FEATURE SELECTION AND FUZZY EXTREME LEARNING MACHINE (FELM) CLASSIFIER FOR HEART DISEASE DIAGNOSIS

G. Sunitha, *Research Scholar, Department of Computer Science, Rayalaseema University, Kurnool. E-Mail: gundasunitha01@gmail.com*
Dr.N. Geethanjali, *Professor, Department of Computer Science & Technology, Sri Krishnadevaraya University, Anantapur.*

ABSTRACT: The considerable growing of cardiovascular disease and its effects and complications as well as the high costs on society makes medical community seek for solutions to prevention, early identification and effective treatment with lower costs. Thus, valuable knowledge can be established by using artificial intelligence and data mining; the discovered knowledge makes improve the quality of service. Until now, different researches have been carried out in order to predict heart disease based on data mining methods such as classification and feature selection methods; however, what has been less noticed is the exact diagnosis of disease with the lowest cost and time. Early detection and treatment of heart disease will reduce the patient mortality rate. Accordingly, herein propose a Particle Swarm Optimization (PSO) and highly accurate hybrid Fuzzy Extreme Learning Machine (FELM) method for the diagnosis of coronary artery disease. The proposed FELM based prediction model is able to detect coronary artery disease based on clinical data without the need for invasive diagnostic methods. Making use of such methodology, we achieved good accuracy, sensitivity and specificity rates on Z-Alizadeh Sani dataset.

Keywords: ***Diagnosis Systems, Heart Disease, Feature Selection, Fuzzy Extreme Learning, Machine Learning, Coronary Artery Disease.***

1. INTRODUCTION

Data mining is the process of examining the raw data set to generate useful information. It looks for pattern of data according to different perspectives [1]. It is capable of delivering the effective course of action by comparing and contrasting the data.The data mining is largely used in retail, financial communication and marketing organization to drill down their transactional data. Machine learning plays a huge role in medical field [7].

The medical domain is the great beneficiary of data mining where huge data are handled easily [1,5,7]. There still prevails a great challenging common disease called Coronary Artery Disease (CAD) which takes off more life in the world. The deadly disease is caused due to the constriction of the blood vessels in the heart. The Angiography is the

widely used diagnosis method which might place the life at risk by causing cancer and other health issues [1].

Therefore a new hybrid model[1] based on artificial neural network and genetic algorithm has been proposed here and is used for detecting the heart disease without employing the angiography. The neural network helps to improve the performance and the genetic algorithm provides the best possible outcome as solution.The rest of the paper is organized as follows. Section 2 discusses the literature review. Section 3 overviews the proposed technique for feature selection and classification. Experimental results of the proposed scheme are presented in Section 4. Concluding remarks with future work are covered in Section 5.

2. LITERATURE REVIEW

ZeinabArabasadiet al., [1] provided the model for diagnosis of heart disease by the data mining technique. The data is retrieved from Z- Alizadeh Sani dataset which contains record of 303 patients out of which 216 where suffered from CAD. The feature process consists of four ranking methods namely Gini index, weight by SVM, information gain and Principal component analysis. Thus the model is based on artificial neural network and genetic algorithm to diagnose the heart disease by clinical means there by eliminating the angiography.

Kale et al., [2] proposed a methodology for handling weighted classification problem and Feature Subset Selection (FSS). They used hybrid FSS approach for F-ELM classifier called H-FELM is designed for clinical dataset. This resulting in H-FELM has the ability to handle problems by selecting features which is related and no redundant. This leads to reduction of computation overhead and minimizing the learning time. In order to validate the efficiency and effectiveness of the proposed algorithm, the comparative performance has been evaluated using different combinations of FELM, FSS, and ELM. Finally these comparisons resulted in validation of effectiveness and accuracy.

Kavipriyaet al., [3] proposed the system that uses comprehensive feature selection which is applicable for selection of attributes that had chosen the dataset of PIMA Indian Diabetes. This mechanism uses the large significance pattern to select the leading features. It also provides way for quantifying feature and target variable, and then informativeness has been identified based on relevant score. Then the SVM classifier is applied over feature selection

for predicting the heart diseases. Performance metrics based on various criteria are taken into account and provides the better performances.

In the paper [4],Chunhong Luet al., had used trace-based separable criterion developed a selection algorithm on the basis of Genetic Algorithm (GA) .In accordance with scores from class and variable separable the significance of feature subset is measured by criterion which is independent of any classification. These processes use the dataset of lung cancer. The solution obtained are undergone verification based on three classifiers such as SVM, BPNN, KNN and being compared with dataset Finally the comparatively obtained result brings out this proposed system has been providing better lung cancer diagnosis.

Xiao Liuet al., [5] proposed the system that uses the Rough Set- (RFRS-) and novel ReliefF based for diagnosis the heart disease. It includes the Systems such as classification and RFRS feature selection subsystem. The Hear dataset has been used for testing. It brought out that superior performance is acquired by RFRS base on specificity and sensitivity. Thus proposed system provides the tool for heart disease diagnosis based on empirical analysis.

Kale et al., [6] bring out the system called Improved Genetic – PSO (IG-PSO) algorithm based on optimal features returned by Extreme Learning Machine (ELM) as well as it provides an optimal features. IG-PSO algorithm that used in this paper paves a way for handling dataset of medical area. This process provides the better accuracy over classification using optimal features. Thus simulation results have been achieved from IG-PSO algorithm which has the capability of reduction of dimensionality, handling optimization.

In the paper [7], Subbulakshmi et al., proposed a methodology on machine learning called hybrid methodology. This helps to integrate the particle swarm optimization (PSO) algorithm along with extreme learning machine (ELM) classifier. This algorithm used to evaluate ELM's optimized group of parameters, then it minimize the list of hidden layer neurons, and to improve performance of the network generalization. The process deals with UCI Machine Learning Repository in order to handle the classification of medical dataset. Thus results helps to achieve better performance.

In the paper [8], Neol Perez Perez aims to enhance the feature selection methods and the performance of AUC which is based on classifier especially for breast cancer. And the

authors also mentioned the related issues such as making the breast cancer dataset that has been used. To increase the various, then enhancing the *uFilter* method's performance in various domains and made the *uFilter* method is made to extend and allow to use on various problems such as multiclass classification.

NithinKumariet al., [9] introduced the system that deals beyond the analysis on health care. Since the health care fields consists of high amount of data which has need of refinement to obtain the beneficial information. This paper compares the various techniques for mining the data and it helps to find out whether one suffers because of coronary disease or not. The technique called Neuro-fuzzy is the one that have advantages of ANN and fuzzy logic. On comparing these all techniques authors conclude that Neuro-fuzzy approach is best among other two techniques especially for diagnosing the heart disease.

In the paper [10], Arun Kumar et al., proposed the system for detecting and classifying the brain tumor based on human by the help of brain images. Noise is available in medical resonance images which are cause by operator that leads the inaccuracies based classification. This system uses Support Vector Machine, PCA and K-mean. Thus it provides fast and tool for accuracy on tumor. In addition K-mean helps for Image Segmentation. SVM use Brain tumor image techniques for testing and training. Thus the system results in to automation in classification of brain tumor.

3. PROPOSED SYSTEM

This study combines the Fuzzy Extreme Learning Machine (FELM) and PSO based on the classifier. In this process, the optimal solution from the dataset is obtained by utilising the PSO. For the classification of the data, FELM is employed. This system is further divided into two phases:

- Phase 1: Feature Selection using PSO
- Phase 2: Classification of Selected Attributes Using Fuzzy Extreme Learning Machine (FELM) based classifier

3.1.Feature selection using particle swarm optimization

The computation technique which is evolutionary, called Particle Swarm Optimisation is inspired by the social behaviour [11]. Swarm is used for monitoring the particle population by PSO and also in search space it encodes the candidate solution. The random position

initialisation in the space and iteration of each particle's position depending on its neighbours and particles' experience is initialised by PSO. The vectorx_i = $(x_i1, \ldots, x_in)$ represents the position of the particle and n stands for search space dimension. The vector v_i = $(v_i1, \ldots, v_in)$ denotes the velocity and the predefined range $[-vmax, vmax]$ limits the vector of each component. The personal best $pbest_i$ = $(p_i1 \ldots p_in)$ in which the particle i is represented as the best previous position (based on the few fitness function) and $gbest = (g1 \ldots gn)$, the global best is the one in the population as a whole is found by best position and recorded. The position of each particle and velocity of the particle is updated by PSO at each iteration and the following equation defines that:

$$\boldsymbol{x_{i,d}^{t+1} = x_{i,d}^{t} + v_{i,d}^{t+1}} \quad \textbf{(1)}$$

$$\boldsymbol{v_{i,d}^{t+1} = w.v_{i,d}^{t} + C_1.r_{1,i}.\left(P_{i,d} - x_{i,d}^{t}\right) + C_2.r_{2,i}.\left(g_d - x_{i,d}^{t}\right)} \quad \textbf{(2)}$$

The velocity vector or the component's position is represented by 0 < d ≤ n, t-th iteration is denoted by t in the algorithm, the inertia weight's predefined constant is represented by w, constants for predefined acceleration is given by c1 and c2. The random values which are uniformly distributed over [0,1] is represented by r1,i and r2,i. The search spaces of the real value is applicable to the description of PSO. The modified algorithm is in demand as the issues occur along with the feature selection in the discrete search in which the Binary PSO has also been included. The restriction of the values to 0 or 1 by the values of all position vectors $(x_i, pbest_i, and\ gbest_i)$ of the components prevails in binary PSO. The corresponding component with the probability in the position vector of each component is being 1 in the Equation (2) and is employed for updating velocity. For the transformation into a unit range, sigmoid function $s(v_i,d)$ is used. Based on the following equation the position of the each particle is updated by binary PSO.

$$x_{i,d} = \begin{cases} 1, & rand() < s(v_{i,d}) \\ 0, & otherwise \end{cases} \quad (3)$$

Where

$$S(X) = \frac{1}{1 + e^{-x}}$$

According to the fitness function PSO places its focus on searching of best classification performance in the feature selection. Besides the performance of classification, PSO demands the assistance for searching the feature subset which is off small size and classification has high accuracy since in most cases, search space is large. Thus the size-controlled PSO is

proposed in which the target size guides for search of PSO. The updating equation's velocity is produced by the influence of T and is given in the equation (4).

$$v_{i,d}^{t+1} = w.v_{i,d}^{t} + C_1.r_{1,i}.\left(P_{i,d} - x_{i,d}^{t}\right) + C_2.r_{2,i}.\left(g_d - x_{i,d}^{t}\right) + C_3.r_{3,i}.S(T - |p|) \quad (4)$$

where the particles *i* for the selection of number of particles is represented by |p|, third generation constant is given by c3, uniform distribution of random values r3,*i* over [0,1].

PSO:

Input: Data Set D

- Divide Dataset into a training set and a test set

Output: Selected Featured Subset

Algorithm:

Random initialization of velocity and position of each particle

1. Do, while the criterion for stopping is not met
2. Fitness evaluation on training set of each particle
3. For each particle p do
4. Update the pbestand gbestof p
5. End for
6. For each particle p do
7. Update the velocity and position of p
8. End for
9. End while
10. Return the position of gbest (the selected feature subset)

3.2.Fuzzy Extreme Learning Machine (FELM) based classifier

This work uses the FELM. Thus combination of relative based metrics of Fuzzy sets and an Extreme based Machine Learning has been done by using classifier. Clinical datasets has been used in proposed System to carry out the above said processes. Mainly three types of subsystems are employs in this proposed system. These Subsystems are in the framework named FELM. The subsystems include fuzzification subsystem, classification subsystem and preprocessing subsystem. The fuzzification subsystem which is one among the three subsystems will map every feature to each fuzzy set. In addition, classification subsystem includes the process of using the various classification algorithms available in extreme machine based learning in order to perform classification.

Fuzzification subsystem

Transformation of features has been performed by using function namely Trapezoidal membership. This process of transformation is applied on clinical datasets which is specially selected in order to obtain fuzzy set additionally with the membership value. Thus fuzzification over clinical datasets is performed using functions of membership.

Classification subsystem

Classification subsystem involves mainly two approaches such as classifier construction and testing. This research aims on the classification which is carry out using Feed forward based neural network. This Feed Forward Neural Network performs classification with the help of separate hidden layer using ELM. ELM stands for Extreme Learning Machine is used in order to identify the weights between output and hidden layer neurons.

Thus amount of neurons in input, hidden and output layers are symbolically represented with the help of notations such as p, q, and r. The Weight vector is measured between hidden and input layer neurons by using the representing as W^{ih}. Additionally W^{ho} is used to represents the weight vector in between the neurons of hidden layer and output layer. The fuzzified based clinical datasets obtains the value that expected is represented as T. Hidden layered neurons uses the function called Sigmoid activation. For each value of q which means hidden layer neurons, the description of SLFNN's training has been obtained as follows,

Step 1: FELM is allowed to take fuzzified features as input which is clinical dataset.

$$I_i = X_i i = 1,2, \dots, p \quad (5)$$

Where, the notation p used to refer the total amount of fuzzified features in X which means clinical dataset.

Step 2: Randomly initialization of weights between input and hidden layer neurons has been applied which ranges from 0 to 1 and can be represented in W^{ih}, Where the letters i and h specifies the neurons such as input layer and hidden layer. On the other hand i have the values that ranging between 1 and amount of fuzzified features which includes in the clinical based data set. The value of letter h, ranging between 1 and q ie number of neurons especially in hidden layer.

Step 3: By using Equation 6, Computation of hidden layer neurons input has been done and denoted as H^i_j ,

$$H_j^i = \sum_{i=1}^{p} \left(I_i^0 W_{ij}^{ih}\right) \quad j = 1,2, \dots, q \quad (6)$$

I_i^o denotes the output of input layer neurons, whereas W_{ij}^{ih} denoted as weight between the hidden and input layer of neuron.

Step 4: Using 7th equation, H^o means hidden layer's output has been computed;

$$H_j^0 = \frac{1}{1+e^{-H_j^i}} \qquad j = 1,2,\ldots.q \qquad (7)$$

Step 5: Between the hidden layer and output layers of neurons, the weights such as W^{ho} has been determined with the help of ELM method as in equation 8.

$$W = H^{\dagger}T \qquad (8)$$

Where$H^{\dagger}$ is Moore-Penrose′s generalized inverse of H .

Here, T is used to denote the target class and H represents the output valueof hidden layer neuron. Weights which could make connection between hidden layer neurons and the output layer neurons has made and said as W.

T=HW (9)

Step 6: Equation 10 helps to get the value for output layered neuron shortly O_k.

O_k is obtained by using the values in H_j and theW_{jk}^{ho} such as

$$O_k = f\left(\sum_{j=1}^{q}(H_j\, W_{jk}^{ho})\right) \qquad \text{k=1,2,…n} \qquad (10)$$

Let f denotes the activation function that we used; number of hidden layer neurons shortly as q and number of dataset used for training is represents as n.

3.3.PSO+FELM

In this section, we describe the proposed PSO-FELM classification method for the diagnosis of coronary artery disease cardiac auscultation. The aim ofthis system is to optimize a set of weighting factors for thefeature set, such that the highest accuracy of the classifier can be achieved.

1. *Initialization*

- Generate an initial swarm of size S.
- Set the velocity vectors v_i (i=1,2,….S) for each particle in the swarm with a value of zero.
- Train an Fuzzy Extreme Learning Machine (FELM) based classifier and compute the corresponding fitness function f(i) (i.e. the accuracy) of each position $p_i(t)$ for each $p_i(t)$ in the swarm.

- Select the best position from each particle with its initial position as in (11):

$$P_{bi} = P_i, (i = 1,2, \dots, S) \quad (11)$$

2. *Optimization process*

- Determine the best global position p_g from particles in the swarm by the fitness function over all explored trajectories.
- Update the velocity and position of each particle.
- Train an Fuzzy Extreme Learning Machine (FELM) based classifier and compute the corresponding fitness function f(i) for each candidate particle p_i (i=1,2,…,S)
- Update the best position p_{bi} of each particle, if the corresponding position has a smaller fitness function value.
- Classify the given data with the trainedFuzzy Extreme Learning Machine (FELM) based classifier.

3. *Stopping Criteria*

- If not at maximum iteration, repeat the optimization process. Otherwise, continue step 2.

4. EXPERIMENTAL DESIGN

Z-Alizadeh Sani dataset is considered from the UCI machine learning repository as benchmark problems to evaluate the performance of proposed method. Table 1 shows the dataset characteristics of the Z-Alizadeh Sani data set. In the experimental design, Accuracy, Sensitivity and Specificity parameters are used. It is evaluated with our proposed method and compared with existing methods.

Table 1. Data Set Characteristics

Name of the Dataset	Z-Alizadeh Sani dataset
Number of Instances	303
Number of Features	54
Number of Groups	4 (Demographic, Symptom and examination ECG, Echo feature)
Classification	Each patient could be in two possible categories Normal or Not based his/her diameter.

Performance Metrics

The following are the performance metrics used up for the evaluation of the model.

- **A confusion matrix** is a classification table used to describe the performance of a classifier.
- **True positive (TP):** Theproportion of actual positives that are correctly identified.
- **False positive (FP):** The proportion of actual positives that areincorrectly identified.
- **True negative (TN):** The proportion ofnegatives that are correctly identified.
- **False negative (FN):**The proportion of negatives that are incorrectly identified.
- **Positive Predictive Value (PPV)** is the probability that the disease is presentgiven a positive test result.
- **Positive Predictive Value (NPV)** is the probability that the diseaseis absent given a negative test result.
- **Sensitivity** defines how well the classifiercan identify a diseased heart correctly, also called truepositive rate.

  $Sensitivity = \frac{TP}{TP+FN}$
- **Specificity** means the tests ability to excludehealthy heart from diseased heart correctly.

  $Specificity = \frac{TN}{FP+TN}$
- The **accuracy** results show the proportion of correctly classified instances and incorrectly classified instances.

Accuracy= (True Positive + True Negative)/(True Positive + False Negative+

False Positive + True Negative)

Results and Discussion

The proposed approach FELM is compared with three algorithms. Naïve Bayes (NB), Neural Network (NN)and Hybrid Neural Network - Genetic Algorithm (HNN-GA) algorithms are taken into consideration.

Table 3. Performance Analysis

Algorithm	***No Feature Reduction***			***Feature Reduction***		
	Accuracy	***Sensitivity***	***Specificity***	***Accuracy***	***Sensitivity***	***Specificity***
Naive Bayes	79.61	0.873	0.675	81.42	0.865	0.762
Neural Network	81.55	0.845	0.750	82.52	0.798	0.947
HNN-GA	83.49	0.849	0.800	84.46	0.871	0.788
FELM	**85.44**	0.835	**0.917**	**87.14**	**0.878**	**0.935**

Table 3 shows the performance analysis of four algorithms. All four algorithms applied based on 2 scenarios. In the first scenario, no features are reduced. In the second scenario Feature Reduction (FR) done using PSO.

Figure 1. Comparison of Accuracy for Different Algorithms

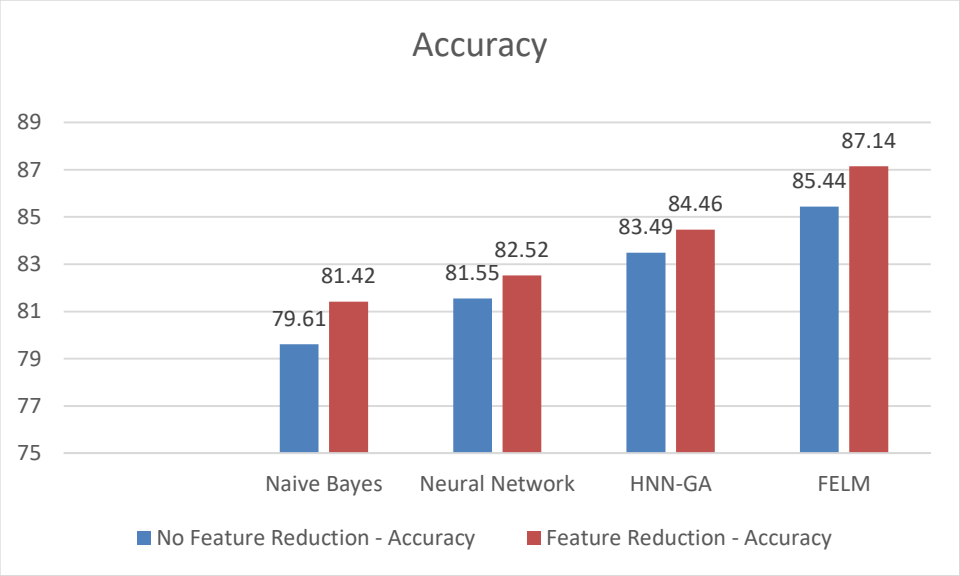
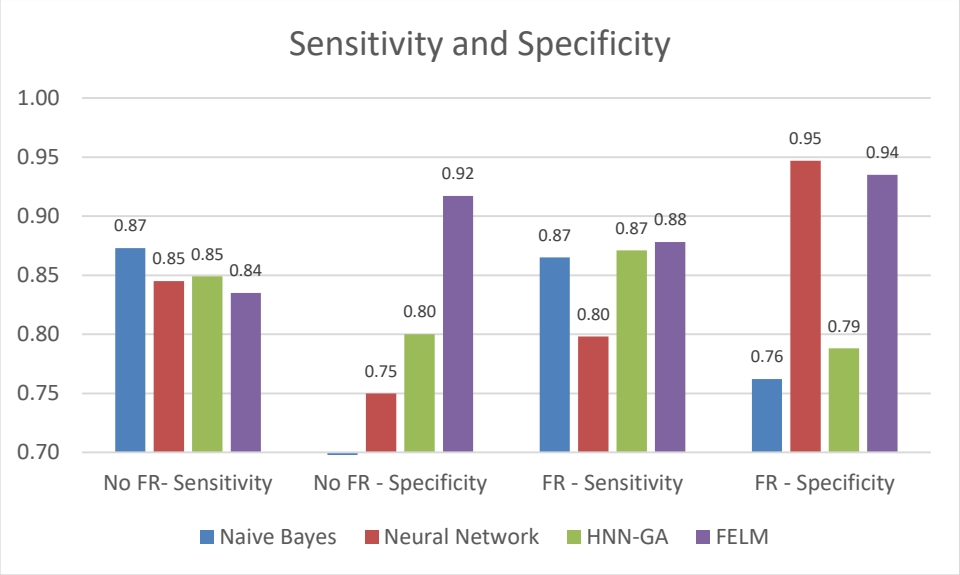


Figure 2. Comparison of Sensitivity and Specificity for Different Algorithms



Comparison of accuracy, sensitivity and specificity for different algorithms are presented in Figure 2 and Figure 3. Compared to the previous methods, the proposed method shows an improvement of classificationin the case of correctly classifies patients. In all cases, the proposed system outperforms existing methods, except sensitivity in the no feature reduction scenario. However, the systemcannot perfectly classify all the instances. It could be concluded that the methodproposed by this study obtains promising results for Z-Alizadeh Sani dataset classification system.

5. CONCLUSION AND FUTURE WORK

Hybridization of Particle Swarm Optimization (PSO) and highly accurate hybrid Fuzzy Extreme Learning Machine (FELM) is proposed for the diagnosis of coronary artery disease in this paper. In the proposed algorithm, the performance of the hybrid Fuzzy Extreme Learning Machine is improved by applying PSO to FELM. The Z-Alizadeh Sani dataset obtained from the UCI machine learning repository is used for evaluating the performance of the proposed method. The performance of the proposed FELM algorithm is estimated based on classification accuracy, sensitivity and specificity. The results are compared with three existing algorithms namely Naïve Bayes, Neural Networks and Hybrid Neural Network - Genetic Algorithm. The proposed algorithm FELM with PSO provides better classification results. In other performance metrics like sensitivity and specificity, the FELM outperforms other three algorithms with better results. In future, EFS might combine the results of three feature selection methods such as Bat Algorithm (BA), and Firefly Algorithm (FA) which give a better approximation to the optimal subset or ranking of features. Parallel Deep Learning Algorithm (PDLA) based prediction model can be proposed to detect coronary artery disease based on clinical data without the need for invasive diagnostic methods.

REFERENCES:

[1] ZeinabArabasadi, RoohallahAlizadehsani, Mohamad Roshanzamir, Hossein Moosaei, Ali AsgharYarifard, "Computer aided decision making for heart disease detection using hybrid neural network-genetic algorithm", Computer Methods and Programs in Biomedicine, January 12, 2017.

[2] A.Kale, S.Sonavane, "Hybrid feature subset selection approach for fuzzy-extreme learning machine", Data-Enabled discovery and applications, September 12, 2017.

[3] S.Kavipriya, T.Deepa, "Comprehensive feature selection based support vector machine classifier (CFS-SVM) for clinical dataset", Journal of theoretical and applied information technology, Vol.96, May 15, 2018.

[4] Chunhong Lu, Zhaomin Zhu XiaofengGu, "An intelligent system for lung cancer diagnosis using a new genetic algorithm based feature selection method", July 4, 2014.

[5] Xiao Liu, Xiaoli Wang, Qiang Su, Mo Zhang, Yanhong Zhu, Qiugen Wang, Qian Wang, "A Hybrid classification system for heart disease diagonosis based on the RFRS method", Hindawi Computational and Mathematical Methods in Medicine, 2017.

[6] A.P.Kale, S.P.Sonavane, "Improved genetic particle swarm optimization and feature subset selection for extreme learning machine", International journal of computer sciences and engineering (IJCSE), Vol.6, February 1, 2018.

[7] C.V.Subbulakshmi, S.N.Deepa, "Medical dataset classification: A machine learning paradigm integration particle swarm optimization with extreme learning machine classifier", Scientific world journal, 30 September 2015.

[8] Neol Perez Perez, "Improving variable selection and mammography-based machine learning classifiers for breast cancer CADx", March 2015.

[9] NithinKumari, Sunita, Smita, "Comparisionof ANNs, Fuzzy Logic and Neuro-Fuzzy integrated approach for diagnosis of coronary heart disease: a survey", International journal of computer science and mobile computing, Vol.2, June 6, 2013.

[10] Arun Kumar, Richika, "A novel approach for brain tumor detection using support vector machine, K-Means and PCA Algorithm", International journal of computer science and mobile computing, Vol.4, August 2015.

[11] Butler-Yeoman, T., Xue, B. and Zhang, M. "Particle Swarm Optimisation for Feature Selection: A Size-Controlled Approach", In Proc. Thirteenth Australasian Data Mining Conference (AusDM 2015) Sydney, Australia. CRPIT, 168. Ong, K.L., Zhao, Y., Stone, M.G. and Islam, M.Z. Eds., ACS. 151-159.