Title: An Analytical Method for Diseases Prediction Using Machine Learning Techniques

Authors: Mehrbakhsh Nilashi, Othman bin Ibrahim, Hossein Ahmadi, Leila Shahmoradi

PII: S0098-1354(17)30257-0
DOI: http://dx.doi.org/doi:10.1016/j.compchemeng.2017.06.011
Reference: CACE 5842

To appear in: *Computers and Chemical Engineering*

Received date: 27-11-2016
Revised date: 20-5-2017
Accepted date: 6-6-2017

Please cite this article as: Nilashi, Mehrbakhsh., Ibrahim, Othman bin., Ahmadi, Hossein., & Shahmoradi, Leila., An Analytical Method for Diseases Prediction Using Machine Learning Techniques.*Computers and Chemical Engineering* http://dx.doi.org/10.1016/j.compchemeng.2017.06.011

An Analytical Method for Diseases Prediction Using Machine Learning Techniques

Mehrbakhsh Nilashi [a,*], Othman bin Ibrahim [a], Hossein Ahmadi [a], Leila Shahmoradi [a,*]

[a] Faculty of Computing, Universiti Technologi Malaysia, Johor, Malaysia

* Corresponding author e-mail address: nilashidotnet@hotmail.com

Highlights

- An analytical method is proposed for diseases prediction.
- We use EM, PCA, CART and fuzzy rule-based techniques in the proposed method.
- Fuzzy rules are extracted from the medical datasets and used for prediction task.
- The method is tested on public medical datasets from UCI.
- The results show that the method is effective in diseases prediction.

Abstract. The use of medical datasets has attracted the attention of researchers worldwide. Data mining techniques have been widely used in developing decision support systems for diseases prediction through a set of medical datasets. In this paper, we propose a new knowledge-based system for diseases prediction using clustering, noise removal, and prediction techniques. We use Classification and Regression Trees (CART) to generate the fuzzy rules to be used in the knowledge-based system. We test our proposed method on several public medical datasets. Results on Pima Indian Diabetes, Mesothelioma, WDBC, StatLog, Cleveland and Parkinson's telemonitoring datasets show that proposed method remarkably improves the diseases prediction accuracy. The results showed that the combination of fuzzy rule-based, CART with noise removal and clustering techniques can be effective in diseases prediction from real-world medical datasets. The knowledge-based system can assist medical practitioners in the healthcare practice as a clinical analytical method.
Keywords: Machine Learning**; ; ; ,** Diseases Classification, Fuzzy Logic, Analytical Method

1. Introduction

As early as 1997, the potential of data mining for improving the problems in the medical domain had been identified by World Health Organization (WHO) (Gulbinat, 1997). The usefulness of knowledge detection from medical data repositories has been emphasized by WHO as it benefits medical diagnosis and prediction.

Data mining is a process of discovering useful knowledge from database to build a structure (i.e., model or pattern) that can meaningfully interpret the data. Data mining is the process of discovering interesting patterns and knowledge from large amount of data (Han et al., 2001). Data mining uses many machine learning techniques to discover hidden pattern in data. These techniques can be in three main categories which are supervised learning techniques, unsupervised learning techniques and semi-supervised learning techniques (Huang al., 2014). Expert systems developed by machine learning techniques can be used to assist physicians in diagnosing and predicting diseases (Kononenko, 2001). Due to diseases diagnosis importance to mankind, several studies have been conducted on developing methods for their classification (see Table 1).

In this paper, we apply machine learning techniques (supervised and unsupervised) and propose a new hybrid intelligent method using Principal Component Analysis (PCA), Gaussian mixture model with Expectation Maximization (EM), Classification and Regression Trees (CART) and fuzzy rule-based techniques. We then evaluate the proposed method on real-world datasets. These datasets are taken from Data Mining Repository of the University of California, Irvine (UCI) (Newman et al., 1998). The datasets are Pima Indian Diabetes, Mesothelioma, Wisconsin Diagnostic Breast Cancer, StatLog, Cleveland and Parkinson's telemonitoring datasets. In comparison with research efforts found in the literature, our work has the following contributions. In this research:

- a knowledge-based system is proposed for disease diagnostic using EM, PCA, CART and fuzzy rule-based reasoning techniques.
- EM is used for the clustering of data in public medical datasets.
- CART is used for rule discovery to be used in the knowledge-based system.

- PCA is used for dimensionality reduction and dealing with the multi-collinearity problem in the experimental data.
- fuzzy rule-based technique is used for diseases prediction task.

Our study at hand is organized as follows: Section 2 provides the research methodology along with all approaches used in the proposed model. Section 3 presents the method evaluation and finally, conclusions and future work are provided in the Section 4.

2. Method

In the present study, PCA, EM, and fuzzy rule-based techniques are used (see Appendix B). These methodologies are addressed in the following sections. The general framework of proposed model is shown in Fig. 1. In this study, EM clustering is used as an unsupervised classification method to cluster the data of experimental dataset into similar groups. We propose to rely on fuzzy rule-based method to learn the prediction models. We also use PCA for dimensionality reduction because the greatest source of difficulties in using classification methods is the existence of multi-collinearity in many sets of data. In the first step, the data is pre-processed (1). In the second step, EM clustering processing steps are performed to cluster the data (2) and then we apply PCA to reduce the dimensionality of the data and filter out potential noise (3). We then apply CART for discovering the decision rules from the data. Next, prediction models are constructed by fuzzy rule-based method in each cluster (4). In this step, we developed the fuzzy rule based system through several consequent steps which are input fuzzification, generating Membership Functions (MFs), extracting fuzzy rules and output defuzzification. In the fuzzification step, Gaussian MFs are used to determine the degree of inputs that they belong to each of the appropriate fuzzy sets. In addition, for the outputs of model, the Triangular MFs are considered. In the defuzzification step the Centroid of Area (COA) which returns the center of area under the curve (Hellendoorn and Thomas, 1993) is used for deffuzification purpose.

We evaluate the proposed method on real-world datasets. The datasets are taken from Data Mining Repository of the University of California, Irvine (UCI) (Newman et al., 1998). The datasets are Pima Indian Diabetes, Mesothelioma, Wisconsin Diagnostic Breast Cancer (WDBC), StatLog Heart Disease, Cleveland Heart Disease and Parkinson's Telemonitoring (see Appendix A).

3. Results and discussion

The results of the proposed method on real-world datasets are explained in this section. Here, the results of applying all incorporated techniques in the proposed system are discussed.

3.1. Clustering with EM algorithm

We applied the EM clustering on Wisconsin Diagnostic Breast Cancer (WDBC), StatLog Heart Disease, Cleveland Heart Disease, Parkinson's Telemonitoring datasets. In every clustering method, choosing the right number of clusters is important. In EM clustering, with the Gaussian mixture model, the likelihood must be optimized. Hence, for this optimization, the best cluster number is selected by evaluating various values for the number of clusters. It should be noted that according to Pelleg and Moore (2000), we used information theoretic criterion like the Akaike Information Criterion (AIC) (Akaike, 1974) to choose the value optimal number of cluster. Accordingly, in the datasets, we have used a resubstitution AIC estimate and evaluated the number of clusters. Hence, as we used resubstitution AIC estimate to choose the optimal number of cluster which could optimize likelihood, we need to test the number of clusters from n=1 to m, in which the lowest criterion value is obtained for x ($n \leq x \leq m$). As can be seen in Fig 2, we found the lowest (best) criterion values for WDBC, Cleveland, StatLog, Parkinson's telemonitoring, PID and Mesothelioma are respectively for x=5, 7, 8, 13, 6 and 7. Note that we applied 10-fold cross validation to obtain unbiased result. Accordingly, in Fig. 2, we present the various numbers of clusters to select the best cluster based on chosen criterion. Fig. 2 shows that the best criterion value (56730.59123) is obtained for WDBC when 5 clusters are generated by EM. Fig. 2 also shows that the best criterion values (12123.916969, 11399.736308 and 275755.9052) are obtained for Cleveland, StatLog and PD when 7, 8 and 13 clusters are respectively generated by EM. For PID and Mesothelioma, 6 and 7 clusters are respectively selected by EM. For visualizing the dataset clusters into the original space, a PCA is used in order to obtain a 2D representation. It was used to visualize clusters in the scatter plot using the first and second PCs. In Figs. 3(a)-(f), the clusters generated by EM are visualized. As can be seen, we project the observations in the clusters of six datasets (WDBC, Cleveland, StatLog, Parkinson's telemonitoring, PID and Mesothelioma) on the first three dimensions (PC1, PC2 and PC3) generated by PCA.

3.2. PCA and CART Evaluation

Choosing the right number of factors is a crucial problem in PCA. If we select too many factors, we include noise from the sampling fluctuations in the analysis. If we choose too few factors, we lose relevant information, the analysis is incomplete. As we know that the eigenvalue associated to a factor corresponds to its variance. Thus, the eigenvalue indicates the importance of the factor. The higher the value, the higher is the importance of the factor. The eigenvalues that are associated with the factors are indicators for their importance. In our work, we decided to use the rule proposed by Cattell (1966) and create ''scree'' plots, where we plot the eigenvalues of the factors to detect ''elbows'' that indicate possible changes in the structure of the data. We applied the PCA on the clusters obtained by EM algorithm for all experimental datasets. For example, according to the rule proposed by Cattell (1966), in Parkinson's telemonitoring for Cluster 1 and Cluster 13, nine PCs are selected as they provide significant percentage of information. For these clusters, Fig. 4 and Fig. 5 respectively shows the PCs selection based on variance explained and eigenvalues. Following this rule, for Cluster 6, six PCs are selected. For Cluster 3, 5, 9 and 11, eight PCs and for Cluster 2, 4, 10 and 12, five PCs are selected.

In Wisconsin Diagnostic Breast Cancer dataset, for Cluster 1, we selected $k = 6$ factors. Indeed, the eigenvalue ($\lambda_6 \geq 0.4315$) associated with the 6th factor was high. It corresponded to about 92.12% of the variance. In addition, the results of applying PCA showed that for Cluster 2-5, 8,5,6 and 7 PCs were selected which provided 94.55%, 98.42%, 96.32% and 96.18 % of the variance, respectively. We also applied PCA on the clusters of StatLog and Cleveland obtained using EM algorithm. For Cluster 1 of StatLog and Cleveland, we included the elbow into the selection i.e. we selected $k = 9$ factors. Indeed, the eigenvalues associated with the 9th factor was high. In Table 2, we present the selected PCs along with the variance explained of last selected PC for StatLog and Cleveland. Similarly, in PID, for Cluster 1, we included the elbow into the selection i.e. we selected $k = 2$ factors. Indeed, the eigenvalues associated with the 2nd factor was high. In addition, three PCs for Clusters 2 and 4 and four PCs for Clusters 3, 5 and 6 were chosen. In Table 3, we present the selected PCs along with the variance explained of last selected PC for Mesothelioma dataset.

After applying PCA, we used datsets for decision rule discovery by generating decision trees. In this step, we used CART and applied this techniques on all clusters. In Table C.1 in the Appendix, the sample of induced decision tree using CART for PD is presented. We can see that the root node of the tree is split with the MDVP:Jitter:PPQ5 with the cut point 0.0047. The other attributes that appear in the tree are DFA, MDVP:Jitter (Abs), RPDE, PPE, HNR and Shimmer:APQ5. In Table C.2 and Table C.3 in the Appendix, the samples of induced decision tree for breast cancer and diabetes are presented. In Fig. C.1, Fig. C.2 and Fig. C.3 in the Appendix, the diagrams of decision tree for Parkinson, Breast Cancer and Diabetes are visualized, respectively.

3.3. Fuzzy Logic Evaluation

The aim of this study is to predict diseases using a set of input parameters of datasets. To do so, we constructed prediction models by discovering the fuzzy rules using CART from the public medical datasets and applying them in fuzzy rule reasoning technique to generalize the relationship between input and output parameters ($Y = f(X_1, X_2, \ldots, X_n)$) for accurate prediction of diseases. In this relationship, $X_1, X_2, \ldots, X_n$ stands for n input parameters of datasets and Y stands for output parameter Class of a disease. For example for PID dataset, input parameters are (X): Number of times pregnant, Plasma glucose concentration in a 2 hour oral glucose tolerance test, Diastolic blood pressure, Triceps skin fold thickness, 2 hour serum insulin, Body mass index, Diabetes pedigree function, Age and output parameter (Y) is the Class of disease. After discovering the fuzzy rules using CART, the appropriate Membership Functions (MFs) have been constructed them to be used in fuzzification step of rule-based technique. Based on clusters generated by EM, the ranges of MFs have been defined for each input and output variables. We considered Gaussian and Triangular MFs (see Appendix B) for input and output variables, respectively. In fact, In addition, for all input and output variables in the FIS models, we have used three linguistic terms which are “Low”, “Moderate” and “High”. In Figs. 6(a)-(d), we present the MFs for four input variables of PID. It should be noted that as we have six clusters for PID, therefore, totally six prediction models have been developed in fuzzy logic toolbox.

We implemented the fuzzy model based on Mamdani algorithm using fuzzy logic toolbox provided by MATLAB software package (Folorunso and Mustapha, 2015). Combining both the input MFs and the output

MFs with the rules above, three-dimensional curve can be obtained to give a snapshot relationship between the inputs and output. Illustrating the interdependency between inputs and output is helpful in revealing level of presenting a disease. Fig. 7 illustrates the interdependency of diabetes level and input variables through control surfaces obtained from the fuzzy rules discovered from the PID dataset and defined membership functions. The sample of fuzzy control surfaces in this figure for output variable (class of disease) is developed from the corresponding rules base induced by CART for different inputs. In Fig. 7a, the fuzzy control surface is visualized on Number of times pregnant and Plasma glucose concentration in a 2 hour oral glucose tolerance test. In Fig. 7b, the fuzzy control surface is visualized on Number of times pregnant and Triceps skin fold thickness. Similarly, in Fig. 7c and d, the fuzzy control surfaces are respectively visualized on Number of times pregnant and Diabetes pedigree function, and Triceps skin fold thickness and Diastolic blood pressure. Fig. 8 illustrates the interdependency of presence of heart disease and input variables through the control surfaces obtained from the fuzzy rules discovered from the Cleveland dataset and defined membership functions. In fact, these figures represent the mapping from each two parameters of PID and Cleveland datasets to respectively diabetes and heart diseases presence. In addition, the surface plots depict the impacts of the diseases parameters on the level of diabetes and heart diseases presence. Note that the colors in the plots show the behavior of the FIS based on the fuzzy rules, inputs and output parameters.

The fuzzy rule viewer of the established prediction models can better demonstrates the presence of diseases over the change in values of all inputs parameters. It displays a roadmap of the whole fuzzy inference process. In Fig. 9, the prediction of diabetes disease is performed by the fuzzy rules from input parameters (X_1-X_8) of PID dataset. It shows that how the prediction is performed using eight input parameters of PID and eleven fuzzy rules. In fact, the rule viewer presented in this figure demonstrates the changes in the defuzzified output parameter (level of diabetes disease) using COA according to the changes in the fuzzy input variables of PID such as Number of times pregnant, Plasma glucose concentration in a 2 hour oral glucose tolerance test, Diastolic blood pressure, Triceps skin fold thickness, 2 hour serum insulin, Body mass index, Diabetes pedigree function and Age. From Fig. 9, it is clear that for X1 (4.83), X2 (81), X3 (65), X4 (30), X5 (105), X6 (39.5), X7 (0.65) and X8 (31.5) the output, level of diabetes disease, is computed as 0.249.

3.4. Evaluation of Method

For evaluating the proposed method on prediction type datasets, two measures of accuracy are used to determine the model capability for predicting the outputs. In this regard, the models are evaluated by two estimators Mean absolute error and coefficient of determination R^2. The coefficient of determination R^2 provides a value between [0, 1] about the training of the proposed network. A value closer to 1 stands for the success of learning. These estimators are determined by Eqs. (1) and (2).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|(\hat{y}_i - y_i)\right| \quad (3)$$

$$R^2 = SSR / SST = 1 - SSE / SST = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y}_i)^2} \quad (4)$$

where *n* is the number of observations or samples, y is the observed value, $\hat{y}$ is the predicted value and $\bar{y}$ is the average of [y_1,y_2, . . . ,y_n].

For classification type datasets, the ROC chart has been defined as a graphical display that provides the measure of the prediction/classification accuracy of the model by two measures of accuracy, the specificity and sensitivity. Specificity is a measure of accuracy for predicting nonevents that is equal to the true negative/total actual negative of a classifier for a range of cutoffs. Sensitivity is a measure of accuracy for predicting events that is equal to the true positive/total actual positive. We perform several experiments and compare the results with the techniques Neural Network (NN) and SVR with PCA.

For SVR, the algorithm of SVR is LIBSVM developed by Chang and Lin (2011). Specifically, we used epsilon-SVR to develop the prediction models. In this research, different epsilon values (epsilon=0.001, epsilon=0.01 and epsilon=0.1) were tested and we found that the best prediction accuracy is obtained for epsilon=0.001. Hence, the epsilon value was set to be 0.001. In addition, we considered RBF kernel for epsilon-SVR. For RBF kernel, the kernel elements were cost (*C)* and γ. To select the best values for these two parameters, we used *k*-fold cross-validation (*k*=10) as a statistical model selection method. For each fold, we also trained the models with 10 trials. Using 10-cross-validation, the data used in the research were divided into 10 equally sized subsets. Accordingly, 9 subsmaples were used as as the training data and a single subsample was retained as the test data and the remaining 9 subsamples were used to test the method. The learning models were then trained on 9 subsamples. After training process, the model was tested on the single subset and the 10 results from each of the folds could be averaged to produce a single generalization estimation. By trying several values for the parameters C and γ, we then set the value of penalty parameter C and γ in RBF kernel equal to the optimal one determined via 10-fold cross-validation. The method of choosing C and γ was exhaustive search as it is the most popular method in determining the SVR parameters. Specifically, we tried exponentially growing sequences of C form 2^{-15} to 2^{10} and γ from 2^{-10} to 2^{9} to find the optimal values. After testing there ranges, we found that the best (C, γ) is (2^{2},2^{-2}).

For NN, feedforward back-propagation with single output is used for the prediction task. The model used in this research has three layer. For training of NN, different back-propagation techniques are used. Resilient back-propagation (Saini, 2008), Conjugate gradient back-propagation (Wang et al., 2007) and Levenberg Marquardt (Vakili et al., 2016) algorithm are some of the techniques which are used for training. In this research, we used NN with resilient back-propagation training algorithm.

For error estimation in the clusters of EM, the averages MAE and R^2 were calculated as presented in Table 4 and Fig. 10. The MAE and R^2 were calculated based on output (Motor-UPDRS and Total-UPDRS) prediction.

The results demonstrate that the accuracy of proposed method using EM, PCA and fuzzy logic is the best on Total-UPDRS and Motor-UPDRS in relation to other methods. Comparison of performance in predicting Motor-UPDRS and Total-UPDRS for PCA-SVR and PCA-NN on experimental datasets show that the proposed method is more accurate for the disease prediction. In relation to the PCA-NN, our method helps to improve the prediction accuracy (R^2) of Motor-UPDRS and Total-UPDRS by more than 15% and 24% for Motor-UPDRS and Total-UPDRS, respectively. In addition, in relation to the PCA-SVR, our method helps to improve the prediction accuracy of Motor-UPDRS and Total-UPDRS by more than 5% and 16% for Motor-UPDRS and Total-UPDRS, respectively. Accordingly, it can be found that the accuracy of methods which uses fuzzy rule-based method with EM and PCA is higher than those methods that only use PCA. These show the effectiveness of incorporating the clustering and PCA techniques for the prediction accuracy of PD progression. In addition, the superiority of EM-PCA-Fuzzy Rule-Based can be explained by the fact that these methods have used clustering and noise removal techniques before the prediction of Motor-UPDRS and Total-UPDRS while the other methods solely rely on prediction techniques with PCA.

In Table 5 and Fig. 11, the results of methods for classification type datasets are presented. Table 5 presents the accuracy results of applying classification techniques on Cleveland, StatLog, PID, Mesothelioma and Wisconsin Diagnostic Breast Cancer datasets. From the results, we can see that proposed method outperforms PCA-SVM and PCA-KNN. In addition, on the classification type datasets, the proposed method which uses PCA and EM obtained a highest accuracy with AUC values 0.914, 0.928, 0.932, 0.929 and 0.936 for StatLog, Cleveland, Wisconsin Diagnostic Breast Cancer, Pima Indian Diabetes and Mesothelioma, respectively. The results show that the difference of accuracy obtained by EM-PCA-Fuzzy Rule-Based is relatively higher than two other methods.

4. Conclusion and future work

In this paper, we propose a new knowledge-based system for disease diagnosis using machine learning techniques. EM and PCA were respectively used for clustering and addressing multi-collinearity in the datasets. We then used Classification and Regression Trees (CART) to generate the fuzzy rules to be used in the knowledge-based system of fuzzy rule-based reasoning method for the disease prediction. In order to analyze the effectiveness of the proposed method and validate the system, several experiments were conducted on public medical datasets. The datasets were taken from Data Mining Repository of the University of California, Irvine (UCI) which are Pima Indian Diabetes, Mesothelioma, Wisconsin Diagnostic Breast Cancer, StatLog, Cleveland and Parkinson's telemonitoring datasets. The results indicated that the method which combines clustering, PCA, and fuzzy rule-based techniques obtain good prediction accuracy.

Our method proposed in this study has been evaluated by the public datasets from UCI which have input and output parameters for a specific disease diagnostic. In addition, compared to the big healthcare data, the nature of the data in these datasets is not complex. In addition, in case of big healthcare data which can be complex datasets with unique characteristics, the future studies need to consider this issue in the development of new methods in order to overcome the challenges of data processing time and take advantage of big data. Furthermore, as big healthcare data include multi-spectral, heterogeneous, imprecise and incomplete observations (e.g., diagnosis) which are derived from different sources, therefore new methods are needed and

relying solely on conventional machine learning techniques may not be a sophisticated way of predicting diseases.

All of the approaches used in this study, may also be applicable to other diseases classification problems which include datasets with same nature used in this study. However, there is still plenty of work in conducting researches on clustering, noise removal and fuzzy rule-based techniques for disease diagnosis in order to exploit all their potential and usefulness. In the future work, more attention should be paid to the datasets for disease classification and prediction using the incremental machine learning approaches. Hence, in our future study, we plan to evaluate the proposed method on additional datasets and in particular on large datasets to show the effectiveness of the method for computation time of large data. In addition, our future work investigates that how the proposed method can be extended to be applicable to the other types of datasets in medical domain.

References

Abdel-Zaher, A. M., & Eldeib, A. M. (2016). Breast cancer classification using deep belief networks. Expert Systems with Applications, 46, 139-144.

Akaike H. (1974). A new look at the statistical model identification. Automatic Control, IEEE Transactions on, 19(6), 716-723.

Anooj, P. K. (2012). Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules. Journal of King Saud University-Computer and Information Sciences, 24(1), 27-40.

Aslam, M. W., Zhu, Z., & Nandi, A. K. (2013). Feature generation using genetic programming with comparative partner selection for diabetes classification. Expert Systems with Applications, 40(13), 5402-5412.

Åström F., & Koker R. (2011). A parallel neural network approach to prediction of Parkinson's Disease. Expert systems with applications, 38(10), 12470-12474.

Avci, D., & Dogantekin, A. (2016). An Expert diagnosis system for Parkinson disease based on genetic algorithm-wavelet kernel-extreme learning machine. Parkinson's disease, 2016.

Babu G. S., & Suresh S. (2013). Parkinson's disease prediction using gene expression–A projection based learning meta-cognitive neural classifier approach. Expert Systems with Applications, 40(5), 1519-1529.

Behroozi, M., & Sami, A. (2016). A multiple-classifier framework for Parkinson's disease detection based on various vocal tests. International journal of telemedicine and applications, 2016.

Bhardwaj, A., & Tiwari, A. (2015). Breast cancer diagnosis using genetically optimized neural network model. Expert Systems with Applications, 42(10), 4611-4620.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984).Classification and regression trees. CRC press.

Buza, K., & Varga, N. Á. (2016). ParkinsoNET: Estimation of UPDRS Score Using Hubness-Aware Feedforward Neural Networks. Applied Artificial Intelligence, 30(6), 541-555.

Çalişir, D., & Doğantekin, E. (2011). An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine Classifier. Expert Systems with Applications, 38(7), 8311-8315.

Cattell R. B. (1966). The scree test for the number of factors. Multivariate behavioral research, 1(2), 245-276.

Chen H. L., Huang C. C., Yu, X. G., Xu, X., Sun, X., Wang, G., & Wang, S. J. (2013). An efficient diagnosis system for detection of Parkinson's disease using fuzzy k-nearest neighbor approach. Expert systems with applications, 40(1), 263-271.

Chen, C. H. (2014). A hybrid intelligent model of analyzing clinical breast cancer data using clustering techniques with feature selection. *Applied Soft Computing*, *20*, 4-14.

Das R. (2010). A comparison of multiple classification methods for diagnosis of Parkinson disease. Expert Systems with Applications, 37(2), 1568-1572.

Das, R., Turkoglu, I., & Sengur, A. (2009). Effective diagnosis of heart disease through neural networks ensembles. Expert systems with applications, 36(4), 7675-7680.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society. Series B (methodological), 1-38.

Dogantekin, E., Dogantekin, A., Avci, D., & Avci, L. (2010). An intelligent diagnosis system for diabetes on linear discriminant analysis and adaptive network based fuzzy inference system: LDA-ANFIS. Digital Signal Processing, 20(4), 1248-1255.

Ephzibah, E. P. (2010). Cost effective approach on feature selection using genetic algorithms and LS-SVM classifier. IJCA Special Issue on Evolutionary Computation for Optimization Techniques, ECOT.

Er, O., Tanrikulu, A. Ç., & Abakay, A. (2015). Use of artificial intelligence techniques for diagnosis of malignant pleural mesothelioma. Dicle Tıp Dergisi, 42(1).

Erkaymaz, O., & Ozer, M. (2016). Impact of small-world network topology on the conventional artificial neural network for the diagnosis of diabetes. Chaos, Solitons & Fractals, 83, 178-185.

Eskidere, Ö., Erta F., & Hanilçi C. (2012). A comparison of regression methods for remote tracking of Parkinson's disease progression. Expert Systems with Applications, 39(5), 5523-5528.

Folorunso, O., & Mustapha, O. A. (2015). A fuzzy expert system to Trust-Based Access Control in crowdsourcing environments. Applied Computing and Informatics, 11(2), 116-129.

Froelich W., Wrobel K., & Porwik P. (2015). Diagnosis of Parkinson's Disease Using Speech Samples and Threshold-Based Classification. Journal of Medical Imaging and Health Informatics, 5(6), 1358-1363.

Ganji, M. F., & Abadeh, M. S. (2011). A fuzzy classification system based on Ant Colony Optimization for diabetes disease diagnosis. Expert Systems with Applications, 38(12), 14650-14659.

Gulbinat, W. (1997). What is the role of WHO as an intergovernmental organisation In: The coordination of telematics in healthcare. World Health Organisation. Geneva, Switzerland at http://www. hon. ch/libraray/papers/gulbinat. html.

Guo P. F., Bhattacharya P., & Kharma N. (2010). Advances in detecting Parkinson's disease. In Medical Biometrics (pp. 306-314). Springer Berlin Heidelberg.

Han J., & Kamber M. (2001). Data mining: concepts and techniques. Morgan Kaufmann, San Francisco, Calif, USA, 2nd edition, 2011.

Hariharan M., Polat K., & Sindhu R. (2014). A new hybrid intelligent system for accurate detection of Parkinson's disease. Computer methods and programs in biomedicine, 113(3), 904-913.

Hayashi, Y., & Yukita, S. (2016). Rule extraction using Recursive-Rule extraction algorithm with J48graft combined with sampling selection techniques for the diagnosis of type 2 diabetes mellitus in the Pima Indian dataset. Informatics in Medicine Unlocked, 2, 92-104.

Hellendoorn, H., & Thomas, C. (1993). Defuzzification in fuzzy controllers.Journal of Intelligent & Fuzzy Systems, 1(2), 109-123.

Huang, G., Song, S., Gupta, J. N., & Wu, C. (2014). Semi-supervised and unsupervised extreme learning machines. IEEE Transactions on Cybernetics, 44(12), 2405-2417.

Jang, J. S. R., Sun, C. T., & Mizutani, E. (1997). Fuzzy inference systems. Neuro-fuzzy and soft computing: a computational approach to learning and machine intelligence, 73

Jang, J. S., & Sun, C. T. (1995). Neuro-fuzzy modeling and control. Proceedings of the IEEE, 83(3), 378-406.

Janssen, J. A. E. B., Krol, M. S., Schielen, R. M. J., Hoekstra, A. Y., & de Kok, J. L. (2010). Assessment of uncertainties in expert knowledge, illustrated in fuzzy rule-based models. Ecological Modelling, 221(9), 1245-1251.

Jung, Y. G., Kang, M. S., & Heo, J. (2014). Clustering performance comparison using K-means and expectation maximization algorithms. Biotechnology & Biotechnological Equipment, 28(sup1), S44-S48.

Kahramanli, H., & Allahverdi, N. (2008). Design of a hybrid system for the diabetes and heart diseases. Expert Systems with Applications, 35(1), 82-89.

Kahramanli, H., & Allahverdi, N. (2008). Design of a hybrid system for the diabetes and heart diseases. Expert systems with applications, 35(1), 82-89.

Karabatak, M. (2015). A new classifier for breast cancer detection based on Naïve Bayesian. Measurement, 72, 32-36.

Kausar, N., Palaniappan, S., Samir, B. B., Abdullah, A., & Dey, N. (2016). Systematic analysis of applied data mining based optimization algorithms in clinical attribute extraction and classification for diagnosis of cardiac patients. In Applications of intelligent optimization in biology and medicine (pp. 217-231). Springer International Publishing.

Chang, C. C., & Lin, C. J. (2011). LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST), 2(3), 27

Kononenko, I. (2001). Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in medicine, 23(1), 89-109.

Li D. C., Liu C. W., & Hu S. C. (2011). A fuzzy-based data transformation for feature extraction to increase classification performance with small medical data sets. Artificial Intelligence in Medicine, 52(1), 45-52.

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

Marcano-Cedeño, A., Quintanilla-Domínguez, J., & Andina, D. (2011). WBCD breast cancer database classification applying artificial metaplasticity neural network. Expert Systems with Applications, 38(8), 9573-9579.

Mitra, P., Pal, S. K., & Siddiqi, M. A. (2003). Non-convex clustering using expectation maximization algorithm with rough set initialization. Pattern Recognition Letters, 24(6), 863-873.

Moore, B. C. (1981). Principal component analysis in linear systems: Controllability, observability, and model reduction. Automatic Control, IEEE Transactions on, 26(1), 17-32.

Nahar, J., Imam, T., Tickle, K. S., & Chen, Y. P. P. (2013). Association rule mining to detect factors which contribute to heart disease in males and females. Expert Systems with Applications, 40(4), 1086-1093.

Naranjo, L., Pérez, C. J., Campos-Roca, Y., & Martín, J. (2016). Addressing voice recording replications for Parkinson's disease detection. Expert Systems with Applications, 46, 286-292.

Nathiya G., Punitha S. C., & Punithavalli M. (2010). An analytical study on behavior of clusters using k means, em and k- means algorithm. International Journal of Computer Science and Information Security, 7(3), 155-190.

Newman, D. J., Hettich, S., Blake, C. L., Merz, C. J., & Aha, D. W. (1998). UCI repository of machine learning databases. Department of Information and Computer Science, University of California, Irvine, CA. In 1998 of Conference, http://archive. ics. uci. edu/ml/datasets. html.

Nguyen, T., Khosravi, A., Creighton, D., & Nahavandi, S. (2015a). Classification of healthcare data using genetic fuzzy logic system and wavelets. Expert Systems with Applications, 42(4), 2184-2197.

Nguyen, T., Khosravi, A., Creighton, D., & Nahavandi, S. (2015b). Medical data classification using interval type-2 fuzzy logic system and wavelets. Applied Soft Computing, 30, 812-822.

Nilashi, M., & Ibrahim, O. B. (2014b). A model for detecting customer level intentions to purchase in B2C websites using TOPSIS and fuzzy logic rule-based system. Arabian Journal for Science and Engineering, 39(3), 1907-1922.

Nilashi, M., Ahmadi, H., Shahmoradi, L., Salahshour, M., & Ibrahim, O. (2017). A Soft Computing Method for Mesothelioma Disease Classification. Journal of Soft Computing and Decision Support Systems, 4(1), 16-18.

Nilashi, M., bin Ibrahim, O., & Ithnin, N. (2014a). Hybrid recommendation approaches for multi-criteria collaborative filtering. Expert Systems with Applications, 41(8), 3879-3900.

Nilashi, M., bin Ibrahim, O., & Ithnin, N. (2014c). Multi-criteria collaborative filtering with high accuracy using higher order singular value decomposition and Neuro-Fuzzy system. Knowledge-Based Systems, 60, 82-101.

Nilashi, M., Jannach, D., bin Ibrahim, O., & Ithnin, N. (2015). Clustering-and regression-based multi-criteria collaborative filtering with incremental updates. Information Sciences, 293, 235-250.

Nilashi, M., Salahshour, M., Ibrahim, O., Mardani, A., Esfahani, M. D., & Zakuan, N. (2016). A New Method for Collaborative Filtering Recommender Systems: The Case of Yahoo! Movies and TripAdvisor Datasets. Journal of Soft Computing and Decision Support Systems, 3(5), 44-46.

Onan, A. (2015). A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer. Expert Systems with Applications.

Ozcift, A. (2012). SVM feature selection based rotation forest ensemble classifiers to improve computer-aided diagnosis of Parkinson disease. Journal of medical systems, 36(4), 2141-2147.

Polat K. (2012). Classification of Parkinson's disease using feature weighting method on the basis of fuzzy C-means clustering. International Journal of Systems Science, 43(4), 597-609.

Polat, K., & Güneş, S. (2007). Breast cancer diagnosis using least square support vector machine. *Digital Signal Processing*, *17*(4), 694-701.

Polat, K., Güneş, S., & Arslan, A. (2008). A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine. Expert systems with applications, 34(1), 482-487.

Quinlan, J. R. (1987). Simplifying decision trees. International journal of man-machine studies, 27(3), 221-23

Rout, S. (2012). Fuzzy petri net application: Heart disease diagnosis. Fuzzy Systems, 4(4), 124-131.

Şahan, S., Polat, K., Kodaz, H., & Güneş, S. (2007). A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis. *Computers in Biology and Medicine*, *37*(3), 415-423.

Saini, L. M. (2008). Peak load forecasting using Bayesian regularization, Resilient and adaptive backpropagation learning based artificial neural networks. Electric Power Systems Research, 78(7), 1302-1310.

Shao, Y. E., Hou, C. D., & Chiu, C. C. (2014). Hybrid intelligent modeling schemes for heart disease classification. Applied Soft Computing, 14, 47-52.

Shilaskar, S., & Ghatol, A. (2013). Feature selection for medical diagnosis: Evaluation for cardiovascular diseases. Expert Systems with Applications, 40(10), 4146-4153.

Soni, J., Ansari, U., Dipesh Sharma, 2011, "Intelligent and Effective Heart Disease Prediction System using Weighted Associative classifiers". International Journal on Computer Science and Engineering, vol 3, No. 6, pp.2385- 2392.

Temurtas, H., Yumusak, N., & Temurtas, F. (2009). A comparative study on diabetes disease diagnosis using neural networks. Expert Systems with applications, 36(4), 8610-8615.

Übeyli, E. D. (2007). Implementing automated diagnostic systems for breast cancer detection. Expert Systems with Applications, 33(4), 1054-1062.

Vakili, M., Karami, M., Delfani, S., & Khosrojerdi, S. (2016). Experimental investigation and modeling of thermal radiative properties of f-CNTs nanofluid by artificial neural network with Levenberg–Marquardt algorithm. International Communications in Heat and Mass Transfer, 78, 224-230.

Wang, C. H., Kao, C. H., & Lee, W. H. (2007). A new interactive model for improving the learning performance of back propagation neural network. Automation in construction, 16(6), 745-758.

Zadeh, L. A. (1965). Fuzzy sets. Information and control, 8(3), 338-353.

Zheng, B., Yoon, S. W., & Lam, S. S. (2014). Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Systems with Applications, 41(4), 1476-1482.

Pelleg, D., & Moore, A. W. (2000, June). X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In ICML (pp. 727-734).

Ordonez, C., & Omiecinski, E. (2002, November). FREM: fast and robust EM clustering for large data sets. In Proceedings of the eleventh international conference on Information and knowledge management (pp. 590-599). ACM.

Bhatia, S., Prakash, P., & Pillai, G. N. (2008, October). SVM based decision support system for heart disease classification with integer-coded genetic algorithm to select critical features. In Proceedings of the World Congress on Engineering and Computer Science, WCECS (pp. 22-24).

Adeli, A., & Neshat, M. (2010, March). A fuzzy expert system for heart disease diagnosis. In Proceedings of International Multi Conference of Engineers and Computer Scientists, Hong Kong (Vol. 1).

Bhattacharya I., & Bhatia M. P. S. (2010, September). SVM classification to distinguish Parkinson disease patients. In Proceedings of the 1st Amrita ACM-W Celebration on Women in Computing in India (p. 14). ACM.

Gudadhe, M., Wankhade, K., & Dongre, S. (2010, September). Decision support system for heart disease based on support vector machine and artificial neural network. In Computer and Communication Technology (ICCCT), 2010 International Conference on (pp. 741-745). IEEE.

Ghumbre, S., Patil, C., & Ghatol, A. (2011, December). Heart disease diagnosis using support vector machine. In International conference on computer science and information technology (ICCSIT') Pattaya.

Peterek T., Dohnalek P., Gajdos, P., & Smondrk M. (2013, December). Performance evaluation of Random Forest regression model in tracking Parkinson's disease progress. In Hybrid Intelligent Systems (HIS), 2013 13th International Conference on (pp. 83-87). IEEE.

Jain, S., & Shetty, S. (2016, April). Improving accuracy in noninvasive telemonitoring of progression of Parkinson'S Disease using two-step predictive model. In Electrical, Electronics, Computer Engineering and their Applications (EECEA), 2016 Third International Conference on (pp. 104-109). IEEE.

Al-Fatlawi, A. H., Jabardi, M. H., & Ling, S. H. (2016, July). Efficient diagnosis system for Parkinson's disease using deep belief network. In Evolutionary Computation (CEC), 2016 IEEE Congress on (pp. 1324-1330). IEEE.

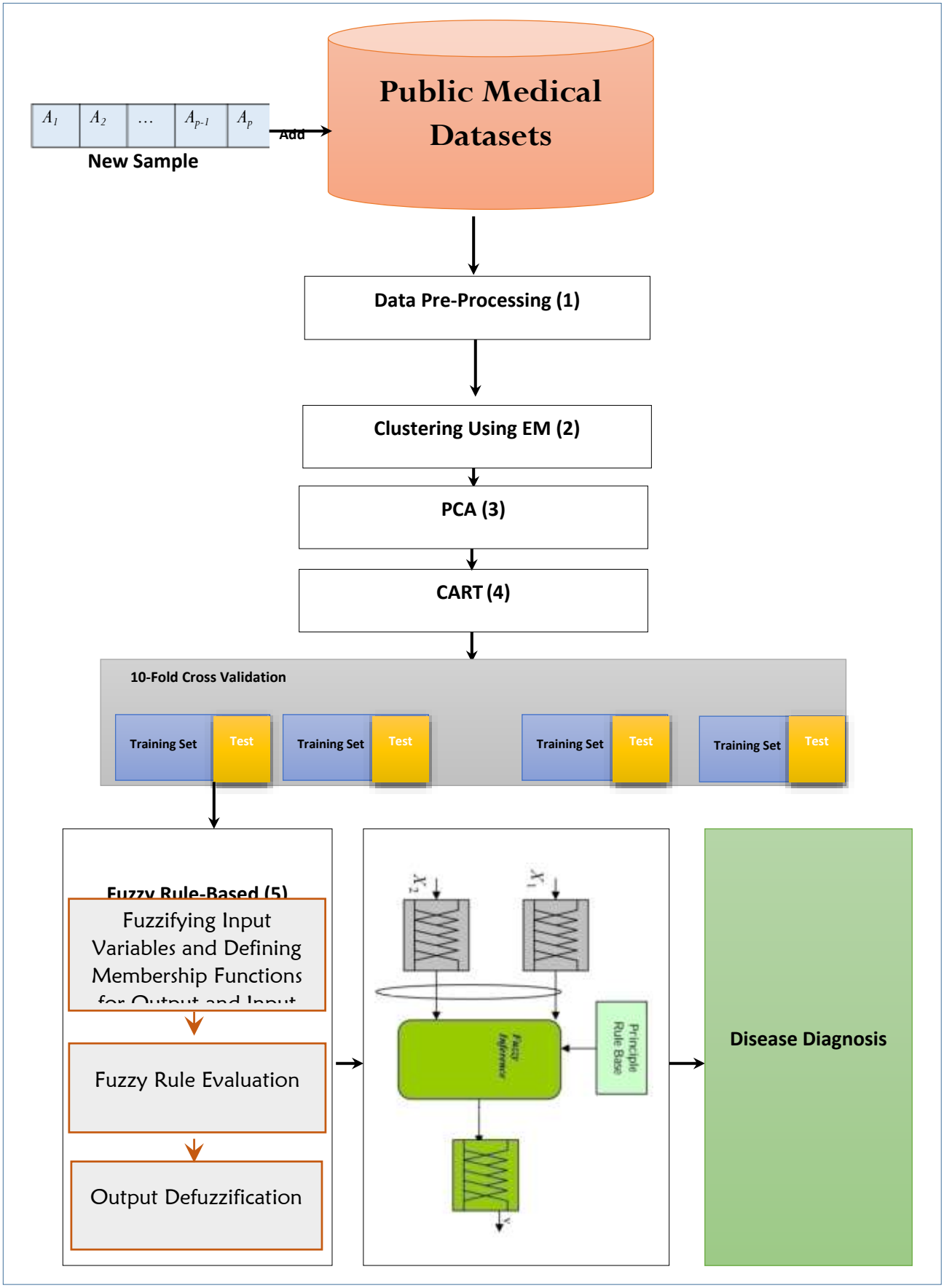
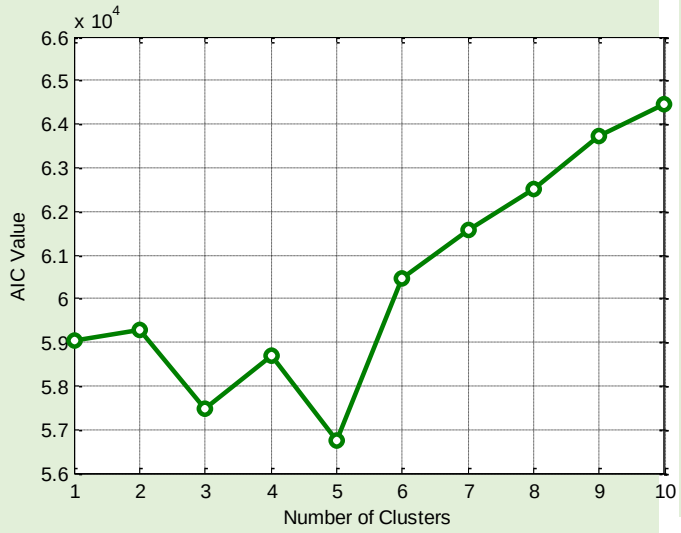
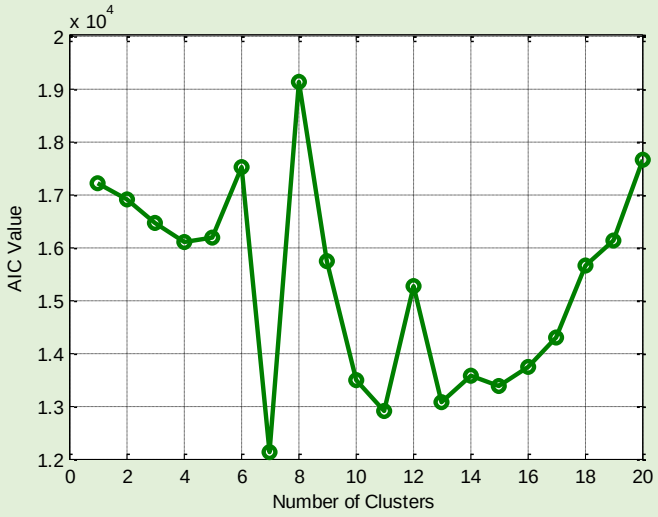


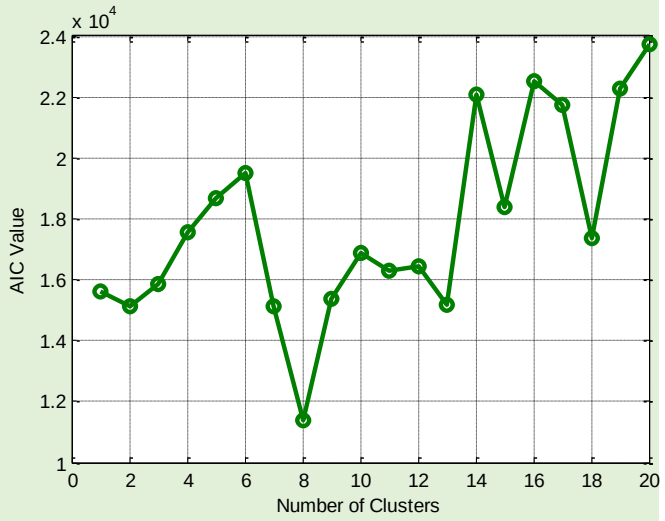
Fig. 1. Proposed method for the diseases diagnosis



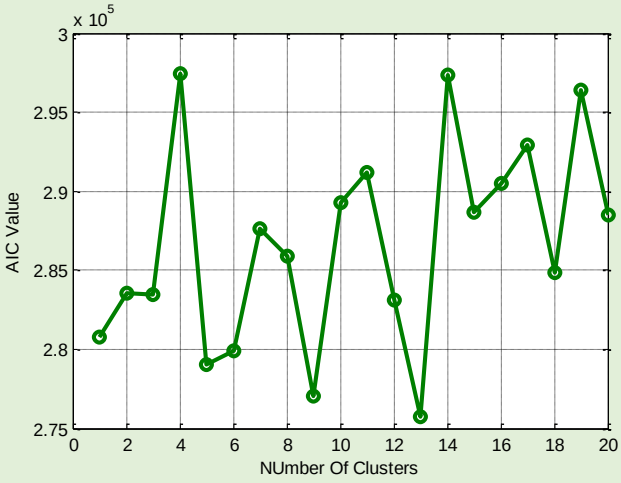
(a)WDBC



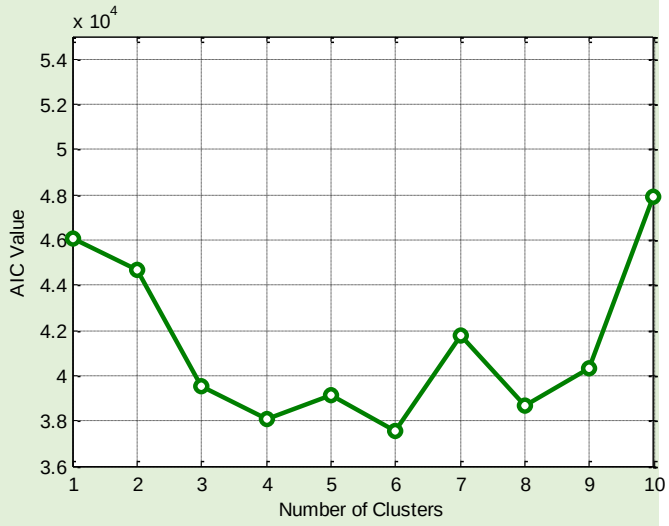
(b)Cleveland



(c)StatLog



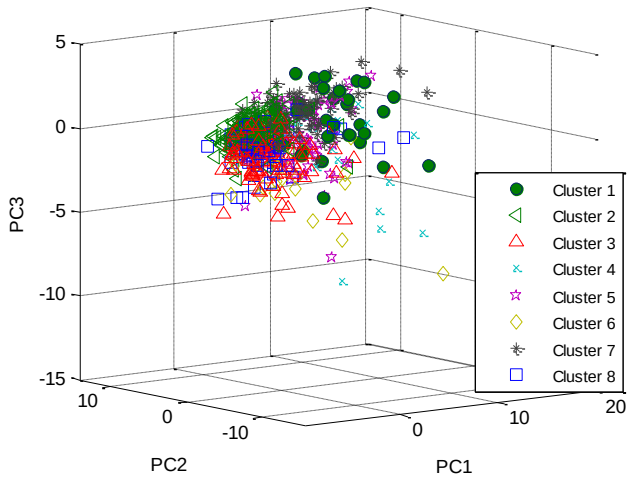
(d)Parkinson's telemonitoring



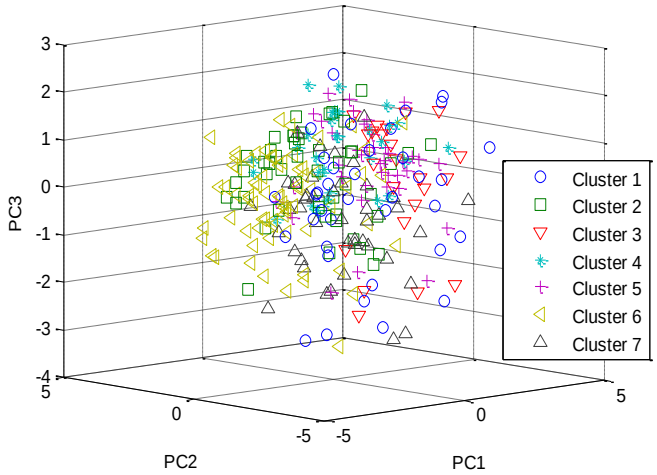
(e)PID

(f)Mesothelioma

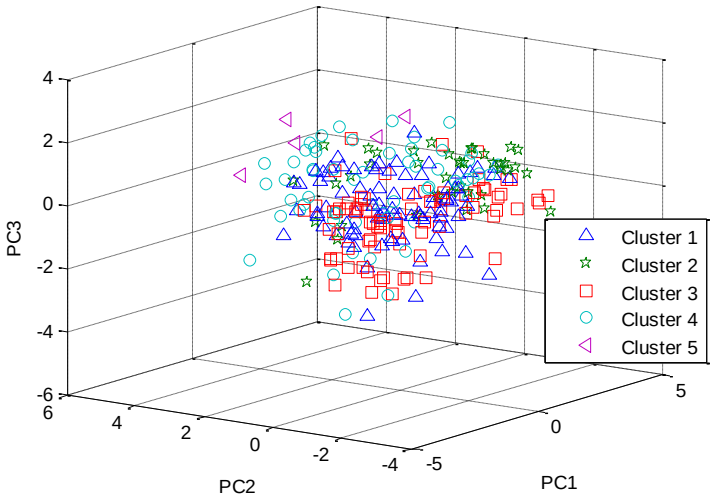
Fig. 2. Best cluster based on chosen criterion for experimental datasets



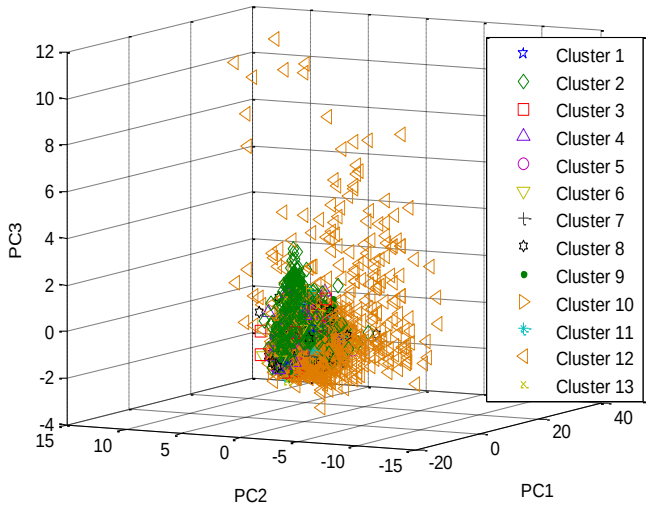
(a)WDBC



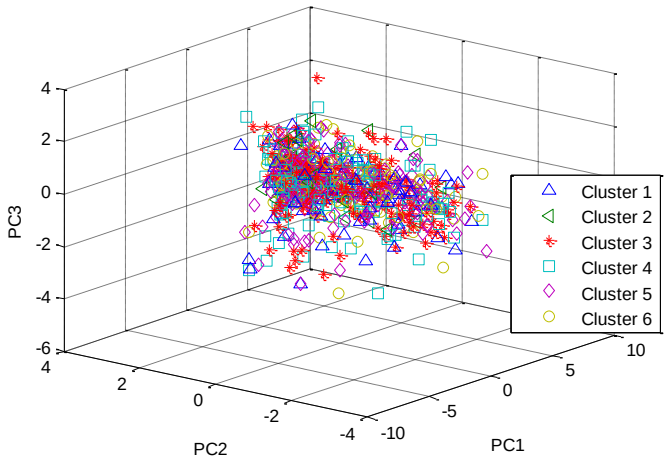
(b)Cleveland



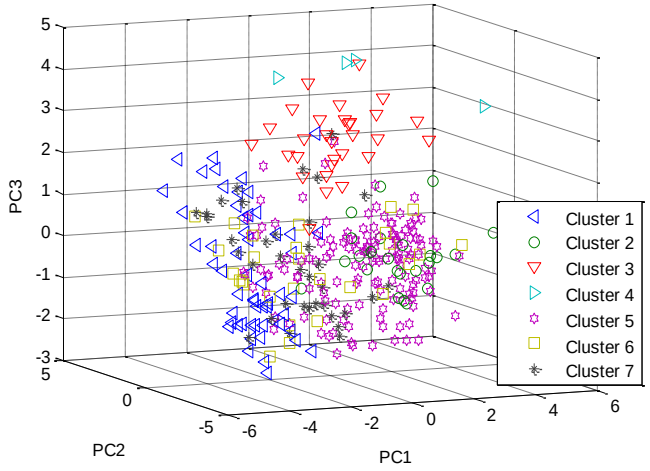
(c)StatLog



(d)Parkinson's telemonitoring

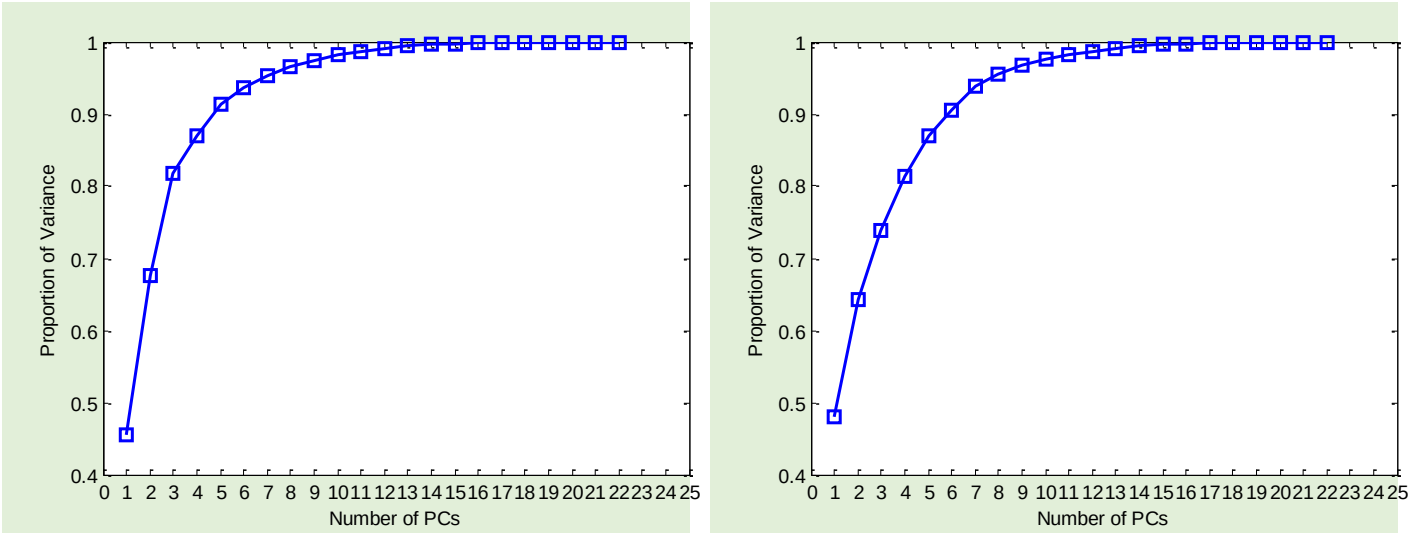


(e)PID



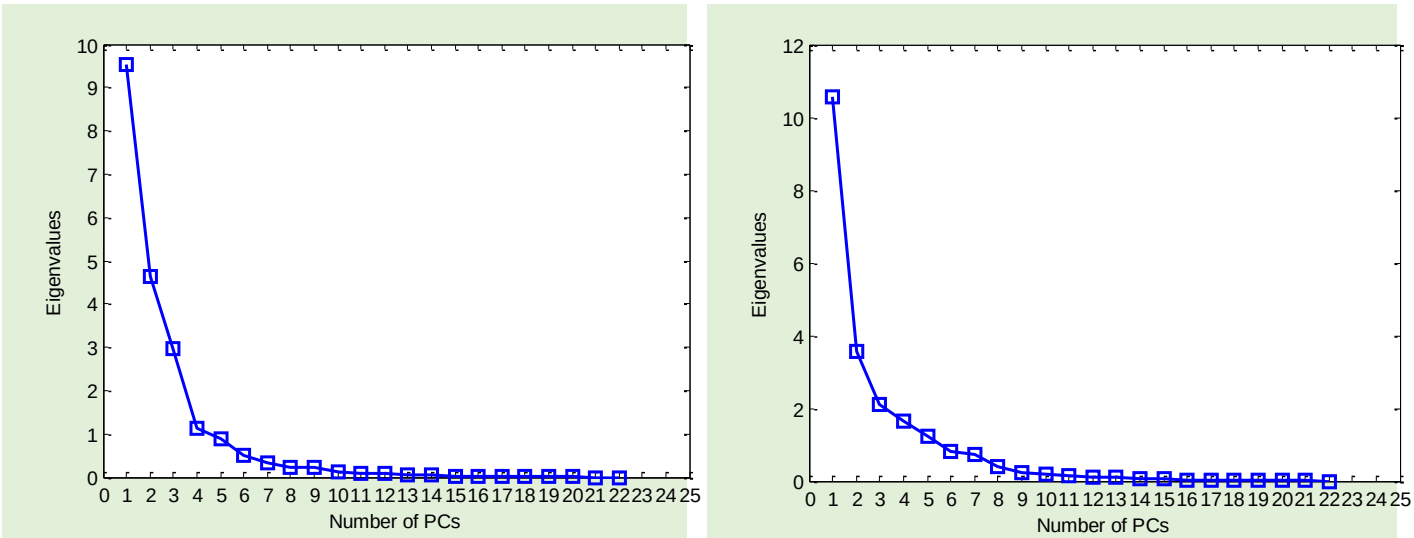
(f)Mesothelioma

Fig. 3 Clusters visualization of experimental datasets



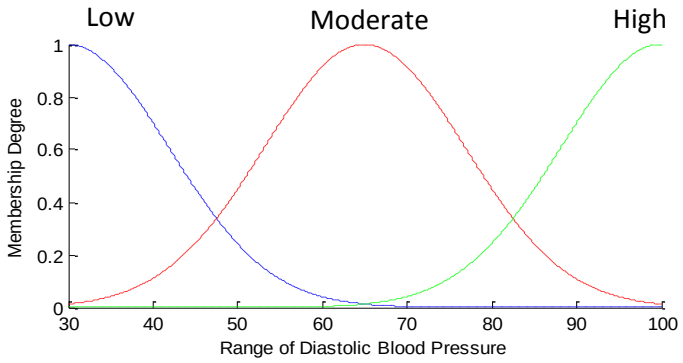
(a) Cluster 1 **(b)** Cluster 13

Fig. 4. Variance explained for (a) Cluster 1 and (b) Cluster 13 of PD

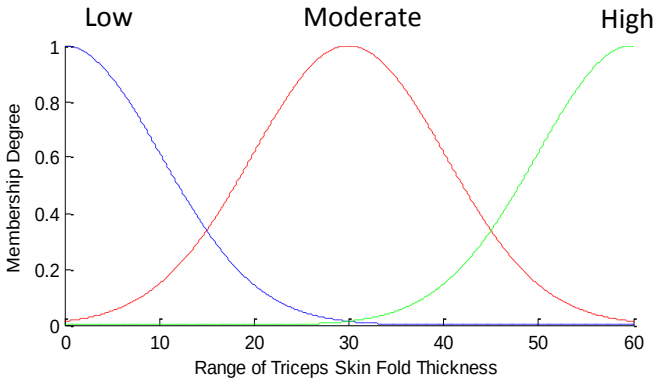


(a) Cluster 1 **(b)** Cluster 13

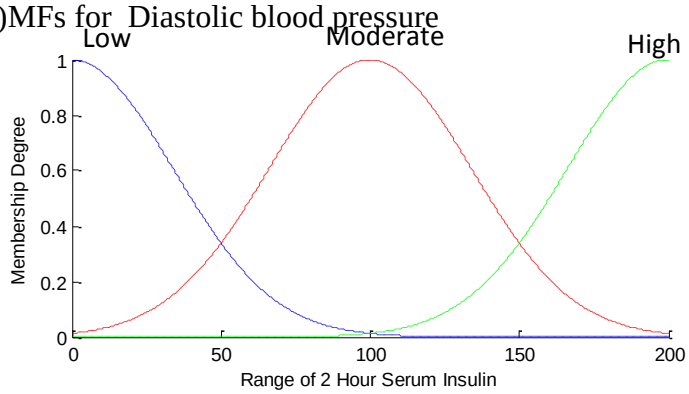
Fig. 5. Eigenvalues for (a) Cluster 1 and (b) Cluster 13 of PD



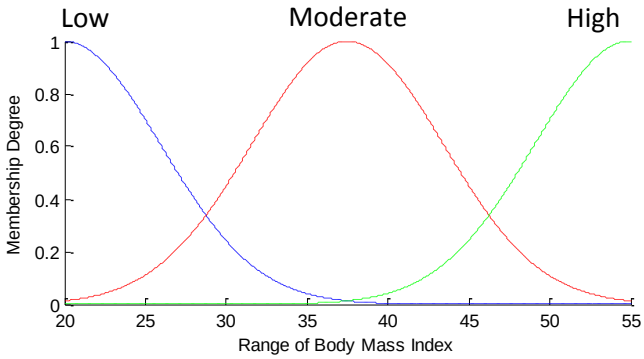
(a)MFs for Diastolic blood pressure



(b)MFs for Triceps skin fold thickness

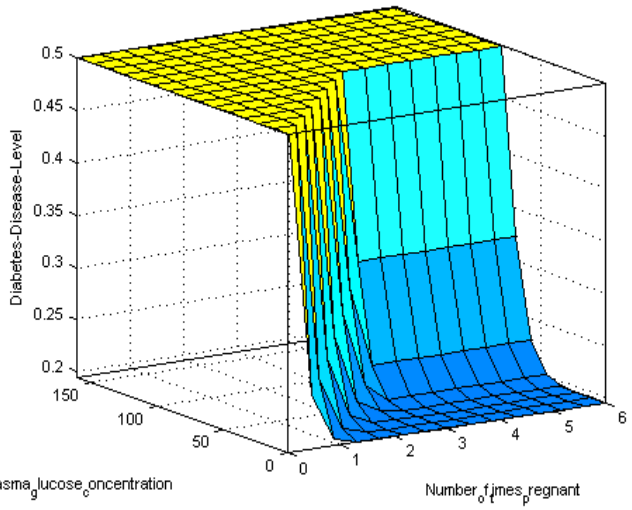


(c)MFs for 2 hour serum insulin

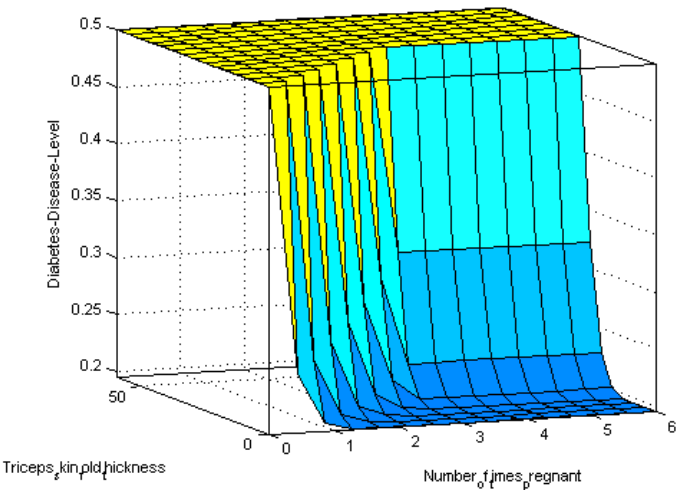


(d)MFs for Body mass index

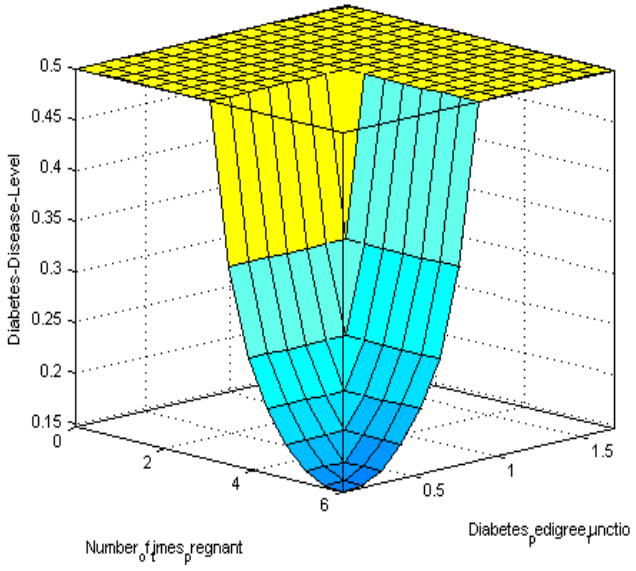
Fig. 6. MFs for PID



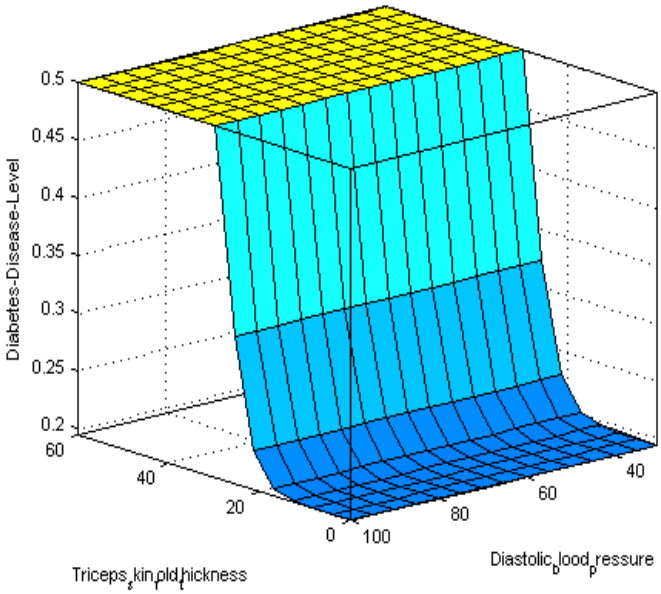
(a) Interdependency of Diabetes Presence and two parameters Number of times pregnant and Plasma glucose concentration in a 2 hour oral glucose tolerance test



(b) Interdependency of Diabetes Presence and two parameters Number of times pregnant and Triceps skin fold thickness

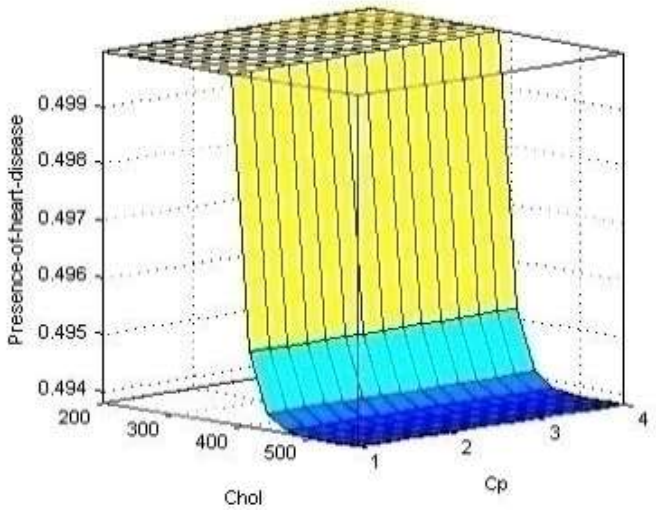


c) Interdependency of Diabetes Presence and two parameters Number of times pregnant and Diabetes pedigree function

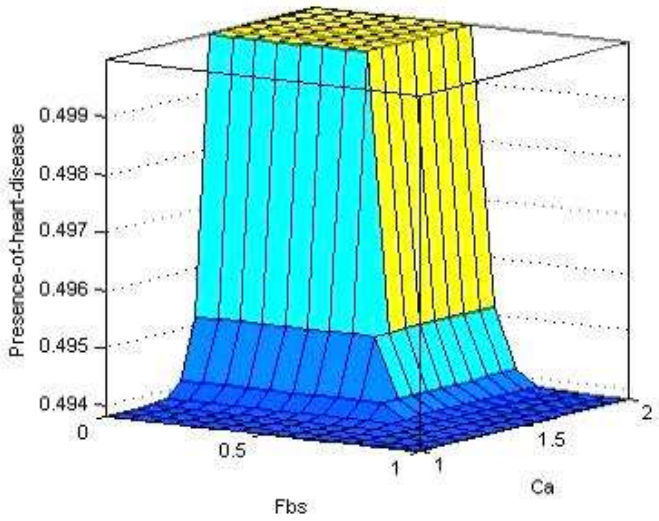


(d) Interdependency of Diabetes Presence and two parameters Triceps skin fold thickness and Diastolic blood pressure

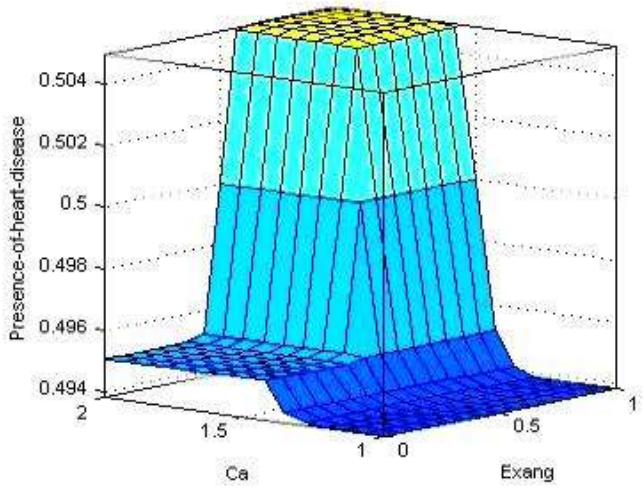
Fig. 7. Relationship between the inputs and output parameters of PID dataset



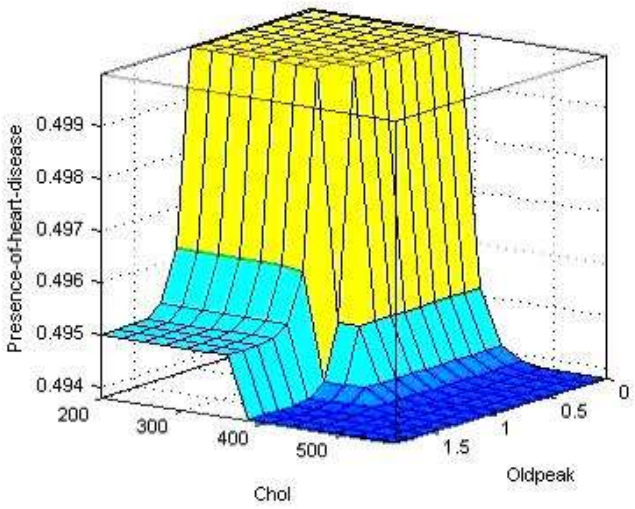
(a) Interdependency of Heart Disease Presence and two parameters Cp and Chol



(b) Interdependency of Heart Disease Presence and two parameters Fbs and Ca



(c) Interdependency of Heart Disease Presence and two parameters Ca and Exang



(d) Interdependency of Heart Disease Presence and two parameters Chol and Oldpeak

Fig. 8. Relationship between the inputs and output parameters of Cleveland dataset

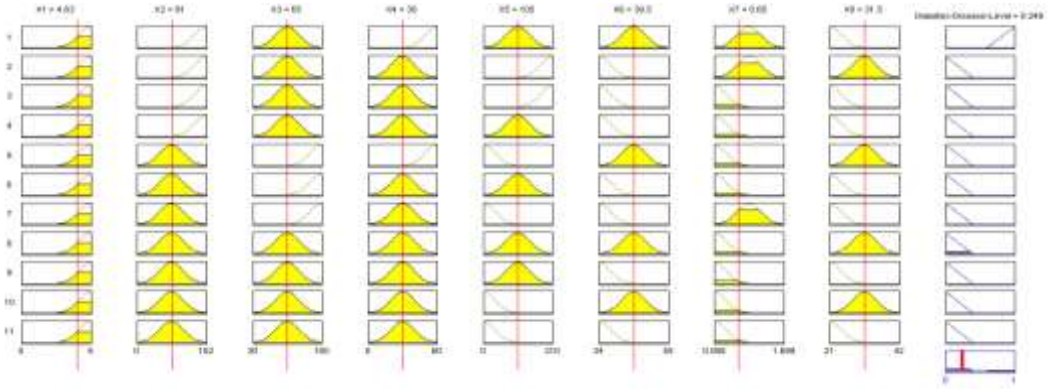
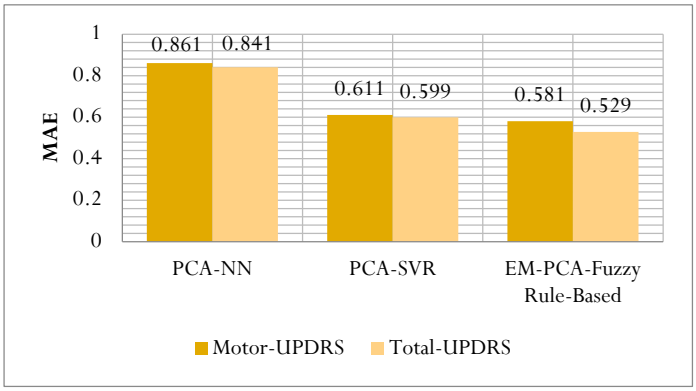


Fig. 9. Diabetes disease presence based on PID parameters



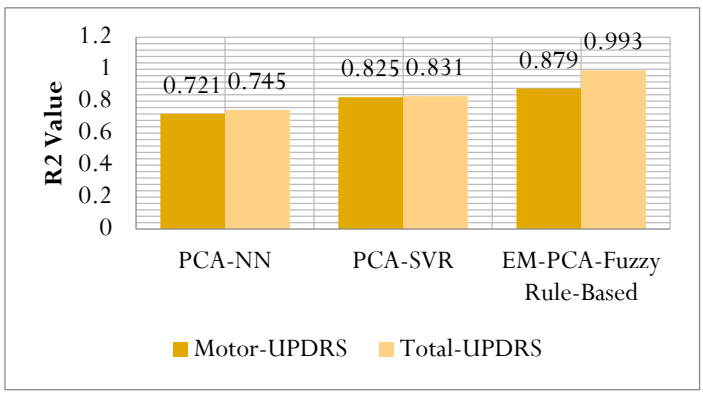


Fig. 10. MAE and R^2 values for Motor-UPDRS and Total-UPDRS predictions

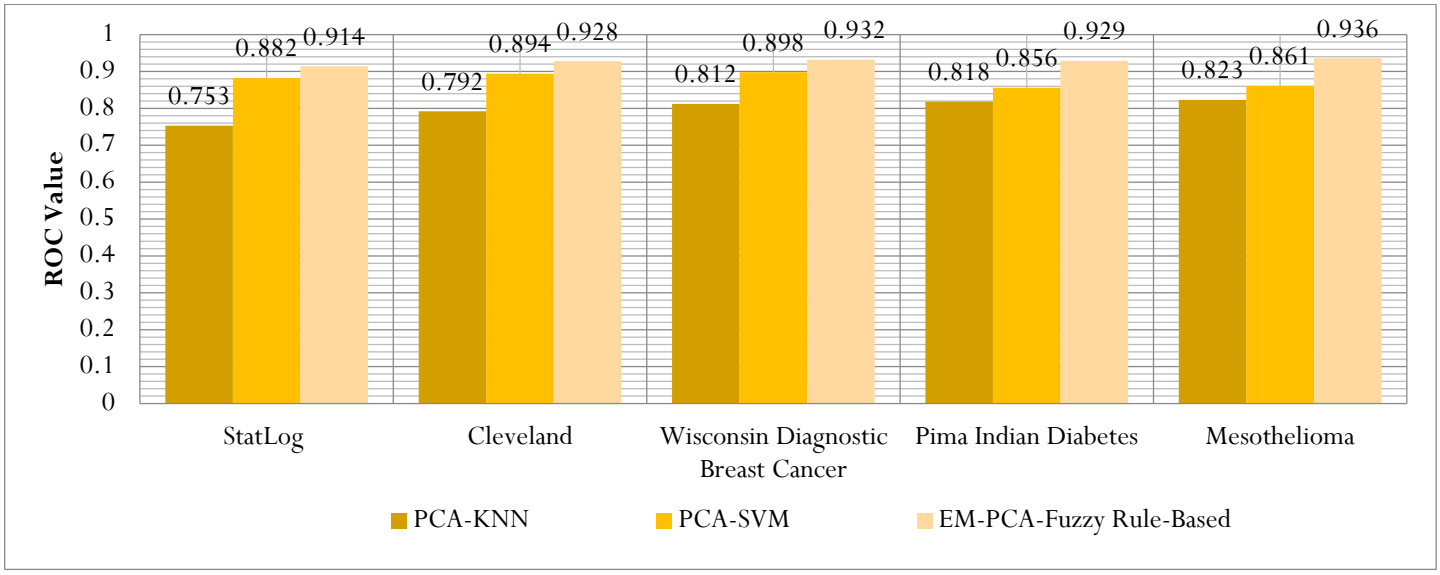


Fig. 11. MAE and R^2 values for Motor-UPDRS and Total-UPDRS predictions

Table 1

Previous studies on diseases classification

Disease	Author	Techniques														
		SVM	KNN	NN	ANFIS	FL	KM	GP	EM	PCA	RF	LDA	DT	AR	PSO	NB
Diabetes	Polat et al. (2008)	*														
	Aslam et al. (2013)	*	*					*								
	Kahramanli and Allahverdi (2008)			*		*										
	Erkaymaz and Ozer (2016)			*												
	Ganji and Abadeh (2011)					*										
	Dogantekin et al. (2010)				*							*				
	Temurtas et al. (2009)			*												
	Çalişir and Doğantekin (2011)	*										*				
	Hayashi and Yukita (2016)												*			
Breast Cancer	Şahan et al. (2007)					*										
	Polat and Güneş (2007)	*														
	Übeyli (2007)	*														
	Marcano-Cedeño (2011)			*												
	Zheng et al. (2014)	*														
	Chen (2014)													*		
	Bhardwaj and Tiwari (2015)			*												
	Onan (2015)		*			*										
	Karabatak (2015)															*

	Abdel-Zaher and Eldeib (2016)			*												*

Abbreviation used in this table: **SVM:** Support Vactor Machine, **KNN:** K-Nearest Neighbor**, NN:** Neural Network, **ANFIS:** Adaptive Neuro-Fuzzy Inference System, **FL:** Fuzzy Logic, **KM:** K-Means, **GP:** Genetic Programming, **EM:** Expectation Maximization, **PCA:** Principal Commponent Analysis, RF: Random Forest, **LDA:** Linear Discriminant Analysis, **DT:** Decision Tree, **AR:** Association Rule, **PSO:** Particle Swarm Optimization and **NB:** Naive Bayes

Table 1

Previous studies on diseases classification (Cont.)

Disease	**Author**	**Techniques**														
		SVM	**KNN**	**NN**	**ANFIS**	**FL**	**KM**	**GP**	**EM**	**PCA**	**RF**	**LDA**	**DT**	**AR**	**PSO**	**NB**
Parkinson	Guo et al. (2010)							*	*							
	Das (2010)			*												
	Bhattacharya and Bhatia (2010)	*														
	Åström and Koker (2011)					*										
	Li et al. (2011)					*										
	Ozcift (2012)	*														
	Polat (2012)		*			*	*									
	Eskidere et al. (2012)	*		*												
	Chen et al. (2013)	*	*			*				*						
	Babu and Suresh (2013)			*												

	Peterek et al. (2013)									*				
	Hariharan et al. (2014)			*				*	*		*			
	Froelich et al. (2015)											*		
	Buza and Varga (2016)			*										
	Naranjo et al. (2016)													*
	Al-Fatlawi et al. (2016)			*										
	Jain and Shetty (2016)		*	*									*	
	Behroozi and Sami (2016)	*	*											*
	Avci et al. (2016)						*							
Heart	Bhatia et al. (2008)	*												
	Allahverdi (2008)			*		*	*							
	Das et al. (2009)			*										
	Adeli et al. (2010)					*								
	Soni et al. (2011)												*	
	Gudadhe et al. (2010)	*		*										
	Ghumbre et al. (2011)	*		*										
	Anooj (2012)					*								
	Rout (2012)					*								
	Nahar et al. (2013)												*	
	Shilaskar and Ghatol (2013)	*												
	Shao et al. (2014)			*										
	Long et al. (2015)					*								
	Nguyen et al. (2015a)					*	*							
	Nguyen et al. (2015b)					*								

Kausar et al. (2016)	*	*	*

Abbreviation used in this table: **SVM:** Support Vactor Machine, **KNN:** K-Nearest Neighbor**, NN:** Neural Network, **ANFIS:** Adaptive Neuro-Fuzzy Inference System, **FL:** Fuzzy Logic, **KM:** K-Means, **GP:** Genetic Programming, **EM:** Expectation Maximization, **PCA:** Principal Commponent Analysis, RF: Random Forest, **LDA:** Linear Discriminant Analysis, **DT:** Decision Tree, **AR:** Association Rule, **PSO:** Particle Swarm Optimization and **NB:** Naive Bayes

Table 2

Result of PCA on clusters of Cleveland and StatLog

Cleveland	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Cluster 1	√	√	√	√	√	√	√	√	√ (96.98%)
Cluster 2	√	√	√	√	√	√	√ (97.61%)	-	-
Cluster 3	√	√	√	√	√	√ (95.43%)	-	-	-
Cluster 4	√	√	√	√	√	√	√ (97.21%)	-	-
Cluster 5	√	√	√	√	√	√ (95.66%)	-	-	-
Cluster 6	√	√	√	√	√	√ (97.82%)	-	-	-
Cluster 7	√	√	√	√	√	√	√	√ (96.41%)	-
StatLog									
Cluster 1	√	√	√	√	√	√	√	√	√ (95.51%)
Cluster 2	√	√	√	√	√	√	√ (94.26%)	-	-
Cluster 3	√	√	√	√	√	√	√	√ (93.85%)	-
Cluster 4	√	√	√	√	√	√	√ (94.67%)	-	-
Cluster 5	√	√	√	√	√	√	√	√ (97.21%)	-
Cluster 6	√	√	√	√	√	√	√	√ (96.52%)	-
Cluster 7	√	√	√	√	√	√	√ (93.56%)	-	-
Cluster 8	√	√	√	√	√	√	√ (92.58%)	-	-

Table 3

Result of PCA on clusters of Mesothelioma

Mesothelioma	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Cluster 1	√	√	√	√	√ (92.31%)	-	-	-
Cluster 2	√	√	√	√	√	√	√ (96.32%)	-
Cluster 3	√	√	√	√	√ (93.52%)	-	-	-

Mesothelioma	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Cluster 4	√	√	√	√	√	√ (94.28%)	-	-
Cluster 5	√	√	√	√	√	√	√	√ (96.23%)
Cluster 6	√	√	√	√	√	√ (94.48%)	-	-
Cluster 7	√	√	√	√	√	√	√ (97.31%)	-

Table 4

MAE and R^2 for the proposed method of predicting Motor-UPDRS and Total-UPDRS in Parkinson's telemonitoring dataset

Method	**Output**	**MAE**	**R^2**
PCA-NN	Motor-UPDRS	0.861	0.721
	Total-UPDRS	0.841	0.745
PCA-SVR	Motor-UPDRS	0.611	0.825
	Total-UPDRS	0.599	0.831
EM-PCA-Fuzzy Rule-Based	**Motor-UPDRS**	**0.581**	**0.879**
	Total-UPDRS	**0.529**	**0.993**

Table 5

Prediction accuracy on classification type datasets

Method	**StatLog**	**Cleveland**	**Wisconsin Diagnostic Breast Cancer**	**Pima Indian Diabetes**	**Mesothelioma**
PCA-KNN	0.753	0.792	0.812	0.818	0.823
PCA-SVM	0.882	0.894	0.898	0.856	0.861
EM-PCA-Fuzzy Rule-Based	0.914	0.928	0.932	0.929	0.936