is working as an Assistant professor in VIT University, India. His research interests are Data Mining in Healthcare, Natural Language Processing, Knowledge Mining etc.

Neelu Khare has received her PhD in the year 2011 from NIT Bhopal, India, she is currently working as Associate Professor at School of Information Technology and Engineering in VIT University, India. She has around 20 publications in International Journals and Conferences. Her research areas are Data Mining- Association and classification, Bio Informatics, Soft Computing.