**T**urkish **J**ournal of **M**athematics and **C**omputer **S**cience
TJMCS

http://tjmcs.matder.org.tr/

@TJMCS
© MatDer
http://www.matder.org.tr/

# A Comparative Study of Heart Disease Prediction Based on Principal Component Analysis and Clustering Methods

Negar Ziasabounchi and Iman N. Askerzade

Abstract. In this article, we propose clustering approach based on Principal Component Analysis (PCA) to diagnosis of heart disease patients. At the first stage, the original dataset is reduced using PCA reduction method. Then, at the second stage, reduced dataset is applied to clustering methods which is based on fuzzy C-means and K-means algorithms. These algorithms are implemented and tested on a Cleveland heart disease dataset. We compared the clustering results with and without PCA. The results are suggesting that the combination of clustering algorithms and PCA was the most effective at heart disease diagnosis.

Keywords: Data Clustering; K-means, Fuzzy C-means, Principal Component Analysis

latexsym

## 1. Introduction

Based on the World Health Organization reports, stated in [1], Cardiovascular disease (CVD) also known as heart disease is the number one cause of death globally and the rate of death annually from CVDs is more than any other cause. CVD is caused by disorders of the heart and blood vessels and comprises coronary heart disease (heart attacks), cerebrovascular disease (stroke), rheumatic heart disease and etc. The most common cause of heart disease is coronary heart disease and most heart attacks occur as a result of this disease. Heart attacks are mainly caused when one of the arteries that supplies blood flow to the heart (the coronary arteries) becomes blocked, therefore the blockage prevents oxygenated blood from flowing to the heart. If blood flow is not restored quickly, the section of heart muscle begins to die. A major problem in medical science is correct diagnosis of disease. Medical diagnosis is considered as a prominent yet complicated task that needs to be executed precisely and efficiently [2]. Data mining in health care has become increasingly popular because it can improve patient care by early detection of diseases, supports helping care providers for treatment programs and reduces the cost of health care .

In recent years, several machine learning techniques and algorithms have been used in prediction of medical diagnosis. The two essential algorithms that are used in data mining are clustering and classification algorithms which classification is used as a supervised learning method and clustering for unsupervised learning [3]. Clustering is a process for classifying data points by partitioning the similar kind in to respective groups. As noted in [4] hierarchical and K-means clustering methods are approved and widely used data mining clustering techniques that can be used on high dimensional datasets. Fuzzy C-means and K-means methods are both unsupervised clustering methods. When the literatures related with fuzzy and clustering methodologies are examined, it can be seen that there are diverse types of studies [5, 6, 7, 8, 9, 10, 11, 12]. In [13], Sundar et al. introduced predictive model for heart disease diagnosis based on K-means clustering technique. Likewise, In [9] Chitra and Seenivasagam carried out a study about heart attack prediction by using fuzzy C-means classifier which gives a classification accuracy of 92%. On the other hand, Yin et al. [14] has Compared the performance of fuzzy C-means and K-means cluster analysis for arterial input function (AIF) detection. One of the problems that high dimensional dataset may arise for machine learning is the risk of overfitting. The other problem which can also lead to reduction of prediction accuracy is the redundancy of the attributes [15]. Hence, for overcoming these problems, PCA can be used to reduce the size of a dataset. PCA is a widely used statistical technique for unsupervised dimension reduction and the main basis of PCA is to pick up the dimensions with the largest variances [5]. When the literature is investigated, it can be seen that many studies have shown the benefits of using PCA via clustering methods [16, 17, 18].

The purpose of this study is to build an intelligent diagnosis system that could accurately classify heart disease patients in to normal or abnormal groups with clustering methods**.** In the set of experiments, to improve the performance of clustering algorithms, we used the combination of clustering algorithms and PCA method. Two Fuzzy C-means and K-means clustering methods have been implemented via PCA for medical diagnosis of heart disease. The dataset that are used in this study is Cleveland Clinic Foundation heart disease dataset which has been obtained from the University of California (UCI) machine learning data repository. This dataset contains 303 samples. The followed analysis methodology and the experiment results are given in the corresponding section. The results indicate that, the use of PCA in combination with clustering methods can stand as a good predicting mechanism for heart disease prediction. Remaining of this paper is organized as follows: The theoretical background of clustering methods and PCA is demonstrated in section 2. Section 3 presents the information of dataset. The experiments and obtained results are given in section 4 and section 5 presents the conclusion of this study.

## 2. Background

In this section a detailed discussion of two clustering techniques and PCA are presented.

2.1. **Clustering.** As it mentioned above, clustering classifies data instances in to subgroups to characterizes the population being sampled [3]. In other words, the goal of clustering is to discover and organize structure in datasets by recognizing and quantifying similarities between data patterns [19,20].

2.1.1. *K-means Clustering Algorithm.* One of the commonly used partitioning based clustering technique is K-means that tries to find a number of clusters ($k$) which defined by user and these clusters are represented by their centroids [18]. The K-means algorithm seeks a partition of $N$ (number of instances) in to $k$(number of clusters) in a way that sum of squared errors are minimized. At first initial set of cluster centers are calculated randomly. each instance is allocated to its nearest cluster center based on the Euclidean distance between the two and then the cluster centers are recalculated [3]. The processes of reassigning the input vectors and the update of the cluster centers is repeated until no more change in the value of any of the cluster centers.

The center of each cluster is calculated as the mean of all the instances belonging to that cluster:

$$\mu_k = \frac{1}{N_k} \sum_{q=1}^{N} {}_k x_q \tag{2.1}$$

Where $N_k$ is the number of instances belonging to cluster $k$ and $\mu_k$ is the mean of the cluster $k$.

The steps of the K-means algorithm are written below:

Input: $S$ (instance set), $K$ (number of cluster)

Output: clusters

Step 1: Initialize $K$ cluster centers.

Step 2: Assign instances to the closest cluster center.

Step 3: Update cluster centers in a way that each cluster using the mean of the instances allotted to that cluster.

Step 4: repeat steps 2 and 3 until the value of the means stabilizes.

2.1.2. *Fuzzy C-means Clustering Algorithm (FCM).* FCM algorithm is one of the well-known and commonly used fuzzy clustering methods. The membership value concept allows FCM algorithm to increase the number of belongings of data points from one to more than one. In other words a data point can partially belong to a cluster and the components of partition matrix $U$, varying within the interval [0, 1][21, 22].The basic idea of FCM algorithm is to find the minimization of objective function. This algorithm is an overlapping data clustering technique where in each data point, $X = \{x_1, x_2\ , \ \dots\ ,x_k\ \}$ belongs *to* a cluster $i$ to some degree specified by a membership grade, $u_{ij}$.

$$J_m\left(U, x_1, \dots, x_k\right) = \sum_{i=1}^{k} \sum_{j=1}^{c} u_{ij}^{m} ||x_i - c_j||^2 , \qquad 1 \leq m \leq \infty \tag{2.2}$$

where $m$ is any real number greater than 1, $u_{ij}$ is the degree of membership of $x_i$ between 0 and 1 in the cluster $j$, $x_i$ is the $i$ th of d-dimensional measured data, $c_j$ is the d-dimension center of the cluster, and $|| * ||$ is the Euclidean distance between the $i$th cluster center and the $j$th data point. Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership $u_{ij}$ and the cluster centers $c_j$ by equation 3 and 4 [9]

$$u_{ij} = \frac{1}{\sum\limits_{k=1}^{c} \left\{ ||x_i - c_j||^{2}/m - 1 \Big/ ||x_i - c_k|| \right\}} \tag{2.3}$$

$$c_j = \sum\nolimits_{i=1}^{N} u_{ij}^{m}.x_i \Big/ \sum\nolimits_{i=1}^{N} u_{ij}^{m} \tag{2.4}$$

According to equation 5 the iteration will stop.

$$max_{ij} \left\{ u_{ij}^{(k+1)} - u_{ij}^{(k)} \right\} < \varepsilon \tag{2.5}$$

where $\varepsilon$ is a termination criterion between 0 and 1 and $k$are the iteration steps.

The algorithm is composed of the following steps:

Step 1: Choose number of clusters

Step 2: Initialize the membership matrix$U$ with random membership values between 0 and 1.

Step 3: At $k$ step, calculate $c$ fuzzy cluster centers using $c_j$and $U^k$.

Step 4: Update $U^{k+1} = U^k$

Step 5: Repeat steps 3 and 4 until $\varepsilon$ is terminated according to equation 5.

2.1.3. *Principal Component Analysis (PCA).* The main idea of PCA is to reduce the size of a dataset by retaining maximum of information about original dataset. PCA includes an arithmetical procedure that converts a number of correlated variables into a smaller number of uncorrelated variables called principal components (PCs). In other words the problem can be stated as follows: given the $d$-dimensional random variable $x = (x_1, ..., x_d)^T$, find a lower dimensional representation of it , $y = (y_1, ..., y_D)^T$ with $D<d$.

The first principal component has the highest possible variance, and each of the succeeding components has the highest possible variance under the restriction that it has to be orthogonal to the previous component [23]. PCA mostly is used as a tool in exploratory data analysis and it can improve the predictive performance of some machine learning methods. The derivation and properties of PCs are based on the eigen vectors and eigen values of the covariance matrix [4]. In [23] the calculation of eigen vectors and values have been described in detail.

## 3. Dataset

The Cleveland heart diseases dataset was taken from the UCI machine learning repository.

This dataset contains 303 samples taken from healthy and unhealthy persons. The number of healthy and unhealthy samples in the dataset is 164 and 139 respectively. For each samples, the number of attributes is 13. These attributes are *age, sex, chest pain type, resting blood pressure, cholesterol, resting blood sugar, resting electrocardiographic, maximum heart rate, exercise, old peak slope, number of major vessels, thalium scan* Information about the attributes can be found in Table 1.

**Table 1. Information about the input variables**

| No | Attribute name | Min | Max | Description |
|---|---|---|---|---|
| 1 | age | 29 | 77 | age in years |
| 2 | sex | 0 | 1 | sex |
| 3 | Cp | 1 | 4 | chest pain tip (4 type) |
| 4 | Trestbps | 94 | 200 | resting blood pressure |
| 5 | Chol | 126 | 564 | cholesterol |
| 6 | Fbs | 0 | 1 | resting blood sugar (0=F, 1=T) is T if fbs > 120 |
| 7 | Restecg | 0 | 2 | resting electrocardiographic(ECG) |
| 8 | Thalach | 71 | 202 | maximum heart rate |
| 9 | Exang | 0 | 1 | exercise induced angina |
| 10 | Old peak | 0 | 6.2 | ST depression induced by exercise relative to rest |
| 11 | Stslope | 1 | 3 | the slope of the peak exercise ST segment |
| 12 | Vessels | 0 | 3 | major vessels (0-3) colored by flourosopy |
| 13 | Thal | 3 | 7 | Thalium scan |

## 4. Experiments and Results

4.1. **Proposed Method.** In this study to improve the performance of clustering methods, we apply Kmeans and FCM clustering methods in the PCA subspace. To reduce the size of database, PCA chooses the dimensions with the largest variances We used MATLAB software to compute principal components and clusters. In our proposed method, first PCA was applied to reduce the dimension of the Cleveland heart disease dataset then K-means and FCM clustering methods were applied to reduced dataset.

The steps of PCA and clustering methods are as follows:

1. Create the input matrix of dataset.

2. Standardize the dataset by subtracting the sample mean from each observation, then dividing by the sample standard deviation, which is given by:

*N = (A - repmat(AMean,[n 1])) ./ repmat(AStd,[n 1])* (6)

3.Calculate the covariance matrix of $N$and find the eigenvalues and eignvectors of the sample covariance matrix.

$[VD]$ = *eig(cov(N))* (7)

Which matrix $V$and $D$include the coefficients for the principal components and variance of the principal components respectively.

4 To calculate the principal components, multiply the standardized data by the principal component coefficients.

5. After obtaining principal components, corresponding variances, percentage of variances and the cumulative variences are calculated to eliminate weaker components and choose the new PCs with largest variances.

6. Create the transformation matrix $W$consisting of those new PCs.

7. Find the reduced dataset $Y$ in a new coordinate axis by applying $W$ to $N$

8. Partition Y into k clusters with Kmeans and FCM clustering algorithms.

4.2. **Performances Evaluation.** Some statistical measures like sensitivity, specificity and accuracy were used to evaluate the performance of the proposed method to indicate efficiency and reliability of the test. Sensitivity is used to find out the

**Table 2. Performance Results Comparision**

| Method | Accuracy rate(%) | RMSE | Time(sec) |
|---|---|---|---|
| K-means | 81.0% | 0.45 | 0.28 |
| FCM | 80.0% | 0.46 | 0,45 |
| K-means via PCA | 87.0% | 0.40 | 0.1 |
| FCM via PCA | 82.0% | 0.44 | 0.32 |

efficiency of test in detecting a positive disease Specificity measures how likely patients without disease can be correctly ruled out [24]. Accuracy of a diagnostic test can be determined from correct classified instances divided by total number of instances. Sensitivity, specificity and accuracy are described in terms of *TP, TN, FN.* The diagnosis is true positive (*TP*) when diseases is present in a patient. The diagnosis is true negative (*TN*) when diseases is absent. Both of the *TP* and *TN* are correct classifications. However, no medical test is hundred percent correct. For instance false positive (*FP*) result is the kind of diagnosis that shows the existence of disease in a patient who actually has no such disease and false negative (*FN*) diagnosis indicates the lack of disease for a patient with disease. Both FP and FN are incorrect classifications.

Sensitivity = $TP/(TP + FN)$ (8)

Specificity = $TN/(TN + FP)$ (9)

Accuracy = $(TN + TP)/(TN+TP+FN+FP)$ (10)

4.3. **Analysis of Results.** By applying PCA procedure, we reduced size of a Cleveland dataset then K-means and FCM clustering algorithms were used to classify the data as healthy or unhealthy. K-means and FCM algorithms initialize the cluster centers randomly Therefore, performances of both clustering methods are related to initial cluster centres In this case to achieve better results, several runs of algorithms are suggested. In this study, for system validation we used 5 cross validation to estimating performance of our proposed method. We also compared achieved results with the results without PCA method

Table 2 shows the results of these clustering methods. The first column represent the classification accuracy of each model. After determining the cluster centers, the evaluation data vectors are assigned to their respective clusters according to the distance between each vector and each of the cluster centers [25]. An error measure is then calculated. As seen from second column of Table 2, we used the root mean square error (RMSE) for this purpose. According to the results K-means via PCA clustering produces fairly higher accuracy and lower RMSE than the other techniques, and requires less computation time.

In Table 3 the performance of the proposed classifier methods are compared with the other classifier methods without using PCA Here performance metrics were calculated with the aid of equ 8, 9 and 10 Each instance is assigned to one of these two classes: healthy (normal) or unhealthy (abnormal). When the results in Table 3 are evaluated, it can be seen that using data clustering methods on low dimensionality dataset, produces quite high classification accuracy. Therefore, the use of PCA in

**Table 3. Performance Results**

| Method | Type | Patients | Healthy | Unhealthy | Sensitivity | Specificity | Accuracy |
|---|---|---|---|---|---|---|---|
| K-means | Healthy | 164 | 132 | 32 | 82.0% | 80.0% | 81.0% |
| K-means | Unhealthy | 139 | 25 | 114 | 82.0% | 80.0% | 81.0% |
| FCM | Healthy | 164 | 130 | 34 | 82.0% | 79.0% | 80.0% |
| FCM | Unhealthy | 139 | 25 | 114 | 82.0% | 79.0% | 80.0% |
| K-means via PCA | Healthy | 164 | 151 | 13 | 81.0% | 92.0% | 87.0% |
| K-means via PCA | Unhealthy | 139 | 26 | 113 | 81.0% | 92.0% | 87.0% |
| FCM via PCA | Healthy | 164 | 135 | 29 | 80.0% | 82.0% | 82.0% |
| FCM via PCA | Unhealthy | 139 | 27 | 112 | 80.0% | 82.0% | 82.0% |

clustering methods improves the performance of the clustering methods in Cleveland dataset According to the results presented in Table 3 we can demonstrate that K-means via PCA provides better classification accuracy than other methods and FCM via PCA yields close results to K-means via PCA method.

Using PCA, we plot the samples in the first two principal components. Fig 1 shows the Kmeans clustering process via PCA and grouping instances in to one of the two clusters. Fig 2 shows FCM clustering via PCA

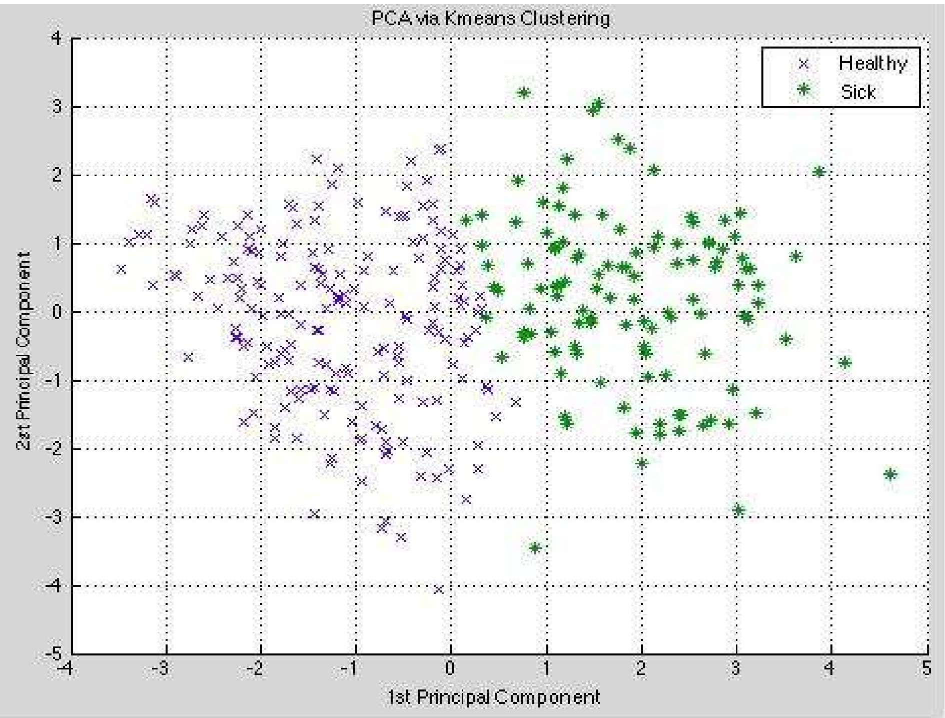

Figure 1. K means clustering via PCA.

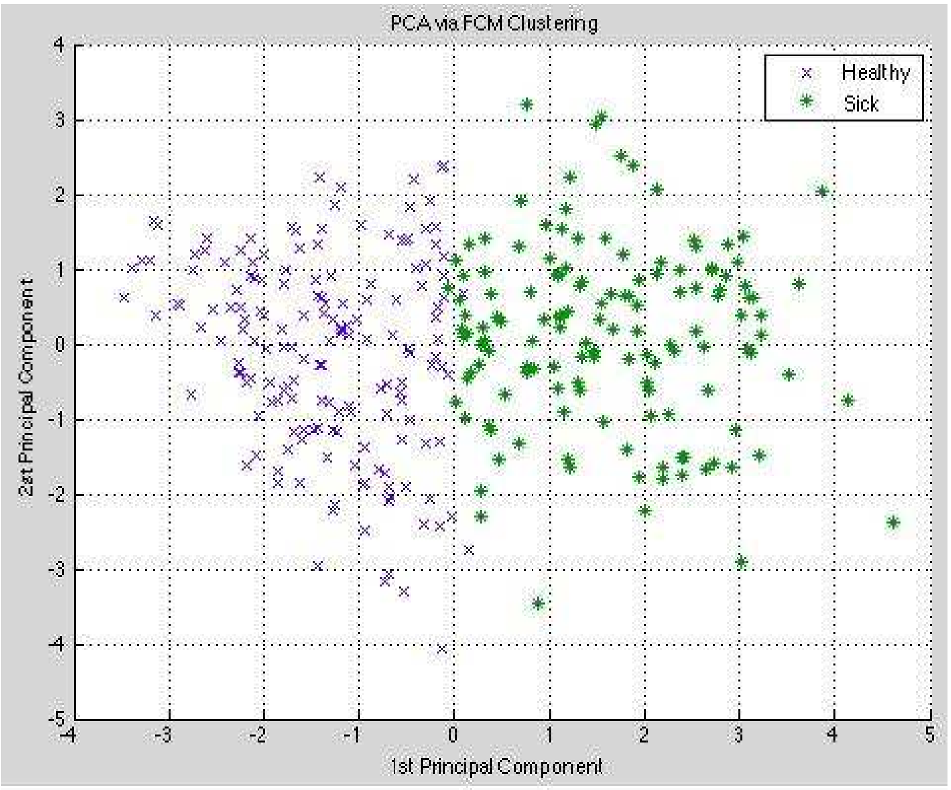

Figure 2. FCM clustering via PCA.

Fig 3 shows the comparison of the sensitivity, specificity and accuracy rates of the four classifier models. The results indicate that the proposed Kmeans via PCA algorithm performed better than the other approached models in terms of accuracy, specificity and sensitivity rates for 303 patients.

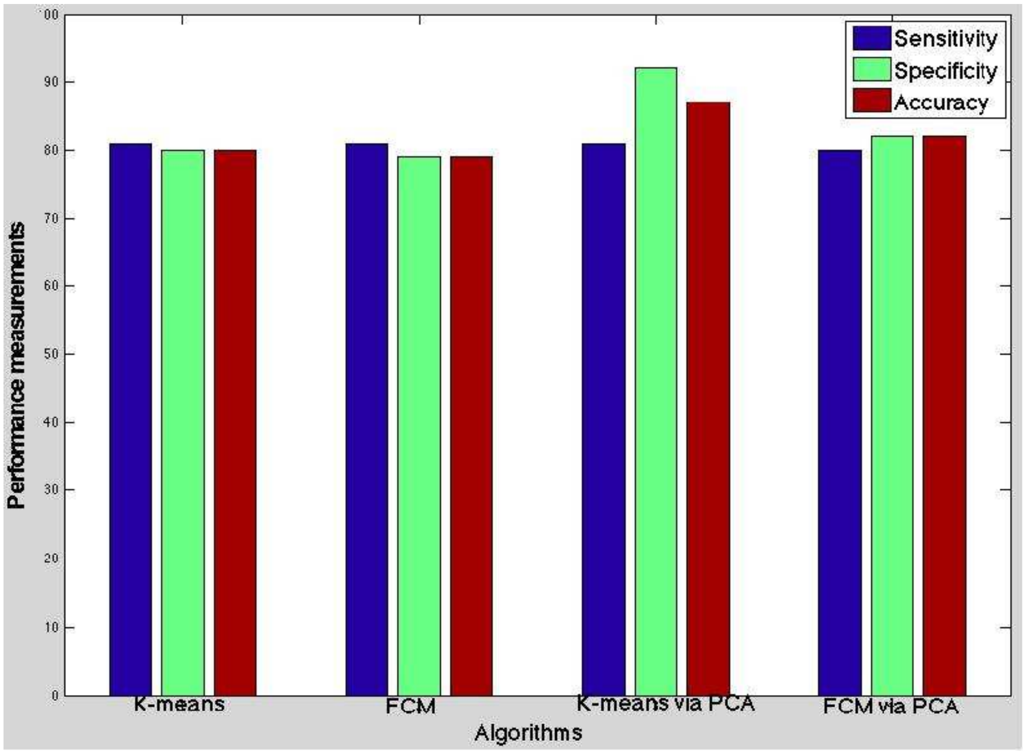

Figure 3. Performance results.

## 5. Conclusion

In this study the performance of clustering algorithms with PCA method was analysed on Cleveland heart disease dataset. For demonstrating and assessing the proposed classification approach, we computed performance metrics results for four methods. The experimental result show that, PCA improves the performance of clustering methods. In addition, using dimension reduction of PCA is important to visualize data in higher dimensional datasets in clustering problems We managed to classify the Cleveland dataset with 87% based on K-means via PCA clustering technique which are satisfying. In the Future an attempt will be done to test the supervised learning methods based on PCA in prediction of medical diagnosis.

## References

[1] World Health Organization, http://www.who.int/topics/cardiovascular_diseases/en/

[2] Patil, S. B., & Kumaraswamy, Y. S., Intelligent and effective heart attack prediction system using data mining and artificial neural network,*European Journal of Scientific Research*,31(4),642-656, 2009.

[3] Maimon,O. &Rokach, L., Data Mining and Knowledge Discovery Handbook, Springer, 2010.

[4] Jolliffe, I., Principal Component Analysis, Springer, 2nd edition, 2002.

[5] Ding, C., He, X., K-means Clustering via Principal Component Analysis, *International Conference on machine learning* , Banff, Canada, 2004.

[6] Ziasabounchi, N. & Askerzade, I., ANFIS based classification model for heart disease diagnosis, *international Journal of Engineering & Computer Sciences*,14(2),7-12, 2014.

[7] Patil, B.M., Joshi, R.C., Toshnival. D., Hybrid prediction model for type-2 diabetic patients, *Expert systems with applications*,37(12), 8102-8108, 2010.

[8] Singh, N., Mohapatra, A.G. & Kanungo, G., Breast cancer mass detection in mammograms using K-means and fuzzy C-means clustering, *International Journal of Computer Applications*,22(2),15-21, 2011.

[9] Chitra, R. & Seenivasagam, V., Heart attack prediction system using fuzzy C-mean classifier, *IOSR Journal of Computer Engineering*,14,23-31, 2013.

[10] Kahramanli, H. & Allahverdi, N., Design of a hybrid system for the diabetes and heart disease, *Expert systems with applications*,35,82-89, 2008.

[11] Askerzade, I.N. & Mahmud, M., Design and implementation of group traffic control system using fuzzy logic, *International Journal of Research and Reviews in Applied Sciences*,6,196-202, 2011.

[12] Askerzade, I.N. & Mahmud, M., Control the extension time of traffic light in single junction by using fuzzy logic, *International Journal of Electrical & Computer Sciences*,10(2),48-55, 2011.

[13] Sundar, B., Devi, T., & Saravanan, N., Development of a clustering algorithm for prediction heart, *International Journal of Computer Application*,48(7),8-13, 2013.

[14] Yin, J., Sun, H., Yang,J., & Gou, Q., Comparison of K-Means and fuzzy C-means algorithm performance for automated determination of the arterial input function, *PLoS ONE*, 9(2),e85884, 2014.

[15] Howley, T., Madden, G.M., Connel, M.L. & Ryder, G.A., The effect of principal component analysis on machine learning accuracy with high dimensional spectral data, *Knowledge Based Systems*,19(5),363–370, 2006.

[16] Indhumathi, R. & Sathiyabama, S., Reducing and clustering high dimensional data through principal component analysis, *International Journal of computer Application*,11(8),1-4, 2010.

[17] Avci, E., Turkoglu, I., An intelligent diagnosis system based on principle component analysis and ANFIS for the heart valve diseases, *Journal of Expert Systems with Application*,36,2873-2878, 2009.

[18] Napoleon, D. & Pavalakodi, S., A new method for dimensionality reduction using K- means clustering algorithm for high dimensional data set , *International Journal of Computer Applications*, 13(7),41-46, 2011.

[19] Fan, J., Han, M., & Wang, J., Single point iterative weighted fuzzy C-means clustering algorithm for remote sensing image segmentation, *Pattern Recognition*,42(11),2527–2540, 2009.

[20] Gath, I. & Geva, A.B., Unsupervised optimal fuzzy clustering, *IEEE Transaction on Pattern Analysis Machine Intelligence*, 11,773-780, 1989.

[21] Ross, T., Fuzzy logic with engineering applications, New York: McGraw Hill Co., 1995.

[22] Altun, S., Okur, V., Goktepe, A.B. & Ansal, A., Comparison of Dynamic Properties of Clays Obtained by Different Test Methods, 4$^{\text{th}}$ International Conference on Earthquake Geotechnical Engineering ,2007.

[23] Subbuthai, P., Periasamy, A., & Muruganand, S., Identifying the character by applying PCA method using Matlab, *International Journal of Computer Applications* ,60(1),8-11, 2012.

[24] Zhu, W., Zeng, N. & Wang, N., Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS Implementations, 2010.

[25] Hammouda, K. & Karray, F., A comparative study of data clustering techniques, SYDE 625: Tools of intelligent system design, course project, 2000.

Negar Ziasabounchi(n.z.sabounchi@gmail.com) –Department of Computer Engineering, Ankara University, Ankara, Turkey

Iman N. Askerzade –Department of Computer Engineering, Ankara University, Ankara, Turkey