International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Optimal Fuzzy Min-Max Neural Network (FMMNN) for Medical Data Classification Using Modified Group Search Optimizer Algorithm

D. Mahammad Rafi[1]* **Chettiar Ramachandra Bharathi[2]**

[1]*Vivekananda Institute of Engineering & Technology, JNTU University, Hyderabad, India*
[2]*Vel Tech University, Chennai, Tamilnadu, India*
*Corresponding author's Email: dmahammadrafi0780@gmail.com

Abstract: The main intension of the research is to classify the medical data with high accuracy value. In order to achieve promising results, we have designed to utilize orthogonal local preserving projection and optimal classifier. Initially, pre-processing will be applied to extract useful data and to convert suitable sample from raw medical datasets. In the proposed method, input dataset will be high dimensional or high features; so the high number of feature is a great obstruction for prediction. Therefore, feature dimension reduction method is used in our proposed technique. Here, orthogonal local preserving projection will be used to reduce the feature dimension. Once the feature reduction is formed, the prediction will be done by optimal classifier. Here modified group search optimizer algorithm combined with Fuzzy Min-Max neural network. The implementation will be done in MATLAB. The performance of the proposed technique is evaluated using accuracy, sensitivity and specificity.

Keywords: orthogonal local preserving projection; group search optimizer algorithm; Fuzzy Min-Max neural network; optimal classifier.

1. Introduction

Data mining is an essential step in the process of knowledge discovery in databases in which intelligent methods are applied in order to extract patterns [1]. It is the process of analyzing data from different perspectives and summarizing it into useful information. The main goal of data mining is to discover new patterns for the users and to interpret the data patterns to provide meaningful and useful information for the users. It applied to find useful patterns to help in the important tasks of medical diagnosis and treatment [2]**.** The algorithms, when appropriately used, are capable of improving the quality of prediction, diagnosis and disease classification [3]. With data technique such knowledge can extracted and accessed transforming the data base tasks form storing and retrieval to learning and extracting knowledge [4]. Classification of medical data for accurate diagnosis is a growing field of application in data mining [5]. It is a process of finding the class model according to their attributes. There are two different category of learning process namely supervised learning and unsupervised learning. Decision tree is a common classification algorithm and widely used in many applications. Decision tree algorithm includes Classification and Regression Tree [6].

The classify panel enables the user to apply classification and regression algorithms to the resulting dataset, to estimate the accuracy of the resulting predictive model, and to visualize erroneous predictions, or the model itself [7]. An ensemble model is defined as a set of individually trained classifier whose predictions are combined when classifying a new data. Ensemble combines the output of several classifiers produced by weak learner into a single composite classification. It can be used to reduce the error of any weak learning algorithm [8]. Classification model could make predication of categorical label include discrete or unordered variables [9]. A data set is imbalanced if the classification categories are not equally represented [4]. Classification divides data samples into target classes. The classification technique predicts the target class for each data points. Binary and multilevel are the two methods of classification. In binary classification the performance and quality of classification algorithms is usually evaluated using predictive accuracy [10]. Not with standing,

this is not suitable when the information is unequivocally imbalanced as disparities in the quantity of items between the classes may prompt serious weakening of the grouping exactness [11]. Data mining and knowledge discovery in databases have been attracting significant amount of research in fields like industry medical, commerce, science which is attracting the media attention of late [12]. After, classifier was built, this model would be evaluated according to the answers accuracy that model will give through testing. Whereas, another data, unlabeled examples that are known test data, provide the model [13]. Many successful applications that which are based on association rule mining algorithms are used to produce very large numbers of rules. In most of the decision support systems the accuracy of classifier is measured on the basis of all attributes [14]. Various algorithms and techniques are used in the medical data mining like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method [15].

2. Related Work

In this section, we have discussed some recent papers about medical data classification.

M. A. Jabbar *et al* [16] have proposed a method that to discover association rules in medical data to predict heart disease for Andhra Pradesh. That approach was expected to help physicians to make accurate decision. Mortality data from the registrar general of India shows that the coronary heart disease (CHD) were a major cause of death in India. They determine the precise cause of death in rural areas of Andhra Pradesh have revealed that CHD cause about 30% death in rural areas.

Indu Saini *et al* [17] have proposed algorithm was evaluated on two manually annotated standard databases such as CSE and MIT-BIH Arrhythmia database. That work tells about digital band-pass filter was used to reduce false detection caused by interference present in ECG signal and further gradient of the signal was used as a feature for QRS-detection. They also found an addition the accuracy of KNN based classifier was largely dependent on the value of K and type of distance metric. The detection rates of 99.89% and 99.81% were achieved for CSE and MIT-BIH databases respectively.

Sina Khanmohammadi and Mandana Rezaeiahari [18] have proposed to choose a machine learning classification that has been used for developing clinical decision support system. They presented ten sample medical datasets. They suggest model SVM as the most desirable classification algorithm for developing CDSS. The research was not to identify a classification algorithm that has been performing best in all medical datasets. The reliability of the model has been improved using more sample datasets.

R. Chitra and V. Seenivasagam [19] have proposed a growing research on heart disease predicting system, that has been become important to categories the research outcomes and provides readers with an overview of the existing heart disease prediction techniques in each category. Neural Networks were one of many data mining analytical tools that have been utilized to make predictions for medical data.

R. Bhuvaneswari and K. Kalaiselvi [20] have proposed the use of Naive Baye's classifier in medical applications. Quality service implies diagnosing patients correctly and administering treatments that are effective. Decision Support in Heart Disease Prediction System was also developed using Naive Bayesian Classification technique. Naïve Bayes classifications have been used as a best decision support system.

P K. Anooj [21] have proposed a weighted fuzzy rule-based clinical decision support system (CDSS) was presented for the diagnosis of heart disease, automatically obtaining knowledge from the patient's clinical data. They consist two phases namely:

(1) Automated approach for the generation of weighted fuzzy rules and

(2) Developing a fuzzy rule-based decision support system.

M. Akhil jabbar *et al* [22] have proposed a combines approach of KNN and genetic algorithm have to improve the classification accuracy of heart disease data set. They used genetic search as a goodness measure to prune redundant and irrelevant attributes, and to rank the attributes which have been contribute more towards classification. Least ranked attributes were removed, and classification algorithm is built based on evaluated attributes.

A. Sudha, *et al* [23] have proposed a principle component analysis algorithm was used for reducing the dimensions and that have been determines the attributes involving more towards the prediction of stroke disease and predicts whether the patient suffering from stroke disease or not. That algorithm deals about stroke was a life threatening disease that has been ranked third leading cause of death in states and in developing countries. The stroke was a leading cause of serious, long term disability in US.

Martti Juhola *et al* [24] have proposed complicated variable distribution of the data although there were only two classes. In addition to a straightforward data cleaning method, they used an efficient way called neighborhood cleaning that solved the problem and improved their classification accuracies 5–10%, at their best, up to 95% of all test cases. That shows important that was first very carefully to study distributions of data sets have been classified and use different cleaning techniques in order to obtain best classification results.

B. Dennis and S. Muthukrishnan [25] have proposed an efficient medical data classification system based on Adaptive Genetic Fuzzy System (AGFS). In their research 1) Generating rules from data as well as for the optimized rules selection, adapting of genetic algorithm is done and to explain the exploration problem in genetic algorithm, introduction of new operator, called systematic addition is done, 2) Proposing a simple technique for scheming of membership function and Discretization, and 3) Designing a fitness function by allowing the frequency of occurrence of the rules in the training data.

Adeniyi et al. [26] has presented a study of automatic web usage data mining and recommendation system based on current user behavior through his/her click stream data on the newly developed Really Simple Syndication (RSS) reader website, in order to provide relevant information to the individual without explicitly asking for it. The K-Nearest-Neighbor (KNN) classification method has been trained to be used on-line and in Real-Time to identify clients/visitors click stream data, matching it to a particular user group and recommend a tailored browsing option that meet the need of the specific user at a particular time. Their result shows that the K-Nearest Neighbor classifier was transparent, consistent, straightforward, simple to understand, high tendency to possess desirable qualities and easy to implement than most other machine learning techniques specifically when there is little or no prior knowledge about data distribution.

3. Problem Definition

In this section discuses the problem definition of my research work

- The data contains redundant and irrelevant attributes classification produce less accurate result [22].
- The straightforward cleaning of medical data set impaired its classification result considerably with some machine learning methods, but not all of them unexpectedly and against intuition compare to the original situation without any data cleaning [24].
- The availability of huge amount of medial data leads to the need for powerful data analysis tools to extract useful knowledge.
- Diagnosis of most of the disease is expensive as many tests are required to predict disease. The cost diagnosis by avoiding many tests by selection of those attribute.
- Number of work carried out for prediction various disease comparing the performance of predictive data mining.
- The physicist diagnose represented by human expertise it can be incurrence to fail. In contrast the data mining can be recruit the extract knowledge from huge of clinical data through data mining and produce a predictive model use the classification task to achieve the diagnostic [13].
- The analysis of data mining process required for medical data mining especially to discover locally frequency disease such as heart alignment lung cancer, breast cancer and so on.
- Classification accuracy is improved by removing most irrelevant features the dataset. Ensemble model is used for improving classification accuracy by combining the prediction of multiple classifiers.
- The healthcare industry gathers enormous of heart disease data which is unfortunately to discover hidden information for effective decision
- The significant progress made in the diagnosis and treatment of heart disease, further investigation is still needed [16].
- Medical professionals need a reliable prediction methodology to diagnose Diabetes.

4. Proposed Methodology

In this research we have intend to propose an efficient method to classify the medical data. In medical data classification, in order to achieve better result orthogonal local preserving projection and optimal classifier will be used in the proposed technique. At first the input data set is selected from the medical database. Then preprocessing will be applied in the input medical data set. In preprocessing stage we have to extract the useful data from the raw medical dataset. After preprocessing the input dataset will be high dimensional or high features; so the high number of feature is a great obstruction for prediction. Therefore, feature dimension reduction method will

be applied to reduce the features space without losing the accuracy of prediction. Here, orthogonal local preserving projection (OLPP) will be used to reduce the feature dimension. Once the feature reduction is formed, the prediction will be done based on the optimal classifier. In the optimal classifier, group search optimizer algorithm will be used with Fuzzy Min-Max neural network. The detailed process of our proposed method is shown in Fig.1.

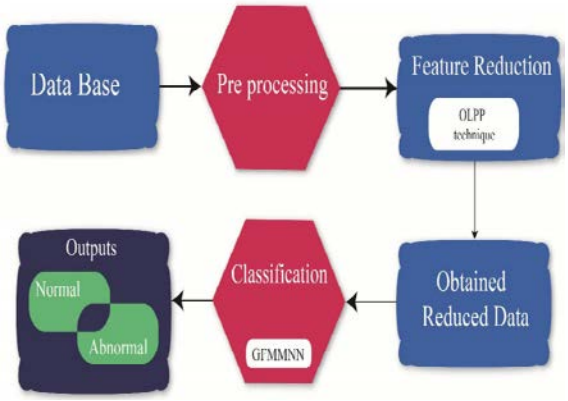

Figure.1 Block diagram of proposed method

The overall process of the proposed framework is divided into three stages,

- Stage1: Preprocessing
- Stage2: Feature reduction using OLPP
- Stage3: Classification using GFMMNN

Stage1: Preprocessing

In preprocessing stage the input dataset is given as the input. Here the input medical data has raw data. This raw data is highly susceptible to noise, missing values and inconsistency. The quality of raw data affects the results of the implemented method. In order to improve the quality of the medical data and consequently, of the results raw data is pre-processed so as to improve the efficiency and ease of mining process. Data pre-processing is one of the most critical steps in a data mining process which deals with the preparation and transformation of the initial dataset. In the paper, pre-processing is applied to the dataset for getting the numerical data from the non-numerical data. In the stage, the non-numerical data are removed and obtained the numerical dataset for proceeding further. The preprocessed output is fed to the further process.

Stage2: Feature Reduction using OLPP

OLPP algorithm differs from Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). The aim of both algorithms is dimensionality reduction, since the first step of the algorithm is PCA which helps in dimensionality reduction. An adjacency graph is built by OLPP and the class relationship between the sample points is best reflected by it. It is not easy to reconstruct the data since Locality Preserving Projections (LPP) is non-orthogonal normally. By means of Orthogonal Locality Preserving Projection method, this problem is overcome which produces orthogonal basis functions and can have more locality preserving power than LPP. The Orthogonal extension of LPP is called as the Orthogonal Locality Preserving Projections (OLPP). The overall steps are shown in the flowchart:

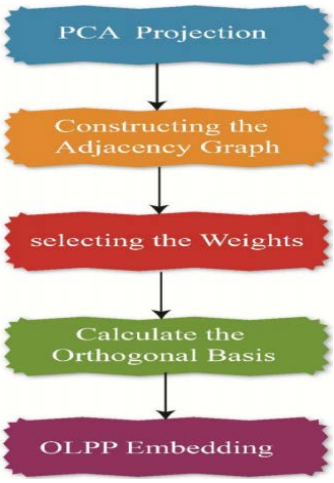

Figure.2 The flowchart for OLPP method

The steps occupied in OLPP:

- **PCA Projection:** Principal Components Analysis is a method that reduces data dimensionality by performing a covariance analysis between factors. The PCA projection involves the following steps:
  (i) Obtain a set of features from the input database
  (ii) Calculate the mean value
  (iii) Compute covariance matrix and then calculate eigen vector and eigen value of covariance matrix
  (iv) The eigen value and eigenvectors are ordered and paired
- **Constructing the Adjacency Graph:** Let $D=[d_1,d_2,...,d_m]$ be a set of input data. Consider G denotes a graph with n nodes. The i^{th} node corresponds to the input data d_i. An edge is put between nodes i and j, if d_i and d_j are "close", i.e. d_i is among p nearest neighbors of d_j or d_j is among p nearest neighbors of d_i. If the

class information is available in any two nodes we simply put an edge between that two nodes belonging to the same class.

- **Choosing the Weights:** if the node i and j are connected, the weight W_{ij} is calculated using the following equation,

$$W_{eij} = e^{-\frac{\|d_i - d_j\|}{t}} \quad (1)$$

Where

$t \rightarrow$ Constant

If the node i and j are not connected means we put $W_{eij} = 0$. The weight matrix W_e of graph G models having the local structure of various input data.

- **Computing the Orthogonal Basis Functions:** After finding the weight matrix W_e we calculate the diagonal matrix M. A diagonal matrix M is defined as, whose entries are column (or row) sums of W_e.

$$M_{ii} = \sum_j W_{e\,ji} \quad (2)$$

After that we calculate the Laplacian matrix L using diagonal matrix M and weight matrix W_e.

$$L = M - W_e \quad (3)$$

Let $O_{r1}, O_{r2}, ..., O_{rm}$ be orthogonal basis vectors, we define

$$A_{m-1} = [O_{r1}, O_{r2}, ..., O_{r\,m-1}], B_{m-1}^T = A_{m-1}^T Z^{-1} A_{m-1} \quad (4)$$

Where;

$$Z^{-1} = DMD^T$$

The orthogonal basis vectors $[O_{r1}, O_{r2}, \ldots, O_{rm}]$ can be computed as follows

(i) Compute O_1 as the eigenvector of $Z^{-1}DLD^T$ associated with the smallest eigenvalue.

(ii) Compute O_m as the eigenvector of associated with the smallest eigenvalue of J_m

$$J_m = \{I - Z^{-1} A_{m-1} B_{m-1}^T\} Z^{-1} \{DLD^T\} \quad (5)$$

- **OLPP Embedding** :

Let $T_{OLPP} = [O_1 O_2 O_3 .... O_l]$ embedding is follow,

$$Y \rightarrow DT^T \quad (6)$$

$$T = T_{PCA}\, T_{OLPP} \quad (7)$$

Where;

$T \rightarrow$ Transformation matrix

$Y \rightarrow$ One dimensional representation of D

This transformation matrix reduces the dimensionality of the feature vectors of the input data. This dimensionality reduced features, given to the classification process.

Stage3: Classification using GFMMNN

Fuzzy Min Max Neural Network

FMNN learning algorithm comprises of three actions: 1) expansion, 2) overlap test, and 3) contraction. Its principle is to locate a suitable hyperbox for each input pattern. If the suitable hyperbox exists, its size cannot surpass the minimum and maximum limits. After expansion, all hyperboxes that have a place with distinctive classes must be checked by overlap test to figure out whether any overlap exists. So a dimension by dimension comparison between hyperboxes of different class is performed. FMNN intends four test cases, at least one of the four cases is satisfied, and then overlap exists between the two hyperboxes. Otherwise, a new hyperbox needs to be added to the network. If no overlaps occur, the hyperboxes are isolated and no contraction is required. Otherwise, a contraction process is needed to eliminate the confusion in overlapped areas.

The training set consist of set of ordered pairs {X, I}, where X= {X_1, X_2,…X_n} is the input data and Iϵ(1,2,…m) is the index of one of the class. The learning process begins by selecting an ordered pair and finding a hyper box for the same class that can expand (if necessary) to include the input. If a hyper box cannot be found that meets the expansion criteria, a new hyper box is formed and added to the neural network. The membership function is defined with respect to the minimum and maximum points of a hyper box. It describes the degree to which a pattern fits in the hyper box. The hyper boxes have a range from 0 to 1 along each dimension. A pattern which is contained in the hyper box has a unity membership function. Mathematically, the definition of each hyper box fuzzy set Hj is defined by,

$$H_j = \{X, V_{\min_j}, W_{\max_j}, F(X, V_{\min_j}, W_{\max_j})\} \quad (8)$$

Where,

X-Input data,

V_{minj} (V_{min1}, V_{min2},….V_{minN}) is the minimum points of H_j

W_{maxj} (W_{max1}, W_{max2},….W_{maxN}) is the maximum points of H_j

F (X, V_{mini}, W_{maxi}) is the membership function

The membership function for the j^{th} hyperbox (Hj) is given below,

$$H_j = \frac{1}{2n}\sum_{i=1}^{n}[\max\ (0,1-\max\ (0,\gamma\min\ (1,X_i - w_{ij}))) + \max\ (0,1-\max\ (0,\gamma\min\ (1,v_{ji} - X_i)))] \quad (9)$$

Where,

γ is a sensitivity parameter that regulates how fast the membership value decreases as the distance between X and H_j increases.

The architecture of fuzzy min max neural network consists of three layers of node. First layer represent the input layer that contains input data. Last layer represent the output layer that contain the number of classes. The middle or hidden layer is called hyper box layer. The overall structure of fuzzy min max neural network is shown in Fig.3.

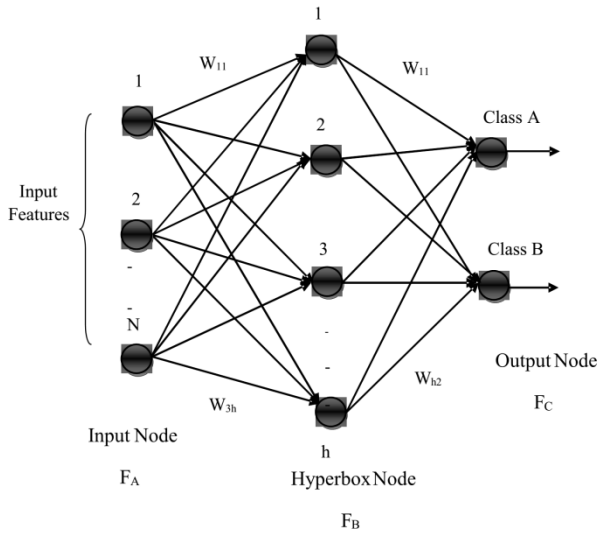

Figure.3 The structure of fuzzy min max neural network

MGSO Algorithm

Group search optimization algorithm (GSO) is proposed to optimize the weight in the neural network. The Group search optimization algorithm is developed with the motivation from the searching activity of animals. The searching activity of animals is mainly done with the intention of discovering resources that include food and shelter. Here the traditional GSO algorithm is improved with the help of velocity updation instead of ranger performance to randomly select the resource.

In this MGSO algorithm, the population is termed as a group and the individuals residing within the group are known as members. The members within a group are of three kinds, namely, the producers, the scroungers and the rangers. The activity of the producers as well as the scroungers relies on the PS model. The rangers move in an arbitrary manner.

Producers: These members go in search of resources.

Scroungers: These members link the resources, which the producer discovers.

Rangers: These are the members that make movements in an arbitrary manner and perform searching in an organized way, so that efficient finding of resources could be achieved.

Step by step procedure of Modified GSO:-

Step 1: Initialize the search solution as well as the head angle

- The head angle can be stated as,

$$\Psi_i^t = \left(\Psi_{i1}^t .....\Psi_{i(n-1)}^t\right) \quad (10)$$

The members direction of search relies on the head angle

$$L_i^t(\Psi_i^t) = \left(l_{i1}^t .....l_{i(n)}^t\right) \quad (11)$$

- Polar and Cartesian coordinate transformation is employed to assess the direction of search

$$L_{i1}^t = \prod_{p=1}^{n-1}\cos(\Psi_{ip}^t) \quad (12)$$

$$L_{ij}^t = \sin(\Psi_{i(j-1)}^t \prod_{p=j}^{n-1}\cos(\Psi_{ip}^t);$$

Where (j=2….n-1) (13)

$$L_{in}^t = \sin(\Psi_{i(n-1)}^t) \quad (14)$$

Step 2: Fitness function is calculated

$$fitness = \min(MSE) \quad (15)$$

Step 3: Find the producer (Z_p) of the group

The member with the best fitness is called as the producer

- Producer performance

During the execution of the MGSO algorithm, the activity of the producer Z_p at't' iteration can be elucidated as follows,

(i) Scanning operation at zero degree

$$Z_z = Z_p^t + \varepsilon_1 d_{\max} L_p^t\left(\Psi^t\right) \quad (16)$$

Where, d_{max} denotes the maximum search distance.

(ii) Scanning operation at the right hand side

$$Z_r = Z_p^t + \varepsilon_1 d_{\max} L_p^t\left(\Psi^t + \varepsilon_2 \Phi_{\max}/2\right) \quad (17)$$

(iii) Scanning operation at the left hand side

$$Z_l = Z_p^t + \varepsilon_1 d_{\max} L_p^t\left(\Psi^t - \varepsilon_2 \Phi_{\max}/2\right) \quad (18)$$

Where, ε_1 points to a normally distributed random number with zero mean and unity standard

deviation. And ε_2 stands for a uniformly distributed random sequence that takes value between zero and one. The computation of maximum search distance d_{max}

Maximum search angle Φ_{max}

$$\Phi_{\max} = \frac{\pi}{c^2} \tag{19}$$

The constant c can be stated as:

$$C = round\ (\sqrt{n+1}) \tag{20}$$

Where, n denotes the dimension of the search space.

$$\therefore \Phi_{\max} = \frac{\pi}{n+1} \tag{21}$$

The present best location would take a new best location, if its resource is found as not better than that in the new location. Else, the producer will maintain its location and turn its head according to the head angle direction that is arbitrarily generated.

- Scrounger performance

In all iterations, several members that exclude the producer are also chosen and they are called as scroungers. The scrounging action of GSO usually involves the area copying activity.

$$Z^{t+1} = Z_i^{\ t} + \varepsilon_3 o(Z_p^t - Z_i^t) \tag{22}$$

Where, o specifies the Hadamard product that computes the product of the two vectors in an entry-wise manner and ε_3 denotes a uniform random sequence lying in the interval of (0, 1).

- Ranger performance

The rangers are the remaining members of the group, which have been displaced from their present location. The rangers can also find the resources effectively through random walks or an organized searching procedure. Random walks are preferred in cases, where the resources are found to be distributed. In our Modified GSO here instead of ranger performance we go for the velocity Updation.

$$V_i^{new} = V_i^t + c_1.r_1.(IP_{best}^{\ t} - S_i^t) + c_2.r_2.(IG_{best}^{\ t} - S_i^t) \tag{23}$$

C1 and C2 constants containing value of 2

r1 and r2 random numbers equally produced in the range [0, 1]

S_i initial position of i^{th} particle

Once the entire process gets completed, the fitness of the updated solution is evaluated. The best solution will be gained, if the process is repeated for‘t’ number of iterations. Based on these the weights are optimized. Finally we classify the medical data with high prediction value.

5. Results and Discussion

The proposed system is implemented using MATLAB 2014 and the experimentation is performed with i5 processor of 3GB RAM.

Dataset description

The proposed method is experimented with the four dataset namely Kidney chronic, Cleveland, Hungarian and Switzerland. These datasets are taken from the UCI machine learning repository.

(i) Mammographic Mass Data Set

The mammographic mass dataset used here has been collected at the Institute of Radiology of the University Erlangen-Nuremberg between 2003 and 2006. The data set is available by http access of the University of California at Irvine (UCI) machine learning repository. Digital Database for Screening Mammography (DDSM) has been used to evaluate the proposed system. The database contains approximately 2,620 cases.

(ii) Pima Indians Diabetes Data Set

The source of Pima Indian diabetes data set is the UCI machine learning repository. The data source uses 768 samples with two class problems to test whether the patient would test positive or negative for diabetes. All the patients in this database are Pima Indian women at least 21 years old and living near Phoenix Arizona, USA.

(iii) Cleveland data

This data base contains 76 characteristics, however all distributed tests refer to utilizing a subset of 14 of them. Specially, ML researchers use only the Cleveland database till today. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1, 2, 3, 4) from absence (value 0). The names and social security numbers of the patients were recently removed from the database, replaced with dummy values.

(iv) Hungarian data

Owing to a vast percentage of missing values three of the characteristics have been rejected however the format of the data is precisely the similar as that of the Cleveland data. Thirty-four examples of the database were rejected on account of missing values and 261 examples were present. Class distributions are 62.5% heart disease not present and 37.5% heart disease present.

(v) Switzerland data

More number of missing values is in Switzerland data. It encloses 123 data instances and

14 features. Class distributions are 6.5% heart disease not present and 93.5% heart disease present.

Evaluation metrics

In order to assess the efficiency of the proposed system an evaluation metric is employed. It contains a set of measures that pursue a general underlying evaluation methodology some of the metrics that we have select for our evaluation purpose are True Positive, True Negative, False Positive and False Negative, Sensitivity, Specificity and Accuracy.

$$Sensitivity = \frac{TP}{TP + FN} \quad (29)$$

$$Specificity = \frac{TN}{FP + TN} \quad (30)$$

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (31)$$

Performance Analysis

The results of proposed work help to analyze the efficiency of the prediction process. The subsequent table.1 tabulates the results. Here the results of four datasets are given in table.1.

Table 1. Performance of the proposed method using various dataset

Dataset	Accuracy	Sensitivity	Specificity
Mammographic Mass data	94.216	0.974886	0.915816
Pima Indians Diabetes data	95.30	0.97032	0.933673
Cleveland data	85.148	0.851485	0.851485
Hungarian data	83.3333	0.952128	0.698113
Switzerland data	84.013	0.946809	0.707547

From table.1, the evaluation metrics are analyzed for the five numbers of datasets, by which we can observe the efficiency of proposed medical data classification system. The accuracy values of five dataset are 94.2%, 95.30%, 85.14%, 83.33% and 84.01%. The sensitivity values for the five datasets are 0.974, 0.9703, 0.851, 0.952 and 0.946. The specificity values for the five datasets are 0.915, 0.933, 0.851, 0.69 and 0.707.

Comparative Analysis

The literature review works are compared in this section with the proposed work to show that our proposed work is better than the state-of-art works. We can establish that our proposed work helps to attain very good accuracy for the medical data classification. In our proposed method we use fuzzy min max neural network with modified group search algorithm for classification. And also we can establish this prediction accuracy outcome by comparing other classifiers. Here the proposed classifier is compared with the traditional neural network. Below specified Fig.4 explains the comparison outcomes of the Sensitivity measures.

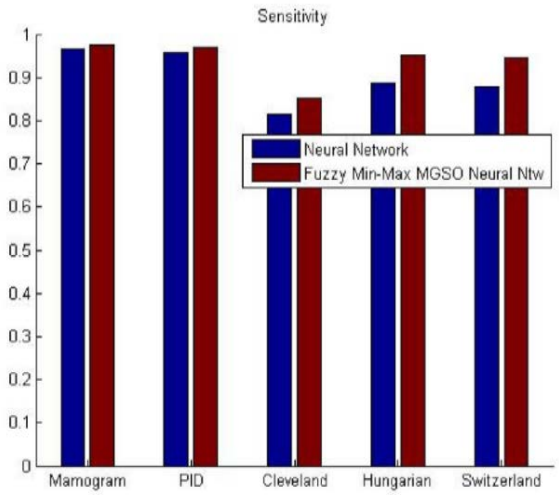

Figure.4 the comparison outcomes of the Sensitivity measures

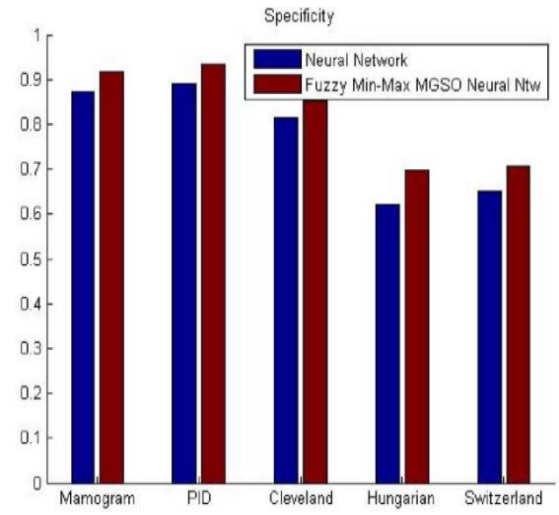

Figure.5 The comparison outcomes of the Specificity measures

The sensitivity values for the existing methods are 0.96, 0.958, 0.81, 0.88, and 0.87 which is low when compared with our optimal classifier, the

sensitivity values of our optimal classifier are 0.974, 0.970, 0.851, 0.952, and 0.94. Below specified Fig.5 explains the comparison outcomes of the Specificity measures.

The specificity values for the existing methods are 0.872, 0.89, 0.81, 0.622 and 0.650, which is low when compared with our optimal classifier, the specificity values of our optimal classifier are 0.915, 0.93, 0.85, 0.69 and 0.707. Below specified Fig.6 explains the comparison outcomes of the Accuracy.

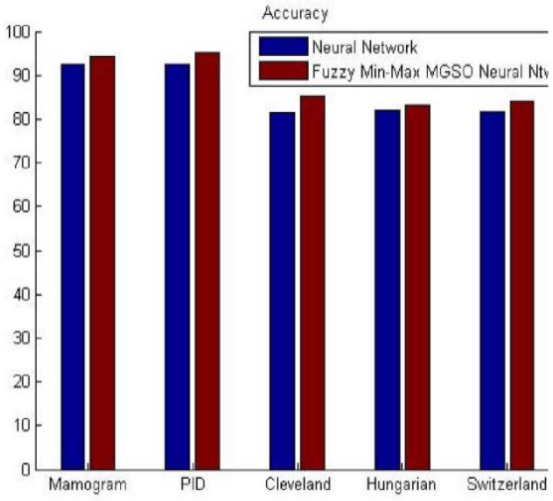

Figure.6 The comparison outcomes of the Accuracy measures

The accuracy values for the existing methods are 92.65%, 92.65%, 81.51%, 81.97% and 81.6%, which is low when compared with our optimal classifier, the accuracy values of our optimal classifier are 94.21%, 95.30%, 85.14%, 83.33% and 84%. When compared to the existing method the proposed method achieves better result. The drawbacks of other existing technique contain both redundant and irrelevant attributes; it leads to less classification accuracy. In order to overcome this drawback, initially the proposed technique removes the irrelevant data and also reduces the dimension with the help of OLPP technique. Then classification is done by optimal classifier. So that, the implemented technique reaches the maximum accuracy value compared to the other existing techniques.

6. Conclusion

In this paper we have proposed a method orthogonal local preserving projection and optimal classifier. Initially, the pre-processing will be applied to extract useful data and to convert suitable sample from raw medical datasets. Feature reduction is done with the help of OLPP and the classification done by optimal classifier which combined MGSO and Fuzzy Min-Max Neural Network (GFMMNN).The implementation of the proposed method was done in MATLAB. For experimentation, the dataset given in the UCI machine learning repository such as, Mammographic Mass data, Pima Indians Diabetes data, Cleveland, Hungarian and Switzerland etc., will be subjected to analyze the performance of the proposed technique in class imbalance problem utilizing accuracy, sensitivity and specificity. The results of our proposed method have shown that, our optimal classifier achieves better result when compared to other method. Our method achieves the maximum accuracy value for Mammographic Mass medical dataset and Pima Indians Diabetes data. Both dataset achieves 92.65% of accuracy value for medical data classification.

Reference

[1] S. Kharya, “Using Data Mining Techniques for Diagnosis and Prognosis of Cancer Disease,” *International Journal of Computer Science, Engineering and Information Technology*, Vol. 2, No. 2, pp. 55-66, April 2012.

[2] K. Rajesh, and V. Sangeetha, “Application of Data Mining Methods and Techniques for Diabetes Diagnosis”, *International Journal of Engineering and Innovative Technology*, Vol. 2, No. 3, pp. 224-229, Sep 2012.

[3] M. A. Khaleel, S. K. Pradham and G. N. Dash, “A Survey of Data Mining Techniques on Medical Data for Finding Locally Frequent Diseases”, *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 3, No. 8, pp. 17-22, August 2013.

[4] Q. A. A. Radaideh and E. A. Nagi, “Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance”, *International Journal of Advanced Computer Science and Applications*, Vol. 3, No.2, pp.144-151, 2012.

[5] S. TR and N. B. K. AR, “Hybrid Feature Reduction and Selection for Enhanced Classification of High Dimensional Medical Data”, *In proceeding of IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, pp.1-4, Dec 2013.

[6] H. I. Elshazly, A. M. Elkorany, A. E. Hassanien and A. T. Azar, “Ensemble classifiers for biomedical data: performance evaluation”, *In proceeding of 8th International Conference on Computer Engineering & Systems (ICCES)*, pp.184 - 189, Nov. 2013.

[7] V. Chaurasia and S. Pal, “Data Mining Approach to Detect Heart Dieses”, *International Journal of Advanced Computer Science and Information Technology*, Vol. 2, No. 4, pp. 56-66, 2013.

[8] P. Pujari and J. B. Gupta, “Improving Classification Accuracy by Using Feature Selection and Ensemble

Model", *International Journal of Soft Computing and Engineering*, Vol. 2, No. 2, pp. 380-386, May 2012.

[9] S. Mutalib, R. A. Razak, S. N. S. A. Rahman and A. Mohamed, "Intelligent Classification in Medical Data", *In proceeding of IEEE EMBS International Conference on Biomedical Engineering and Sciences*, pp. 120-124, Dec 2012.

[10] D. Tomar and S. Agarwal, "A survey on Data Mining approaches for Healthcare", *International Journal of Bio-Science and Bio-Technology*, Vol. 5, No. 5, pp. 241-266, 2013.

[11] B. Krawczyk and G. Schaefer, "Ensemble fusion methods for medical data and classification", *In proceeding of 11th symposium on neural network application in electrical engineering*, pp.143-146, 2012.

[12] V. S. Latha, P.Y.L. Swetha, M. Bhavya, G. Geetha and D. K.Suhasini, "Combined Methodology of the Classification Rules For medical Data-Sets", *International Journal of Engineering Trends and Technology*, Vol. 3, No. 1, pp. 32-36, 2012.

[13] L. A. N. Muhammed , "Using Data Mining technique to diagnosis heart disease", *In proceeding of statistic in science, business and engineering*, pp. 1-3, Sep 2012.

[14] K. Rajeswari, V. Vaithiyanathan and S. V. Pede "Feature Selection for Classification in Medical Data Mining", *International conference of emerging Trent and technology in computer science*, Vol. 2, No. 2, pp. 492-497, April 2013.

[15] A. A. and S. A. Hannan, "Data Mining Techniques to Find Out Heart Diseases: An Overview", *International Journal of Innovative Technology and Exploring Engineering*, Vol. 1, No. 4, pp. 18-23, Sep 2012,

[16] M. A. Jabbar, P. Chandra and B. L Deekshatulu, "Knowledge Discovery From Mining Association Rules For Heart Disease Prediction", *Journal of Theoretical and Applied Information Technology*, Vol. 41, No. 2, pp. 166-174, July 2012

[17] I. Saini, D. Singh and A. Khosla, "QRS detection using K-Nearest Neighbor algorithm (KNN) and evaluation on standard ECG databases", *Elsevier Journal of Advanced Research*, Vol. 4, No. 4, pp. 331-344, July 2013

[18] S. Khanmohammadi and M. Rezaeiahari, "AHP based classification algorithm selection for clinical decision support development", *Elsevier Procedia Computer Science*, Vol. 36, pp. 328-334, 2014.

[19] R. Chitra and V. Seenivasagam, "Review Of Heart Disease Prediction System Using Data Mining And Hybrid Intelligent Techniques," *Ictact Journal On Soft Computing*, Vol. 03, No. 04, pp. 605-609, July 2013.

[20] R. Bhuvaneswari and K. Kalaiselvi, "Naive Bayesian Classification Approach in Healthcare Applications", *International Journal of Computer Science and Telecommunications*, Vol. 3, no. 1, pp. 106-112, January 2012.

[21] P. K. Anooj, "Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules", *Elsevier Computer and Information Sciences*, Vol. 24, No. 1, pp. 27-40, January 2012.

[22] M. A. Jabbar, B.L. Deekshatulu and P. Chandra, "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm", *Elsevier Procedia Technology*, Vol. 10, pp. 85-94, 2013.

[23] A. Sudha, P. Gayathri and N. Jaisankar, "Effective Analysis and Predictive Model of Stroke Disease using Classification Methods", *IEEE International Journal of Computer Applications*, Vol. 43, No.14, pp. 26-31, April 2012.

[24] M. Juhola, H. Joutsijoki, H. Aalto and T. P. Hirvonen, "On classification in the case of a medical data set with a complicated distribution," *Elsevier Applied Computing and Informatics*, Vol. 10, No. 2, pp. 52-67, January 2014.

[25] B. Dennis and S. Muthukrishnan, "AGFS: Adaptive Genetic Fuzzy System for medical data classification", *Applied Soft computing*, Vol. 24, pp. 242-252, 2014.

[26] D. A. Adeniyi, Z. Wai, Y. Yongquan, "Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method", *Applied Computing and Informatics*, Vol. 3, 2014.