Linear Discriminant Analysis for An Efficient Diagnosis of Heart Disease via Attribute Filtering Based on Genetic Algorithm

Rania Salah El-Sayed*
Department of Math &Computer Science, Faculty of Science, Al-Azhar University, Cairo. Egypt.

* Corresponding author. Email: Rania5salah@yahoo.com
Manuscript submitted May 10, 2018; accepted July 8, 2018.
doi: 10.17706/jcp.13.11.1290-1299

Abstract: Predicting of the heart disease is one of the important issues and many researchers develop intelligent medical systems to enhance ability of the physicians. In this paper we offer an intelligent system that diagnose and classify the severity of the disease due to heart failure. This system will use attribute filtering techniques genetic algorithm that has been known to be a very adaptive and efficient method of feature selection and reduce number of attributes which indirectly reduces the number of diagnosis tests which are needed to be taken by a patient. The classification techniques such as Support Vector Machines, Naive Bayesian Theorem, nearest neighbor and Linear discriminant analysis are used in this paper to know the classification accuracy of the techniques in the prediction of the heart disease. Apply proposed system on the Cleveland Heart Disease database. Then compare the results with other techniques according to using the same data.

Keywords: Genetic Algorithm (GA), Linear Discriminant Analysis (LDA), Support Vector Machines (SVM), Nearest Neighbor (KNN) and Principle Component Analysis (PCA).

1. Introduction

Heart diseases diagnosis is considered a significant task that needs to be carried out precisely and efficiently. The automation of the same would be highly beneficial. Predicting heart disease still remains the biggest cause of deaths worldwide.

The classification of medical data is still challenge issue due to recent advances in medical mining technology. So we develop system to improve classification model by using genetic algorithm.

In 2013 [1], Syed Umar Amin at al. used neural network Based on genetic algorithm in Prediction of Heart Disease Using Risk Factors and their system was implemented in Matlab for 50 people was collected from surveys done by the American Heart Association and predicts the risk of heart disease with an accuracy of 89%.

Mythili T. *et al.* in 2013 [2], introduce a heart disease prediction model using SVM-Decision Trees-Logistic Regression (SDL) and not apply on dataset.

In 2013[3], M.Akhil jabbar *et al.* introduces Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm. In 2012 LI Zhi-bin *et al*. build the model of transformer fault diagnosis which is based on the rough sets and support vector machine with accuracy 85.7143%.

In 2014 [4], Yuehjen E. Shao *et al.* used Hybrid intelligent modeling schemes for heart disease classification in their system apply logistic regression (LR), multivariate adaptive regression splines (MARS), artificial neural network (ANN), and rough set (RS) techniques. And the result of their model

78.57 % (LR+MARS) ,78.60% (RS) , 78.57% (MARS).

Babatunde Oluleye *et al.* 2014 [5], explain A Genetic Algorithm-Based Feature Selection and apply on Matlab use dataset Flavia Dataset and Ionosphere.

In 2015 [6], K. Sudhakar *et al.*, used artificial neural network classifier and PCA filtering method.

In 2015 [7], Kaur R. and Kaur S. classified Heart Disease Based on Risk Factors Using Genetic and apply model on 50 people collected based on the risk factors through different case studies.

Our an intelligent heart disease prediction system built with an efficient feature selection GA and data mining techniques such as naïve bayes, Linear discriminant analysis, and support vector machine for prediction and diagnosis of heart diseases. The result illustrated the peculiar strength of each of the methodologies in comprehending the objectives of the specified mining using UCI Machine Learning Repository.

This paper is structured as follows. Section 1 discusses relevant related works from the literature, Section 2 feature selection, Genetic Algorithm will be illustrated and methodology for classification, LDA, SVM and K-Nearest Neighbors (KNN) are reviewed. Section 3 provides methodology of proposed system. Experimental results using UCI Machine Learning Repository: Heart Disease Data Set are discussed in Section 4. A summary of this paper can be found in Section 5 where we provide its main conclusions and address future developments.

2. Preliminaries

In this section we discuss two concepts first about methodology of feature selection and second about classification.

2.1. Feature Selection

In machine learning, feature selection, which is also called variable selection, attribute selection is the process of obtaining a subset of relevant features for use in machine model construction. There are lots of techniques available such as PCA and GA.

2.1.1. Genetic algorithm

In feature selection, GA is used as a random selection algorithm, that exploring large search spaces [5], which is usually required in case of attribute selection. Such as; if the original feature set contains N number of features, the total number of competing candidate subsets to be generated is 2N, which is a huge number even for medium-sized N. (Fig. 1) illustrates the concept of GA attribute selector.

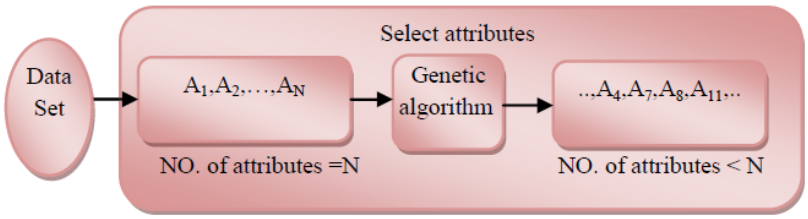

Fig. 1. Block diagram for GA feature selection.

As mention in [8], the five important problems in the GA are chromosome encoding, selection mechanisms, fitness evaluation, genetic operators and criteria to stop the GA. In comparative terminology to human genetics, chromosomes →bit strings, genes→ Features, locus → bit position, allele → feature value, phenotype →decoded genotype and genotype encoded string.

Genetic algorithms operate on a population of individuals to result better approximations. At each generation, a new population is created mainly by selection, crossover, and mutation see (Fig. 2) [9]. The

process of selecting individuals according to their level of fitness in the problem domain and recombining them together using operators borrowed from natural genetics [10].

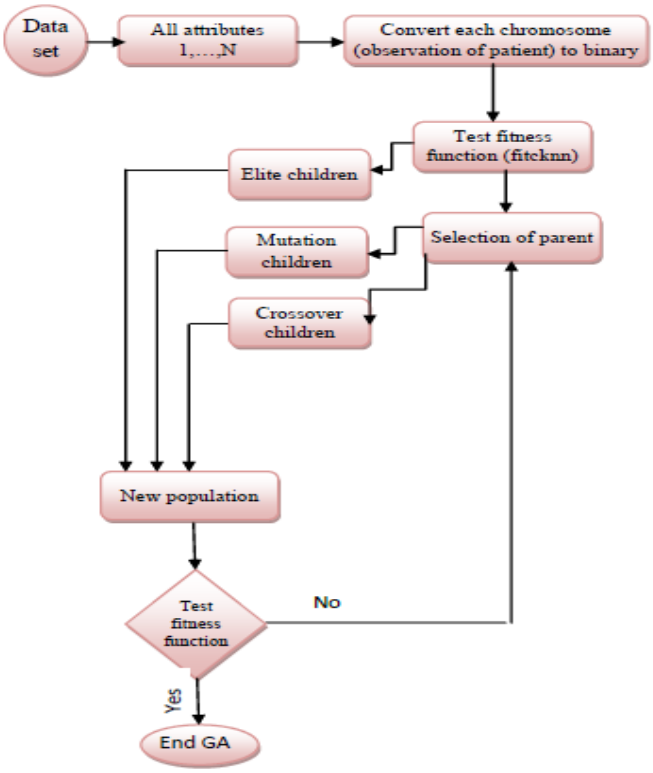

Fig. 2. Summary of GA process.

2.1.2. Principle Component Analysis (PCA)

The aims of principle component analysis PCA is extract the most important information from the database and keeping only important information [11]. And then simplify the dataset description and to analyze the variables and structure of the observations. We can represent procedure of PCA in algorithm 1 [6].

Algorithm 1 PCA
1. Input all observations(Patient data) in matrix
2. Calculate mean then subtract it from dataset
3. Calculate covariance matrix
4. Calculate the Eigen vector and Eigen values from covariance matrix
5. Form a feature vector as new dataset

2.2. Feature Classification

2.2.1. Support Vector Machine (SVM)

The Support Vector Machine (SVM) [12], is a supervised learning method for Data analysis, Classification and Regression analysis. It is a classification method based on statistical learning theory; SVM is a learning system that separates a set of pattern vectors into two classes with an optimal separating hyperplane. we can represent SVM schematically as shown in (Fig. 3) [9].

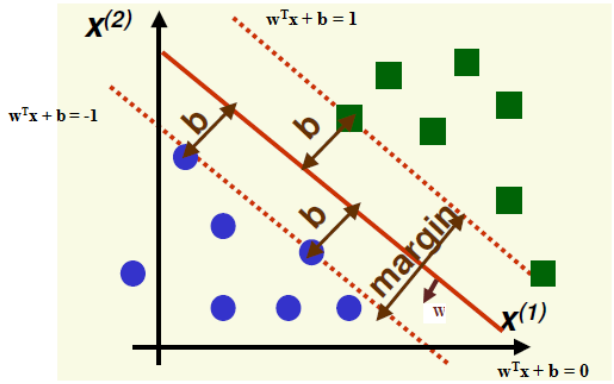

Fig. 3. Choosing the hyperplane that maximizes the margin.

where *w*: is the decision hyperplane normal vector, x_i: is the data point *i* , and y_i: is the class of data point i (+1 or -1). And the margin m is twice the absolute value of distance b of the closest example to the separating hyperplane. And the aim of SVM is to maximize this margin.

2.2.2. K-Nearest Neighbors (KNN)

The KNN classification algorithm tries to find the *K* nearest neighbors of the current sample and uses a majority vote to determine the class label. The KNN method is formally defined as follows [13], [14].

Suppose we have training dataset of labeled samples,

D= {(x_1, y_1) , (x_2, y_2) , … , (x_n, y_n)} ∈ R_px C

And a test vector $x^* \in R_p$, where *C* is some set of class labels. To classify $x*$, we find the *k* nearest training vectors in Euclidean distance (1) [15], say x_{r1}, …, x_{rk}. We then classify $x*$ by choosing the label $c \in C$ that appears most often in $y_{r1},\ldots,y_{rk}$.

$$d(x\,, x*) = \sqrt{\sum_1^n (x_n - x*)^2} \tag{1}$$

The class label of test data are the class label that represents the maximum of the k instances.

2.2.3. Linear Discriminant Analysis (LDA)

Linear discriminant analysis is a generalization of Fisher's linear discriminant, a method used in statistics and machine learning to separates two or more classes [16]. The procedure of DA shown in algorithm 2.

Algorithm 2 (DA)

1. Input all observations(Patient data) in matrix
2. Calculate Prior Probabilities Pi
3. Estimate the parameters of the conditional probability density functions
4. Compute discriminant functions.
5. Use cross validation to estimate misclassification probabilities
6. Classify observations (patients) with unknown group.

3. Methodology of Proposed System

The proposed system is generating reduced optimal attributes using genetic algorithm then using the new attributes in classification. (Fig. 4) represents the block diagram for proposed system.

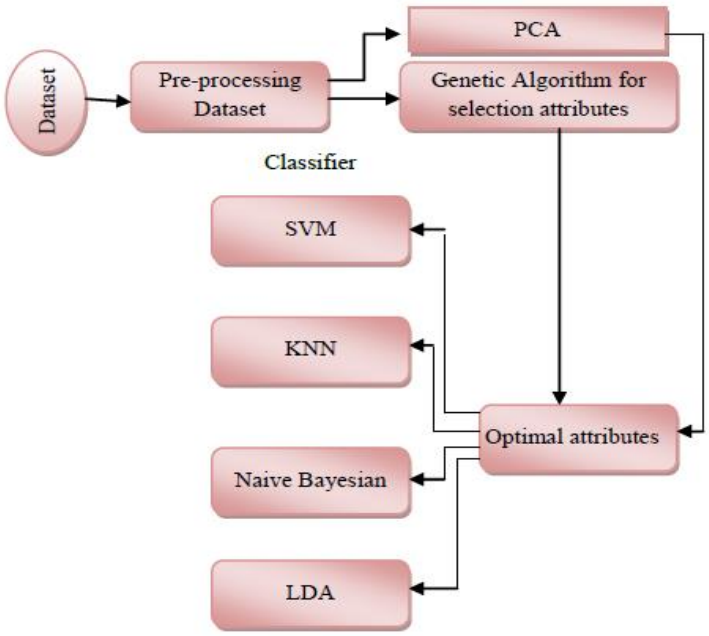

Fig. 4. System architecture for proposed heart diseases diagnosis.

3.1. Dataset Description

Apply our algorithm on UCI Machine Learning Repository: Heart Disease Dataset [17], that contains important risk factor related to heart diseases as shown in Table 1. The data collected from 303 patients with number of attributes: 14 (where 14th is the predicted attribute uses a value "1" for patients with heart disease low risk level, "2" median risk level, 3" for high risk level, "4" serious risk of heart disease and value "0" for patients with no heart disease (absence)).

We use 297 observations (patients) due to this data are complete samples and six are samples with missing attributes [17].

Table 1. Analysis of Attributes in Dataset

No.	Attribute(Risk factor)	Value of Risk factor
1	age	age in years
2	sex	(1 = male; 0 = female)
3	Cp (chest pain type)	1: typical angina 2: atypical angina 3: non-angina pain 4: asymptomatic
4	Trestbps (resting blood pressure)	Below 120 mm Hg-Low 120 to 139 mm Hg- Normal Above 139 mm Hg- High
5	Chol (serum cholestoral)	Below 200 mg/dL - Low 200-239 mg/dL - Normal 240 mg/dL and above - High
6	Fbs (fastin blood sugar)	>120 1 = true; 0 = false
7	Restecg (resting electrocardiographic results)	0: normal 1:having ST-T wave abnormality 2:showing probable or definite left ventricular

		hypertrophy by Estes' criteria
8	Thalach	maximum heart rate achieved
9	Exang(exercise induced angina	1 = yes 0 = no
10	Oldpeak	ST depression induced by exercise relative to rest
11	Slope(the slope of the peak exercise ST segment)	1: upsloping 2: flat 3: downsloping
12	Ca(number of major vessels colored by flourosopy)	Between 0 - 3
13	thal	3 = normal; 6 = fixed defect; 7 = reversable defect
14	Num (the predicted attribute)	0: No heart diseases 1:low risk level 2:median risk level 3:high risk level 4:serious risk level

3.2. The Proposed Procedure

- **Input**: n training samples partitioned into m classes, S_1, S_2,...,S_m and a test sample P.
- Set matrix training = $\begin{bmatrix} A_{1,1} & A_{1,2} & A_{1,r-1} \\ A_{2,1} & A_{2,2} & A_{2,r-1} \\ \vdots & \vdots & \vdots \\ A_{n,1} & A_{n,1} & A_{n,r-1} \end{bmatrix}$

where *r* no. of attributes and each row in A represent data of one patient.

- Set matrix group=$\begin{bmatrix} A_{1,r} \\ A_{2,r} \\ \vdots \\ A_{n,r} \end{bmatrix}$
- Set matrix test=$[A_{p,1} \quad A_{P,2} \quad \dots \quad A_{p,r-1}]$
- Set parameters of genetic algorithm
  - Attributes No. (6,8,9)
  - PopulationSize 20
  - Generations 80
  - Fitness function K-mean
  - CrossoverFraction 0.8000
  - MigrationInterval 20
  - MigrationFraction: 0.2000
  - EliteCount 1
- **Compute** matrix vector with the indexes of the attributes that composes the optimum set of attributes by apply genetic algorithm selection
- Set new matrix with reduced attributes
- **Classify** test matrix with LDA
- **Output** class to which test belongs

4. Experimental Results

The proposed methodology has been simulated in MATLAB 8.3.0 (R2014a) with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0) [17].

The sample size became 297 cases. Among the 297 cases used in the present study, the first 178 cases approximately 60% of the total cases were selected as the model building set (training sample), while the

remaining 119 cases (approximately 40% of the total cases) were retained as the validation set (testing sample). Details of number of sample in each class are shown in Table 2.

The experimental results have two main approaches. The first applying GA for reduce attributes from dataset and then LDA as classifier. Second approaches applying GA , and using MultiLDA and MultiSVM as classifier for multiclass dataset. Then compare results with other approaches in published literature. As shown in Table 3.

Table 2. Details of Training and Testing Dataset Sample

No. of class	Name of class	No. of sample ∀ class	No. of training ∀ class	No. of Testing ∀ class
2	0: No heart disease	160	98	62
	1 : heart disease	137	80	57
5	0 : No heart disease	160	98	62
	1 : Low Risk level	54	34	20
	2: Median Risk level	35	16	19
	3 : High Risk level	35	20	15
	4: Serious Risk level	13	10	3

Table 3. Accurate Identification Rate Comparisons for Heart Disease Dataset

algorithm	No. of attribute	No. of class	Name of class	No. of test sample	No of Recognition ∀ class	Accuracy
K-NN	13	2	0	62	45	60.50%
			1	57	27	
SVM	13	2	0	62	56	81.51%
			1	57	41	
Naive Bayesian	13	2	0	62	57	84.87%
			1	57	44	
PCA+KNN	13	2	0	62	44	61.34%
			1	57	29	
PCA+SVM	13	2	0	62	51	77.31%
			1	57	41	
PCA+ Naive Bayesian	13	2	0	62	40	69.75%
			1	57	43	
GA+SVM	8	2	0	62	58	85.71%
			1	57	44	
GA+SVM	6	2	0	62	59	87.39%
			1	57	45	
GA+ Multi SVM	6	5	0	62	59	**65.54%**
			1	20	12	
			2	19	5	
			3	15	2	
			4	3	0	
GA+LDA	6	2	0	62	62	**89.07%**
			1	57	44	
GA+ Naive Bayesian	6	2	0	62	58	86.55%
			1	57	45	
GA+LDA	6	5	0	62	62	**67.22%**
			1	20	7	
			2	19	5	
			3	15	5	
			4	3	1	

In this experimental work Table 3, Performance results were presented based on the prediction outcomes of the test set. We evaluate twelve approaches for diagnosis of heart diseases and compare the accurate identification rate of the previously published work with proposed approach.

In Table 3 we apply first single-stage classifiers such as K-NN, SVM and Naive Bayesian. The classification result 60.50%,81.51% and84.87% respectively. And second we apply hybrid model by using attribute filtering PCA with the same classifier, the results 61.34%, 77.31% and 69.75% respectively so we note that using PCA as attribute filter reduce the accuracy rate but when using genetic algorithm as attribute filter with classifier SVM, Naive Bayesian and LDA give accurate identification rate 87.39%, 86.55% and 89.07%. This prove that the proposed approach (GA+LDA) achieves the highest performance in diagnosis of heart diseases (89.07%) because we used the properties of GA in the stage of attribute filter that selected 6 attribute from 13 and then using the properties of Linear discriminant analysis (LDA) in classification. the classification table chart shown in (Fig. 5) indicaties the prediction classification of models and Accurate identification rate.

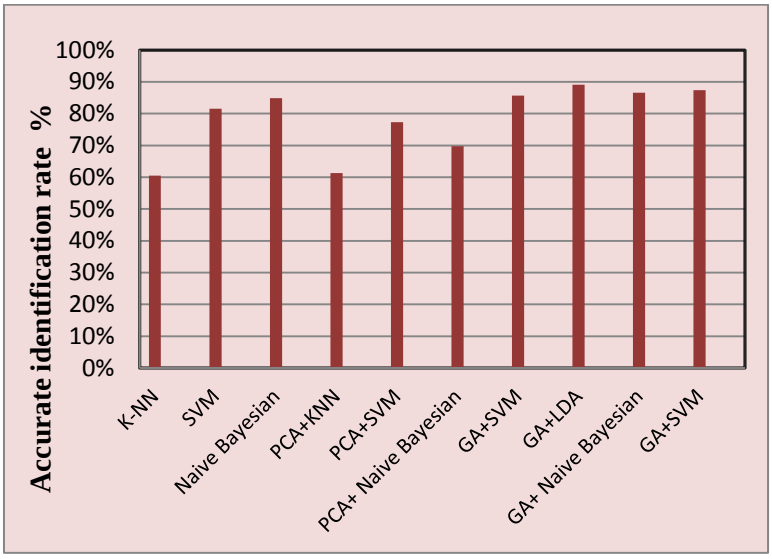

Fig. 5. Accuracy comparison of various algorithms.

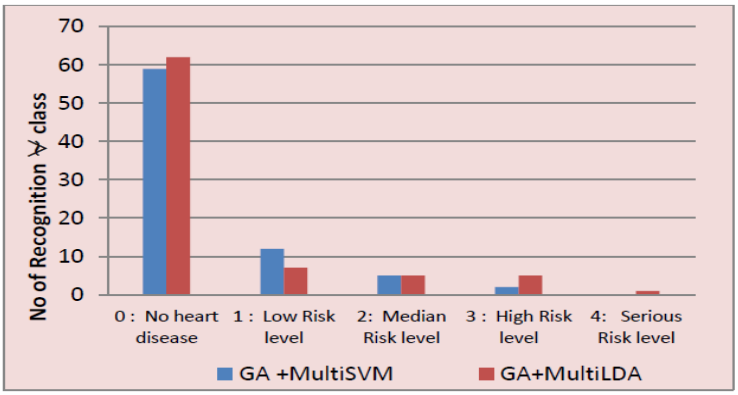

Fig. 6. No of recognition ∀ class for multiclass models.

The third part in our experimental work used multiclass for heart diseases. The previous studies that use multiclass classification approach [18] is hybrid model combine (Binary tree +SVM) and reported accuracy of (61.86%). In our study we apply GA in attribute filter stage that select 6 attributes from 13 then using MultiSVM and MultiLDA in classification, the result 65.54% and 67.22% respectively. This result prove that

our proposed approach using LDA as classifier after select attributes by GA improve the accurate identification rate not only when using 2 class but also when using multiclass. (Fig. 6) plots number of observations actually belong to class in models of multiclass.

5. Conclusion

We have presented an efficient hybrid model for diagnosis of heart diseases not only on two classes but also for multiclass. In our studied using genetic algorithm to optimize classification by LDA enhance the performance of linear discriminant analysis. With our system we predict not only if patient healthy or not but it predict the intensity of risk level of heart disease as shown in (Fig. 6) And we done comparative study with classification techniques such SVM, Naive Bayesian Theorem and KNN. An average recognition rate (89.07%) & (67.22%) is achieved for two class and multiclass heard diseases respectively. This means that our approach achieves a high recognition rate compared to other approaches in published literature. The occurrence of inaccurate data on diagnosis of heart disease data for risk levels, resulting in low performance of the system.

We believe that diagnosis the severity of heart diseases is still an interesting area of research, and we anticipate that there will be many further advances in this area.

References

[1] Amin, S. U., Kavita, A., & Rizwan, B. (2013). Genetic neural network based data mining in prediction of heart disease using risk factors. *Proceedings of IEEE Conference on Information & Communication Technologies (ICT)* (pp. 1227-1231).

[2] Mythili, T., Dev, M., Nikita, P., & Abhiram, N. (2013). A heart disease prediction model using svm-decision trees-logistic regression (sdl). *International Journal of Computer Applications, 16(68),* 11-15.

[3] Jabbar, M. A., Deekshatulu, B. L., & Chandra, P. (2013). Classification of heart disease using artificial neural network classification of heart disease using artificial. *Online, 13(3).*

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

[5] Oluleye, B., Leisa, A., & Dean, D. (2014). A genetic algorithm-based feature selection. *Online, 5(4),* 899-905.

[6] Sudhakar, K., & Manimekalai, M. (2015). An ensemble optimization for heart disease classification and attribute filtering. *International Journal of Engineering, 4*, 318-323.

[7] Kaur, R., & Kaur, S. (2015). Prediction of heart disease based on risk factors using genetic SVM classifier. *International Journal, 5(12),* 205-208.

[8] Sivanandam, S. N., & Deepa, S. N. (2008). *Introduction to Genetic Algorithms.* Berlin, Heidelberg: Springer-Verlag.

[9] Kumar, G. R., & Ramachandra, G. A. (2014). An efficient feature selection system to integrating SVM with genetic algorithm for large medical datasets. *International Journal, 4(2),* 272-277.

[10] Santhanam, T., & Padmavathi, M. S. (2015). Application of K-Means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis. *Procedia - Procedia Computer Science, 47,* 76-83.

[11] Negar, Z., & Iman, N. A. A comparative study of heart disease prediction based on principal component analysis and clustering methods. *Turkish Journal of Mathematics and Computer Science*, 1-11.

[12] Boswell, D. (2002). Introduction to support vector machines. *History,* 1-15.

[13] Nitin Bhatia, V. (2010). Survey on nearest neighbor techniques. *IJCSIS, 80(2).*

[14] Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information*

Theory, 13(1), 21-27.

[15] Akhil, M., Deekshatulu, B. L., & Chandra, P. (2013). Classification of heart disease using K- nearest neighbor and genetic algorithm. *Procedia Technology, 10,* 85-94.

[16] Tahmasebi, P., Hezarkhani, A., & Mortazavi, M. (2010). Application of discriminant analysis for alteration separation. *Australian Journal of Basic and Applied Sciences, 6(4),* 564-576.

[17] UCI machine learning repository. Retrieved from http://archive.ics.uci.edu/ml/datasets/Heart+Disease

[18] Kusnanto, H., & Wiharto, H. (2015). Performance analysis of multiclass support vector machine classification for diagnosis of coronary heart diseases. *International Journal, 5(5),* 27-37.

Rania Salah El-Sayed is a lecturer in the Department of Mathematics & Computer Science, Faculty of Science, Al-Azhar University, Cairo, Egypt. She received her Ph.D and M.Sc in pattern recognition and network security from Al-Azhar University in 2013 and 2009 respectively. Her B.Sc degree in math & computer science was received in 2004 from Al-Azhar University. In 2012, she received CCNP security certification from Cisco. Her research interests include pattern recognition, application & network security.