International Journal of Engineering & Technology, 7 (1) (2018) 79-84

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i1.8652
Research paper

Comparative study on dimensionality reduction for disease diagnosis using fuzzy classifier

R.Sujatha [1] *, E.P.Ephzibah[1], Sree Dharinya [1], G. Uma Maheswari [1], V.Mareeswari [1], Vamsi Pamidimarri [2]

[1] School of Information Technology & Engineering, VIT , INIDA
[2] Canadian National (CN) Railways, Canada
**Corresponding author E-mail:*

Abstract

Machine learning is the worldwide recent research technique for various systems as they are intelligent enough to find the solution for classification and prediction problems. The proposed work is about a hybrid genetic fuzzy algorithm that performs an optimal search as well as classification upon uncertain data. The data which is uncertain is suitable for fuzzy classifiers to predict the disease. The hybrid genetic fuzzy system applied on the attributes selects relevant attributes. The selected attributes are fed into the fuzzy classifier. The fuzzy rules are again generated using genetic algorithms. This algorithm is applied on three of the important and bench marking data sets taken from the UCI machine learning repository. The heart disease, Wisconsin breast cancer and Pima Indian diabetes datasets produce classification accuracy as 89.65%, 99.5% and 88.93% respectively. In this article there is a comparative study on few of the feature selection and feature reduction techniques.

Keywords: *Feature Selection; Feature Extraction; Genetic Algorithms; Disease Diagnosis; Fuzzy Classifier.*

1. Introduction

Disease diagnosis is a task that determines the disease or condition that a person has in the form of signs and symptoms. It is of course a very challenging task for the medical practitioners. Some of the diagnostic methods are inspection, interrogation and palpation [1]. There are various tools and techniques available for disease diagnosis. Among these tools and techniques, the experts' knowledge about the field creates a strong impact of the diagnosis. Human knowledge plays a vital role in disease diagnosis. Experts are the health care professionals like the physicians, optometrists, chiropractors, dentists, podiatrists and so on. Computers aid these professionals for correct disease diagnosis. These professionals use their experience in recognizing a pattern of clinical characteristics [2].

Medical data mining and knowledge discovery is a relatively young and growing field that attracts many researchers [3]. Even though there are challenges that hinder the mining process, disease diagnosis is still an ever green field of study. The data obtained from the patients is the resource for medical data mining process. The critical task is to identify the pattern that is hidden within the data and help in differentiating the signs and symptoms between a healthy and a sick person.

Modern diagnostics tools and techniques that make use of the intelligent diagnosis have paved the way for effective healthcare systems. Intelligent machines help the pathologists to predict and classify the diseases. The soft computing techniques like the Genetic Algorithms (GA), neural networks, and Fuzzy Logic (FL) have been extensively used in the medical field for disease diagnosis[4] [5].

Genetic algorithm is a widely used technique for optimization. This algorithm helps to generate a solution for the problems whose solution space is very large. For instance, problems like traveling sales man problem, job scheduling problem, etc., [6] [7] [8] have been solved using genetic algorithms. Genetic algorithm is an iterative stochastic searching technique that works on the data called as population. The individuals into the population become better and more efficient in the successive iterations thereby providing an optimal solution to the problem [9].

Principal Component Analysis (PCA) is an important analysis for many numbers of problems in the medical data mining [10] [11]. It has been extracted from applied linear algebra. This analysis helps in bringing down the higher-dimensional data into lower dimensional data called principal components or artificial variables. The number of components extracted can be equal to the number of observed variables. The extracted components can further be used for classification of data. Even though the original data is reduced the variation present in the sample (Correlation between the variables) is not reduced. Linear Discriminant Analysis (LDA) [12] [13] is a classification method that performs pattern recognition using the discrimination between within and other class features.

Fuzzy logic is a system that allows partial truth to an object apart from the crisp value like either true or false. Partial truth can be in the range 0 and 1 thus permitting a human way of thinking and solving problems [14]. Fuzzy logic systems help to solve 90% of the problems in the control field. Fuzzy logic provides a fuzzy set which expresses the degree of membership of an element in the set. The element is capable of holding a membership value between 0 and 1. Fuzzy inference system has attracted many researchers for classification and prediction of diseases as it is capable of handling uncertain data [15] [16] [17].

Copyright © 2018 R.Sujatha et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

2. Realted works

In the paper on image analysis and disease diagnosis the authors have compressed the medical images and have used a hashing technique to match the incoming image with the other images available in the database, thereby diagnosing the disease [18]. The work on Alzheimer's disease detection performs feature extraction based on the anatomical differences. The methods namely vocal based morphometry and deformation based morphometry have been used. The Lattice computing K-NN classifier has been applied using 10 fold cross validation produces an accuracy of 84% [19]. Researchers have done dimensionality reduction using correlation feature selection as well as CHI squared selection for the lung cancer disease diagnosis. The experimental study shows about 16 classifier models in which the radial basis function network, Naïve bayes, Naïve bayes updateable, and multiplayer perceptron produces a maximum accuracy of 90.24% for the former and the multilayer perceptron produces an accuracy of 90.24% for the latter dimensionality reduction methods [20] .

A predictive model for the lung cancer based on structural and physiochemical properties of proteins using data mining models was carried on and the used gain ratio for ranking the relevant features and correction feature selection for subset selection producing a hybrid feature selection method and further classified using the Bayesian network classifier. The produced accuracy was 87.6% [21]. Principal components analysis as the feature extractor for diagnosing heart disease is one of the stages in the paper [22].

The features extracted were considered for classification using support vector machine (SVM). Their proposed method has given the classification accuracy over the range 91.7% to 99.8% for various types of heart-related diseases. The have taken the heart beat classification system for predicting the ECG signals [23]. They have proposed a 3 stage model that helps in predicting the presence of heart disease. They are: firstly, to de-noise the signal data using multi scale principal components analysis, secondly to extract the important subset of features using auto regressive modeling and thirdly to classify the data using various classifiers namely SVM and K-NN. The experimental results show that the model that has been developed using the multi scale principal components analysis and auto regressive modeling produced an accuracy of 99.93%.

The work is carried on chronic kidney data set considering the missing value attributes and classifiers applied [24]. The work has proven the effectiveness of computerized processing of ECG signals for heart disease diagnosis. They have used PCA for features extraction and SVM for classification [25]. The prediction accuracy is more promising and effective with the sensitivity and specificity as 99.93% and 100% respectively. 99.96% are the final accuracy of their proposed method. Fuzzy classifier implemented for the predicted of arrhythmias in patients with the help of ECG signal patterns is the proposed work. Genetic algorithm has also been implemented for better accuracy. The obtained accuracies with and without genetic algorithm are 98.67% and 93.34% respectively [26].

3. Methodology

3.1. Feature selection using genetic algorithm

Feature selection in the field of bioinformatics eliminates irrelevant features from the medical data and chooses the features that appropriately help for classification and prediction especially in gene selection, micro array analysis and disease diagnosis [27]. Genetic algorithms are the much prominently used evolutionary models that help in selecting the features that contribute more for the classification and prediction in intelligent systems. These algorithms find their role in learning and evolution combined to produce an optimal solution from a wide solution space. As genetic algorithm is an appropriate choice for feature selection the proposed work has taken into account the effective usage of features individually and also as groups. The choice of selecting the best rules is also handled using genetic algorithms. Irrespective of the domain genetic algorithms are capable of identifying important features.

Genetic algorithms involve five important components like:

- Evaluation function
- Selection operator
- Crossover operator
- Mutation operator
- Chromosome representation

The evaluation function for our approach is given in equation (1).

$$\text{Fitness function} = 100 - \left[\left(\sum_{i=1}^{n} S_i * W_i - \sum_{j=n+1}^{m} S_j * W_j\right) / TW * 100\right] \quad (1)$$

Where $\sum_{i=1}^{n} S_i * W_i$ the weighted sum of the match is scores of all the correct recognitions and $\sum_{j=n+1}^{m} S_j * W_j$ is the weighted sum of the match score of all the incorrect recognitions.

TW is the total weights of the samples.

The table (1) gives the values of the genetic operators used for the proposed work.

Table 1: Genetic Operators Value

S. No	Genetic Operator	Value
1	Selection	Roulette Wheel selection
2	Crossover	0.85
3	Mutation	0.025
4	Chromosome representation	Real numbers

3.2. Decision tree for feature selection

Decision trees are simple and easy to understand. The decision tree helps to select important and relevant features using the measures like entropy and information gain [28]. The entropy measure or the expected information based on the partitioning into subsets by A, an attribute in the dataset can be calculated using the equation (2).

$$E(A) = \sum_{j=1}^{v} \frac{S_{ij} + ... + S_{mj}}{s} I(S_{ij},..., S_{mj}). \quad (2)$$

Here in equation (2), E (A) is the entropy, term $\frac{S_{ij} + ... + S_{mj}}{s}$ acts as the weight of the jth subset and it is the number of samples in the subset divided by the total number of samples in the set S. The entropy values decide the accuracy and purity of partitioning the set into subsets. For a subset Sj equation (3) helps to find the expected information, I.

$$I(S_{ij},...,S_{mj}) = -\sum_{i=1}^{m} Pij \log_2 (Pij) \quad (3)$$

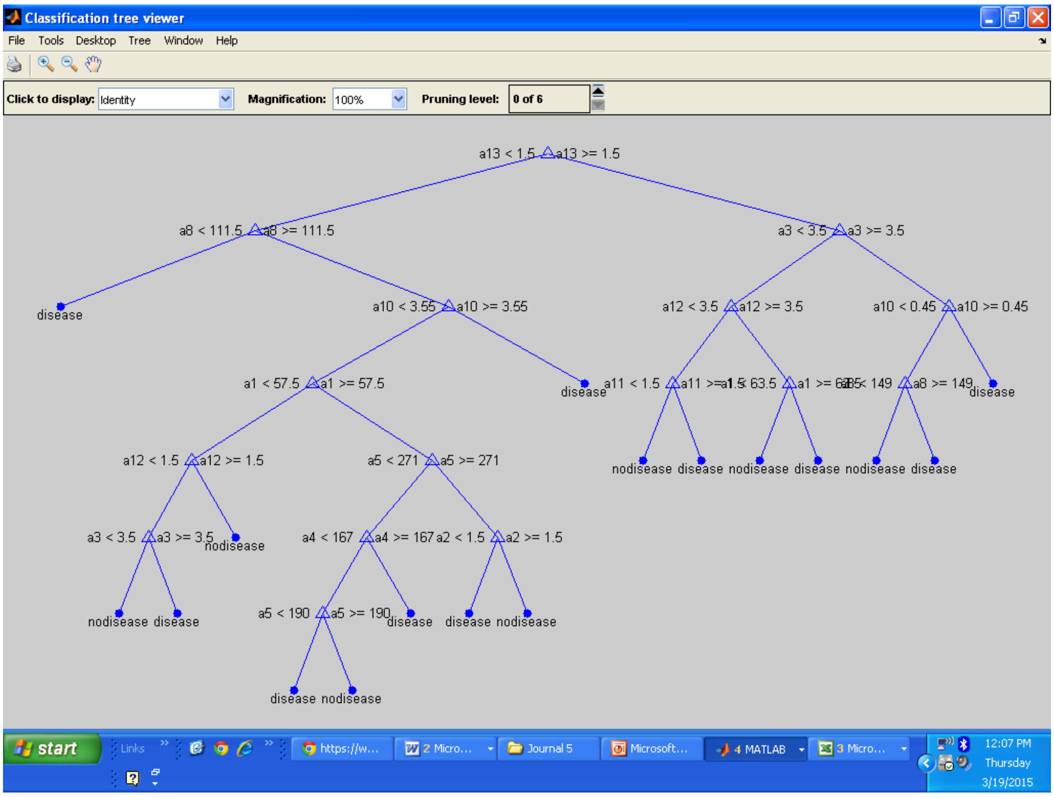
Where

$Pij = \frac{Sij}{|Sj|}$ is the probability that a sample Sj belongs to class Ci.

Thus the information gain of an attribute A can be evaluated using equation (4).

$$Gain(A) = I(S_{ij},..., S_{mj}) - E(A) \quad (4)$$

Based on the Gain value of each and every attribute the tree is split. The figure (1) represents the decision tree generated for the heart disease dataset.



Click to display: Identity Magnification: 100% Pruning level: 0 of 6

Fig. 1: Decision Tree – Heart Disease Data Set.

3.3. Dimensionality reduction using PCA and LDA

Principal component analysis and linear discriminant analysis are good ways of feature extraction techniques which can be used for pattern recognition, machine learning and statistics. The proposed work has taken both of these techniques separately first and together next for feature extraction.

3.4. Principal component analysis

The principal component analysis helps to convert the correlated variables into linearly uncorrelated ones. These uncorrelated variables are called as principal components. These components can be reduced based on the arrangement of the Eigen values and vectors. The heart disease dataset contains the data that has been projected using a box plot as in figure 2. Figure 3 projects the first and second principal components of the heart disease dataset.

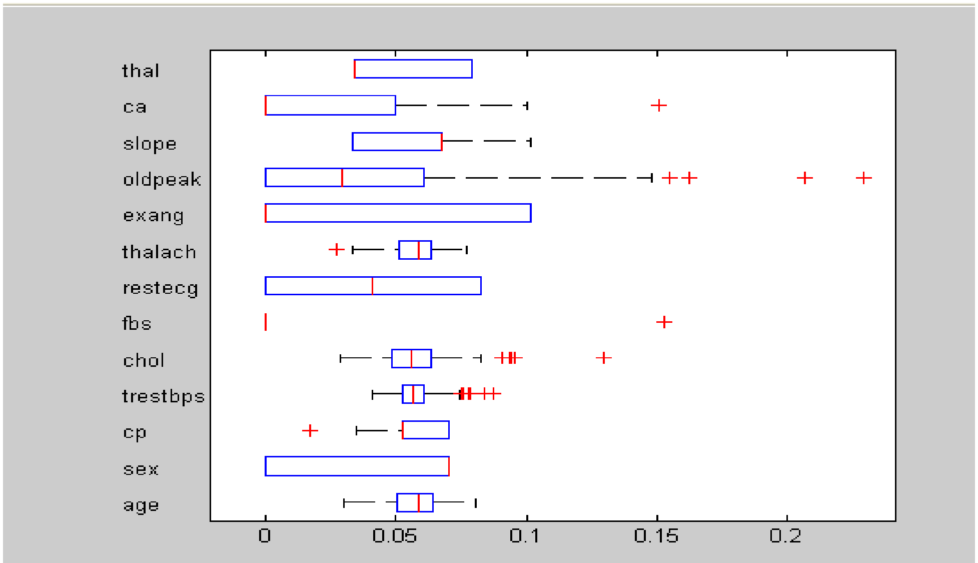


Fig. 2: Box Plot – Attributes.

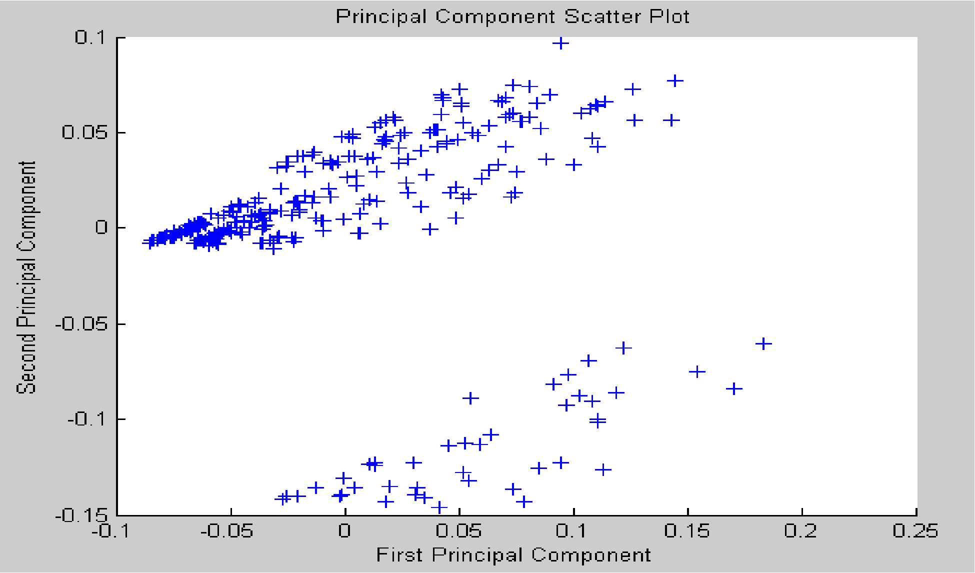


Fig. 3: Principal Components.

3.5. Linear discriminant analysis

LDA is a supervised classification model that performs the task to find the base vectors that contribute more towards the class label. A linear projection of the sample data gives the base vectors as they are orthogonal to each other. LDA when compared to other statistical methods proves to provide maximum separation between the features belonging to different classes thereby paves a way for the solution.

3.6. Classification using fuzzy classifier

Fuzzy classification involves fuzzification, fuzzy inference, fuzzy rule generation and finally defuzzification. The proposed work considers the fuzzy region as labels in the range from 3 to 11 with a step value of 2. Fuzzy inference engine with the knowledge base provides a correct mapping between the input and output. The knowledge base has been developed with the help of the data available. Every record is considered. When these records are analyzed using genetic algorithm a few of the records fall in the solution space thereby reducing the number of rules. The proposed work represented in figure 4.

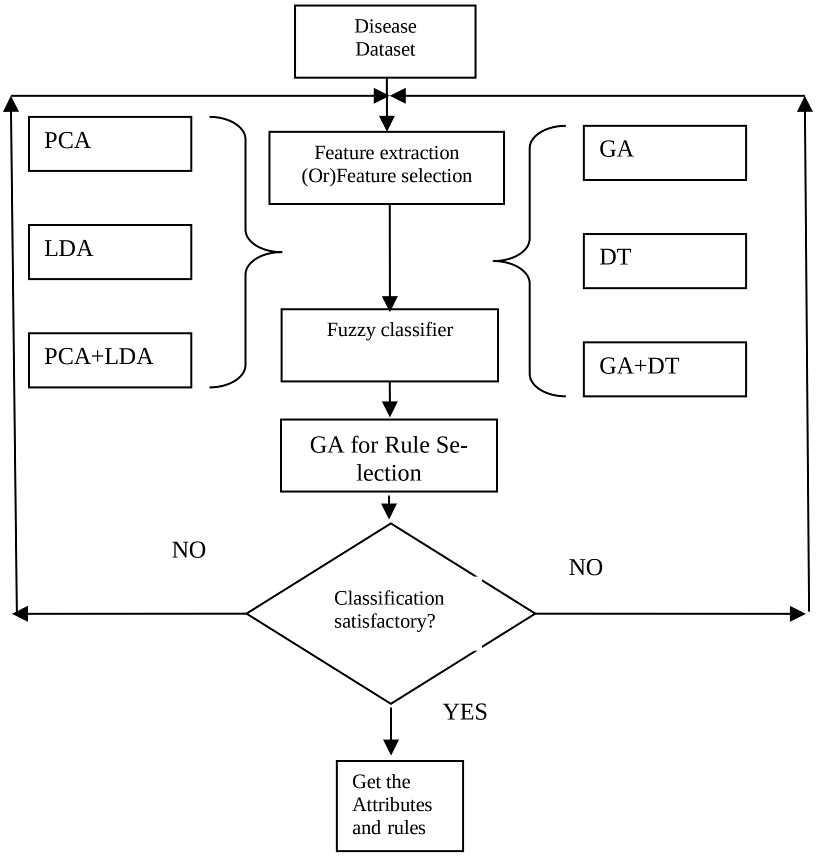

Disease Dataset
PCA
LDA
PCA+LDA
Feature extraction (Or)Feature selection
GA
DT
GA+DT
Fuzzy classifier
GA for Rule Selection
NO
NO
Classification satisfactory?
YES
Get the Attributes and rules

Fig. 4: Proposed Work.

4. Datasets

4.1. Heart disease dataset

In case of heart disease the features are the attributes in the dataset. The attributes provide values related to the patient's history like age, sex, cp-chest pain types, trestbps-resting blood pressure in mmHg, cholesterol in mg/dl, fbs-fasting blood sugar in mg/dl, restecg-resting ECG result, thalach- maximum heart rate achieved, exang-exercise induced angina, oldpeak, slope, ca- number of major vessels colored by fluoroscopy and thal.

4.2. Breast cancer dataset

Wisconsin breast cancer dataset taken from UCI machine learning repository is the second dataset considered. The set contains 699 instances with 16 instances containing missing values. As the missing entries are comparatively lesser in number they were eliminated in the beginning. Excluding the class label there are 9 attributes like clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli and mitoses. The values are in the range from 1 to 10.

4.3. PIMA Indian diabetes dataset

Pima Indian diabetes dataset is the third dataset considered for this work. There are 768 instances available in the dataset. The total number of input attributes is 8. The attribute headings are number of times pregnant, plasma glucose concentration, diastolic blood pressure in mm/Hg, triceps skin fold thickness in mm, 2-Hour serum insulin in U/ml, body mass index in kg/m, diabetes pedigree function and age in years.

4.4. Experimental results with fuzzy classifier

The experimental results in terms of accuracy are tabulated in table 2. FE1 and FE2 are Feature Extraction 1 and Feature Extraction 2 respectively. FS1 and FS2 are Feature Selection1 and Feature Selection 2 respectively. The proposed work is represented as GA^2+Fuzzy. GA^2 means GA applied twice in the work. Table 3 shows the comparative analysis with contemporary work and proposed model shows optimized accuracy.

Table 2: Accuracy of Proposed Work

S.No	Disease Dataset	PCA (FE1)	LDA (FE2)	(FE1+FE2)	GA (FS1)	DT (FS2)	(FS1+FS2)	GA^2+Fuzzy
1	Cleveland Heart Disease data	88.50	86.50	89.00	88.00	89.00	89.45	89.65
2	Wisconsin Breast Cancer Data	96.45	96.00	98.75	96.66	97.00	98.88	99.50
3	Pima Indian Diabetes data	84.45	83.27	85.75	86.50	87.90	88.00	88.93

Table 3: Comparative Analysis

S No	Disease	Paper &method &year	Accuracy
1	Heart disease	MARS-LR, RS-LR[4]	83.93
2	Heart disease	Neural network ensemble [6]	89.01
3	Heart disease	PCA+ANFIS[10]	96.00
4	Heart disease	GA^2+Fuzzy*	89.65
5	Breast Cancer	RS –SVM[11]	96.87
6	Breast Cancer	AR-NN[13]	95.60
7	Breast Cancer	AMMLP[16]	99.26
8	Breast Cancer	GA^2+Fuzzy*	99. 50
9	Breast Cancer	SMO[18]	96.14
10	Diabetes	LDA-ANFIS [7]	84.61
11	Diabetes	GA^2+Fuzzy*	88.93

* Results of Proposed Work.

5. Conclusion

The proposed work compares the dimensionality reduction in terms of feature extraction and feature selection. Feature extraction has been performed using principal component analysis and linear discriminant analysis. The feature selection has been implemented using genetic algorithm and decision tree method. The accuracy of the system has been improved when feature selection and rule selection using genetic algorithm was implemented. It is observed that feature selection using genetic algorithm with the fuzzy classifier produces better results for the diseases considered throughout this work. The proposed work is named as a hybrid technique that uses genetic algorithm for two tasks. Primarily used as an attribute selector that filters the relevant and efficient features that help for better accuracy in classification. It is used as a rule selector that filters the rules required for classification. Thus the proposed work is a useful resource for medical mining, especially for disease diagnosis.

References

[1] Jingfeng, C, Medicine in China. Encyclopedia of the History of Science, Technology, and Medicine in Non-Western Cultures ,(2008), 1529–1534

[2] Thompson, Carl, and Dawn Dowding. Essential Decision Making and Clinical Judgement for Nurses E-Book. Elsevier Health Sciences, 2009.

[3] Cios, Krzysztof J., et al. Data mining: a knowledge discovery approach. Springer Science & Business Media, 2007.

[4] Ahmad, Fadzil, et al., Intelligent breast cancer diagnosis using hybrid GA-ANN, Computational Intelligence, Communication Systems and Networks (CICSyN), 2013 Fifth International Conference on. IEEE, 2013.

[5] Jaganathan, P., and R. Kuppuchamy, A threshold fuzzy entropy based feature selection for medical database classification, Computers in Biology and Medicine 43, 12, (2013), 2222-2229.

[6] Wang, Peng, Cesar Sanin, and Edward Szczerbicki, Evolutionary algorithm and decisional DNA for multiple travelling salesman problem, Neurocomputing, 150, (2015), 50-57. https://doi.org/10.1016/j.neucom.2014.01.075.

[7] Pham, Dinh Thanh, and Thi Thanh Binh Huynh, An Effective Combination of Genetic Algorithms and the Variable Neighbor-

hood Search for Solving Travelling Salesman Problem, Technologies and Applications of Artificial Intelligence (TAAI), 2015 Conference on. IEEE, 2015.

[8] Shen, Zhonghua, Keith J. Burnham, and Leonid Smalov, Optimised job-shop scheduling via genetic algorithm for a manufacturing production system, Progress in Systems Engineering. Springer, Cham, (2015), 89-92.

[9] Sivanandam, S. N., and S. N. Deepa. Introduction to genetic algorithms. Springer Science & Business Media, 2007.

[10] Veenstra, Michelle Anne, et al, Raman spectroscopy in the diagnosis of ulcerative colitis, European Journal of Pediatric Surgery 25, 01 (2015), 56-59.

[11] Hariharan, Muthusamy, Kemal Polat, and Ravindran Sindhu, A new hybrid intelligent system for accurate detection of Parkinson's disease, Computer methods and programs in biomedicine, 113,3, (2014), 904-913.

[12] Giri, Donna, et al., Automated diagnosis of coronary artery disease affected patients using LDA, PCA, ICA and discrete wavelet transform, Knowledge-Based Systems, 37, (2013), 274-282. https://doi.org/10.1016/j.knosys.2012.08.011.

[13] Çalişir, Duygu, and Esin Doğantekin, An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine Classifier, Expert Systems with Applications, 38,7, (2011), 8311-8315.

[14] Alavala, Chennakesava R. Fuzzy logic and neural networks: basic concepts & application. New Age International, 2008.

[15] Santhi, D., D. Manimegalai, and S. Karkuzhali. Diagnosis of diabetic retinopathy by exudates detection using clustering techniques, Biomedical Engineering: Applications, Basis and Communications, 26, 06, (2014), 1450077.

[16] Assadi, Ava, and Saman Harati Zade, UGA: A new genetic algorithm-based classification method for uncertain data, Mid-Est J Scient Res 20.10, (2014), 1207-1212.

[17] Shamshirband, Shahaboddin, et al, Tuberculosis disease diagnosis using artificial immune recognition system, International journal of medical sciences 11, 5 (2014), 508.

[18] Zhang, Xiaofan, et al., towards large-scale histopathological image analysis: Hashing-based image retrieval, IEEE Transactions on Medical Imaging, 34, 2, (2015), 496-506.

[19] Papakostas, George A., et al., A lattice computing approach to Alzheimer’s disease computer assisted diagnosis based on MRI data, Neurocomputing 150, (2015), 37-42. https://doi.org/10.1016/j.neucom.2014.02.076.

[20] Balachandran, K., and R. Anitha, Dimensionality reduction based on the classifier models: Performance Issues in the prediction of Lung cancer, Software Engineering (CONSEG), 2012 CSI Sixth International Conference on. IEEE, 2012.

[21] Ramani, R. Geetha, and Shomona Gracia Jacob, Improved classification of lung cancer tumors based on structural and physicochemical properties of proteins using data mining models, PloS one 8,3, (2013), e58772.

[22] Sun, Shuping. An innovative intelligent system based on automatic diagnostic feature extraction for diagnosing heart diseases, Knowledge-Based Systems 75, (2015), 224-238. https://doi.org/10.1016/j.knosys.2014.12.001.

[23] Alickovic, Emina, and Abdulhamit Subasi, Effect of multiscale PCA de-noising in ECG beat classification for diagnosis of cardiovascular diseases, Circuits, Systems, and Signal Processing 34, 2, (2015), 513-533.

[24] Sujatha .R Ezhilmaran. Performance analysis of data mining classification techniques for chronic kidney disease. International Journal of Pharmacy and Technology, (2016), 8, 2, 13032-13037.

[25] Zhi, Koh Yi, Oliver Faust, and Wenwei Yu, Wavelet based machine learning techniques for electrocardiogram signal analysis, Journal of Medical Imaging and Health Informatics, 4, 5, (2014), 737-742.

[26] Vafaie, M. H., M. Ataei, and Hamid R. Koofigar, Heart diseases prediction based on ECG signals’ classification using a genetic-fuzzy system and dynamical model of ECG signals, Biomedical Signal Processing and Control 14, (2014), 291-296. https://doi.org/10.1016/j.bspc.2014.08.010.

[27] I.Guyon, J.Weston, S.Barnhill, V.Vapnik, Gene Selection for cancer classification using support vector machine, Machine Learning, 2002, 389-422. https://doi.org/10.1023/A:1012487302797.

[28] Han, Jiawei, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. Elsevier, 2011.