Advances In Multidisciplinary & Scientific Research
A Multidisciplinary & Interdisciplinary Journal

Vol. 4 No. 3, September 2018

An Improved Heart Disease Classification System Using Probabilistic Principal Component Analysis and K-Nearest Neighborhood

Akinrotimi, Akinyemi Omololu & Aremu, Dayo Reuben
Department of Computer Science
Faculty of Communication and Information Sciences
University of Ilorin
Ilorin Nigeria
, PMB 1515, Ilorin, Kwara State, Nigeria.
E-mail: timiakin2011@yahoo.com[1], draremu2006@gmail.com[2]

ABSTRACT

One of the most widely used techniques often employed in mining knowledgeable information from medical data bases are Data Mining techniques. While Data mining techniques have proved to be effective in building models illustrating important data classes, especially where class attribute is involved in the construction of the classifier, K-Nearest Neighbor (KNN) is has proved to be an indispensible tool, in pattern recognition. The volume of Medical data bases are huge in nature and getting the right component from them, is essential for data scaling and better data normalization before presenting them for classification. Heart disease is one of the leading causes of death in the world today and based on statistics, the number of adults living with heart failure, increased from about 5.7 million between 2009-2012, to about 6.5 million between 2011-2014 (according to the American Heart Association's 2017 Heart Disease and Stroke Statistics Update).There is therefore need to provide medical practitioners with decision support systems that can help them in the early detection of heart conditions. In this paper, the Probabilistic Principal Component Analysis (PPCA), is used to preprocess and extract components from a clinical dataset, in order to provide a more organized and detailed data to be passed into the KNN classifier, thereby allowing for better classification. Experimental results shows that the combination of the PPCA and KNN is very productive for the prediction of heart diseases as it achieved an accuracy of 98.08%.

Keywords*:* Feature Extraction, Heart Disease, Performance Evaluation, Classification.

Aims Research Journal Reference Format:
A.O. Akinrotimi & D.R. Aremu (2018): An Improved Heart Disease Classification System Using Probabilistic Principal Component Analysis And K-Nearest Neighborhood. Advances in Multidisciplinary & Scientific Research Journal. Vol. 4. No.2, Pp 45-52.

1. INTRODUCTION

Heart and blood vessel diseases called cardiovascular diseases comprise of several problems, many of which are related to the atherosclerosis process, which is a condition develops when a substance called plaque builds up inside the walls of arteries. This condition shrinks the size arteries and causes problems for blood flowing through the vessels and thus increases the risk of heart diseases. The World Health Organization report estimated that more than 80% death from cardiovascular disease take place in low middle income countries [6]. The use of intelligent methods and algorithms (e.g. neural network, fuzzy logic and genetic algorithm) has started to play crucial role in complex and uncertain medical tasks such as diagnosis of diseases, making it possible for medical practitioners to predict the occurrence of heart diseases.

In last decade, information about the use of intelligent methods in medicine domain has seen immense. As reported in [4], studies on computer aided applications and tools for diagnosis and treatment of patients seems to be a more recent area of interest. In addition, medical practitioners are also employing computerized technologies to assist in diagnosis and opinions as medical diagnosis is full of uncertainties [4]. On the other hand, the use of fuzzy logic and neural networks also stand as efficient methodologies in dealing with these uncertainties [4]. In cases where vague data or prior knowledge is involved, both of them serve certain advantages over classical methods.

According to [2], clinical decisions are often made based on doctors' intuition and heuristic experience, rather than on the knowledge rich data hidden in the database. These lead to unwanted biases, errors and excessive medical costs which affect the quality of treatment provided to patients [3]. In this research work, a method that can be used in efficiently diagnosing heart diseases, leading to decreased medical errors and superfluous practice variation, thereby decreasing diagnostic time and enhancing patient safety and satisfaction is proposed.

The use of machine learning algorithms such as neural network, fuzzy logic, genetic algorithm and neuro-fuzzy systems has highly helped in complex and uncertain medical tasks such as diagnosis of diseases [7]. This paper investigates applying KNN in the diagnosis of heart diseases on these benchmark dataset, with improved technique for data extraction using Probabilistic Principal Component Analysis (PPCA).

2. LITERATURE REVIEW

Over the years, numerous works related to heart disease prediction system using different data mining algorithms, has been done by different authors. They tried to achieve efficient methods and accuracy in finding out research, diseases related to the heart, including datasets and different algorithms along with the experimental results and future work that can be done on the system to achieve more efficient results. This paper aims at analyzing different data mining techniques that has been introduced in recent years for heart disease prediction system by different authors.

Probabilistic PCA (PPCA) is a probabilistic formulation of PCA based on a Gaussian latent variable model.It was first introduced by Tipping and Bishop in 1999 [8]. The PPCA model reduces the dimension of high-dimensional data by relating a *p*-dimensional observed data point to a corresponding *q*-dimensional latent variable through a linear transformation function, where $q \ll p$. Principal components analysis (PCA) is probably the most widely used technique for analyzing metabolomic data . The popularity of PCA in metabolomics is due to the fact that it is a simple non-parametric method which can project the NMR or MS spectra into lower dimensional space, revealing inherent data structure, and providing a reduced dimensional representation of the original data. Despite its widespread use in metabolomics, PCA has several shortcomings. Most significantly, PCA does not have an associated probabilistic model, which makes assessing the fit of PCA to the data difficult and limits the potential to extend the scope of application of PCA. Additionally, PCA can fail to reveal underlying groups of subjects in the data, therefore providing a spurious view of the underlying data structure [9, 10]. Other limitations include the inability of PCA to deal with missing data appropriately [11].

K-nearest neighbor (KNN) is a simple algorithm, which stores all cases and classifies new cases, based on similarity measures.KNN algorithm also refffered to as: 1) case based reasoning 2) K-nearest neighbor 3)example based reasoning 4) instance based learning 5) memory based reasoning 6) lazy learning [14]. KNN algorithms have been used since 1970 in many applications like statistical estimation and pattern recognition etc.KNN is a non parametric classification method which is broadly classified into two types1) structure less NN techniques 2) structure based NN techniques. In structure less NN techniques whole data is classified into training and test sample data. From training point to sample point distance is evaluated, and the point with lowest distance is called nearest neighbor. Structure based NN techniques are based on structures of data like orthogonal structure tree (OST), ball tree, k-d tree, axis tree, nearest future line and central line. K-Nearest-Neighbor is one of the most widely used data mining techniques in classification problems. Its simplicity and relatively high convergence speed make it a popular choice. However a main disadvantage of KNN classifiers is the large memory requirement needed to store the whole sample [15].
In [16] C4.5 algorithm, MAFIA and K means clustering was used in analyzing multifarious data, using 13 attributes in the dataset and achieving 89 percent accuracy. [17] Shows the classification techniques in data mining and the performance of classification amongst them. A decision tree and SVM was used in performing classification, proofing more accurate than the other methods by achieving 91% accuracy. In [18] a prediction system for heart diagnosis using decision tree, Neural Network and Naive Bayes techniques using 15 attributes was developed and in [19], a computer aided heart disease prediction system that helps the physician as a tool for heart disease diagnosis was developed. From this analysis it was concluded that neural network with offline training is good for disease prediction at an early stage and an accurate performance can be obtained by pre-processing and normalizing the dataset.

3. METHODOLOGY

3.1 Data Collection

The data set used in this study is the benchmark Cleveland Clinic Foundation Heart disease data set available at http://archive.ics.uci.edu/ml/datasets/Heart+Disease. The dataset consists of 8 attributes with seven predicting variables and one class label and total observation of 209 instances.

3.2 Data Filtering.

At this stage every inconsistent data was removed, including outliers and string variables. There were converted into numeric variables for proper labeling.

3.3 Feature Extraction.

The probabilistic principal component analysis was used to extract feature scores from the data so as to scale and center the data in a well normalized format for a better output.

3.3.1 PPCA

Let $x_i = (x_{i1}. . . x_{ip})^T$ be an observed set of variables (eg. an NMR spectrum) for observation *i* and $u_i = (u_{i1}, . . . , u_{iq})^T$ be a latent variable corresponding to observation *i* in the latent, reduced dimension space. In terms of traditional PCA, u_i can be viewed as the principal score of subject *i*. The PPCA model can be expressed as follows

$$\overline{x_i} = W\, u\overline{i} + \overline{\mu} + \epsilon\overline{i} \qquad (3.1)$$

Where W is a $p \times q$ loadings matrix, μ is a mean vector and ϵ i is multivariate Gaussian noise for observation *i*, i.e. $p\,(\epsilon.\overline{i}) = MV\,N_p(0, \sigma^2 I)$ where I denotes the identity matrix. The latent variable u_i is also assumed to be multivariate Gaussian distributed, $p\,(u_i) = MV\,N_q\,(0, I)$.The conditional distribution of the observed data given the latent variable can then be expressed as

$$p(\overline{x_i}\,|\overline{u_i}) = MVNp(W\,\overline{u_i}+\mu,\sigma 2\overline{I}) \qquad (3.2)$$

The distribution of the observed data, $p\,(x_i)$, also known as the predictive distribution, can be derived from the convolution of $p\,(u_i)$ and $p\,(x_i|u_i)$ giving

$$p(\overline{x_i}) = MVNp(\mu, WWT+\sigma 2I) \qquad (3.3)$$

In contrast to the more conventional view of PCA, which is a mapping from a high dimensional data space to a low dimensional latent space, the PPCA framework is based on a mapping from a latent space to the data space. The observed data x_i is generated by first drawing a value for the latent variable u_i from its unit variance multivariate Gaussian distribution, $p\,(u_i)$. The observed variable x_i is then sampled, conditioning on the generated value for u_i, from the isotropic distribution defined in (1).

Any observed data point x_i can be represented in a latent space by its corresponding *q*-dimensional latent variable u_i. The distribution of the latent variable given the observed data can be derived using Bayes' Theorem to give:

$$p(\overline{u_i}\,|\,\overline{x_i}) = MVNq(M{-}1WT(\,x_i{-}\mu), \sigma^2\overline{M}{-}\overline{1}) \qquad (3.4)$$

Where M is a $q \times q$ matrix defined as $M = W^T W + \sigma^2 I$.

A key benefit of the PPCA approach is that, not only is an estimate of the location of each observation in the lower dimensional space available through its expected value $(u_i) = M^{-1}W^T(x_i - \mu)$, an estimate of its associated uncertainty is also available through the covariance matrix $\sigma^2 M^{-1}$ in (2). This is in contrast to conventional PCA where the lower dimensional location (i.e. the score) of an observation is available, but the uncertainty associated with it is not. The parameters (W, μ and σ^2) of the PPCA model can be estimated using maximum likelihood. Maximizing the (log) likelihood function with respect to model parameters is non-trivial; in [13] it is demonstrated that the estimates do however have closed form solutions. Crucially, the log likelihood of the PPCA model is maximized when the columns of W span the principal subspace of conventional PCA. Thus the maximum likelihood estimate (MLE) of the loadings matrix $\hat{W}$ in PPCA corresponds exactly to the loadings matrix in conventional PCA. Hence the model output in PPCA is exactly that obtained in conventional PCA, but with the additional advantages of uncertainty assessment and potential model extensions.

3.4 Data Partitioning.

After the feature extraction the data was partitioned into the training and testing set so as create a robust knowledge discovery for the KNN classifier and validate the classifier. The data was grouped at 75% training and 25% testing.

3.5 Classification

After data partitioning into training and validating set, the training set was introduced into the KNN Euclidean classifier for pattern discovery and deep learning, this helps to create an experimental self-learning of the dataset. Theoretically, in this method, first of all, the KNN classifier will be trained with given dataset and known output. Then, after finishing the training session, the KNN will be taken into its application session and then, new unknown data that was totally different from the data given in the training session will be provided. Threafter, the KNN classifier will try to find out the distance between the training result and present result. The result, which will be minimum, will be considered as the closest neighbor.

Mathematically, it works based on the Euclidian Distance:-

The vector $X'_n \in \{ X_1, X_2, \ldots, X_N \}$ is the nearest neighbor of X_{test} if X'_n is the class of X_{test}.

3.6 Testing Phase

The model was validated with 25% of the dataset that was held out during data partitioning. This will help to determine the performance of the model.

3.7 System Flow Chart

The system algorithm flowchart is shown below.

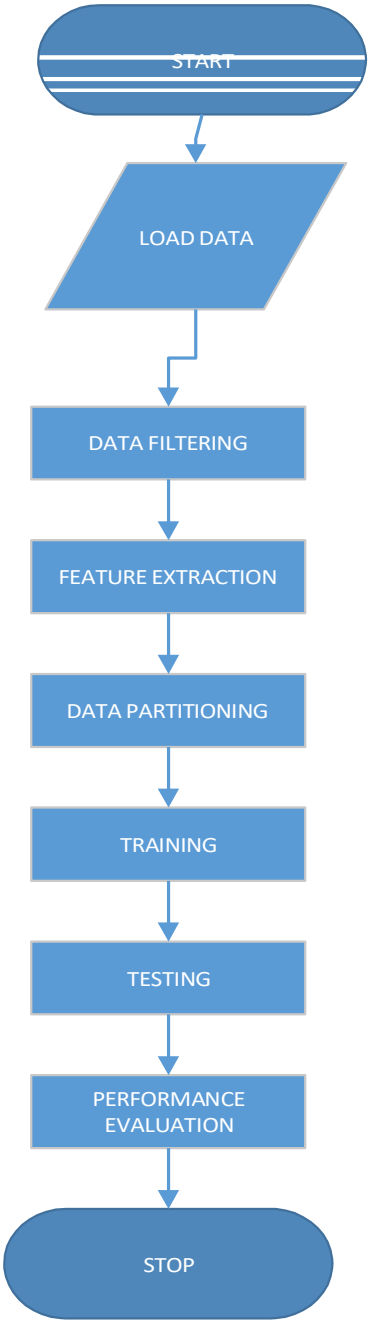

Figure 1.0: System Flow Chart

4. IMPLEMENTATION AND RESULTS

The implemented application for the heart disease using PPCA and KNN Euclidean is shown and emphasized below:

4.1.1 Data Filtering

The data was normalized by converting from string variable to numeric variable. A sample of the normalized data is shown below. Example is the class label that has 1= positive and 2 is negative. For the purpose of analytical structure, system performance and evaluation the dataset was partitioned into two, the training set and testing set at the ratio of 75% and 25%. The 75% of the data was used to create discover knowledge from the dataset. The 25% of the dataset was used to validate the system.

Table 1.0 Dataset filtering

Age	ches2_pain	res2_bpress	blood_sugar	res2_ elec2ro	max_hear2_ra2e	Exercise _angina	Disease
43	1	140	1	1	135	1	1
39	2	120	1	1	160	1	2
39	3	160	2	1	160	2	2
42	3	160	1	1	146	2	2
49	1	140	1	1	130	2	2
50	1	140	1	1	135	2	2
59	1	140	2	2	119	1	1
54	1	200	1	1	142	1	1
59	1	130	1	1	125	2	1
56	1	170	1	3	122	1	1
52	3	140	1	3	170	2	2
60	1	100	1	1	125	2	1

4.1.2 Feature Extraction

Principal component scores are a group of scores that are obtained following a Principle Components Analysis (PCA). In PCA the relationships between a groups of scores is analyzed such that an equal number of new "imaginary" variables (aka principle components) are created. The first of these new imaginary variables is maximally correlated with all the original group of variables. The PPCA helps to transform the data into a better normalized format, extract the component score before presenting it into the classifier. The extracted PPCA is shown below and the extracted results is saved with the save button.

	age	ches2_pain	res2_bpress	blood_sugar	res2_elec2ro	max_hear2_...	exercice_an...	di
1	-3.4162	5.1391	-5.7934	-0.8893	-0.3945	-0.5045	-0.1486	
2	26.6466	-6.5003	-4.1869	-0.1472	-0.1305	-0.3279	-0.8613	
3	14.9838	31.6372	-7.2375	1.0973	-0.2099	-0.1617	0.0020	
4	1.2986	27.4991	-6.5304	1.3364	-0.2214	0.0019	-0.0211	
5	-9.1544	3.7866	-0.6897	-0.2743	-0.4627	0.8285	0.0202	
6	-4.6302	5.3163	1.0943	-0.3698	-0.4636	0.8168	0.0167	
7	-21.2719	0.7341	7.3999	-0.8270	0.5009	-0.3800	-0.0417	
8	-16.2923	64.7216	1.5530	-0.9731	-0.6958	-0.3313	-0.0365	
9	-12.6975	-6.9995	9.1173	-0.5548	-0.4674	0.1604	0.6413	
10	-26.6897	30.1620	2.6256	-0.9227	1.3880	-0.1651	4.4060e-04	
11	27.9340	15.9134	8.6874	0.5694	1.8156	0.0831	0.0021	
12	-4.1113	-35.5735	12.4004	-0.6147	-0.3429	0.0984	0.5894	
13	-3.8345	26.9215	5.7784	-0.1662	-0.3967	-0.9057	-0.0947	
14	-1.1488	7.0036	8.7985	0.4012	-0.3600	0.3526	0.0325	
15	35.3714	-14.2601	-3.4563	-1.1976	-0.2598	-0.1457	0.5443	
16	-0.5248	-13.4849	12.4910	1.2321	0.8524	0.1282	-0.0265	
17	8.6036	9.9603	8.4247	0.2411	-0.3324	0.4046	-0.0184	
18	0.0455	6.8270	1.8999	-1.2624	1.5562	-0.3291	-0.0760	
19	-17.8567	-34.6684	-2.2625	-0.2700	-0.3383	-0.0018	0.6395	
20	-44.5475	43.9828	-18.1912	1.2145	-0.5082	0.5798	0.0924	
21	-45.9952	-6.6160	10.2810	-0.2945	-0.5250	-0.2321	-0.0847	
22	4.0174	24.2819	8.2550	-1.2158	-0.5171	-0.5473	-0.0818	
23	-2.1231	0.3969	-4.4271	-0.6343	-0.4158	0.0522	0.6238	

Figure 2.0: PPCA Extracted Component Scores

4.2 System Evaluation

4.2.1 Confusion Matrix

The confusion matrix is an indication of the correctly and incorrectly classified class. The class 2 which represents the positive shows a total observation of 22 observations for the validation set. A total of 22 were actually classified as positive and 0 were classified as incorrect, while the class 1 represents the negative class shows a total of 30 observations for the validation set. A total of 29 were actually classified as correct and 1 was classified as incorrect.

4.2.2 True Positive Rate/ False Negative Rate

The true positive rate shows the Positive and Negative class correctly identified as Positive and Negative class and the Positive and Negative class incorrectly identified Positive and Negative class.

Table 2.0 Statistical Metrics

CLASSIFICATION METRICS	RESULTS
True Positive	22
False Positive	0
True Negative	29
False Positive	1
SENSITIVITY = True positive rate = TP / (TP + FN)	0.956522
SPECIFICITY = True negative rate = TN / (TN + FP)	1
ACCURACY = TP + TN / (FP + FN + TP +TN)	98.0769%
RECALL	0.956522
FSCORE	0.977776
TRAINING TIME	1.2363 s

4.4 Graphical Results

The graphical result of the statistical metrics is shown below.

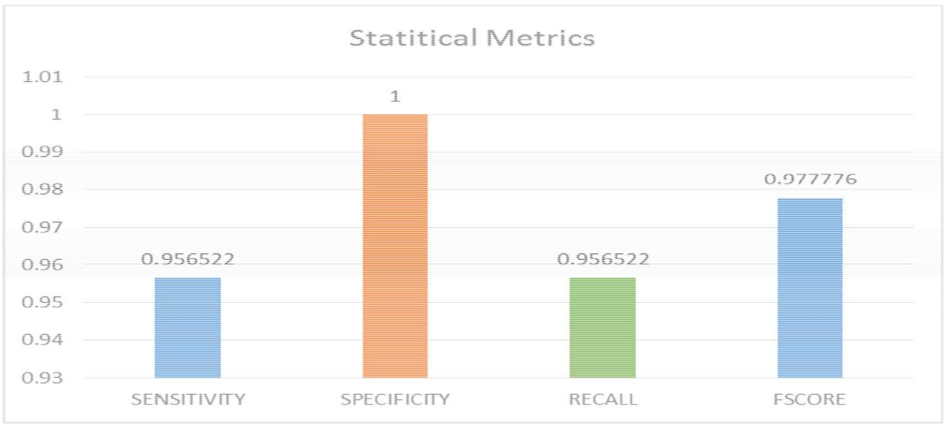

Figure 3.0: System Evaluation

The training time of the system seems to very fast the system was able to train finish at 1.263 seconds.

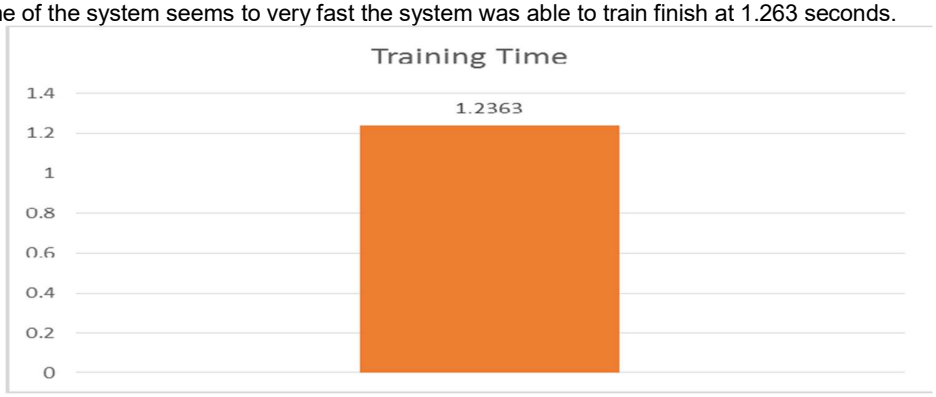

Figure 4.0: Training Time

The system also achieved a very high classification accuracy thereby increasing the true positive rate and decreasing the false positive rate. An accuracy of 98.0769% was achieved.

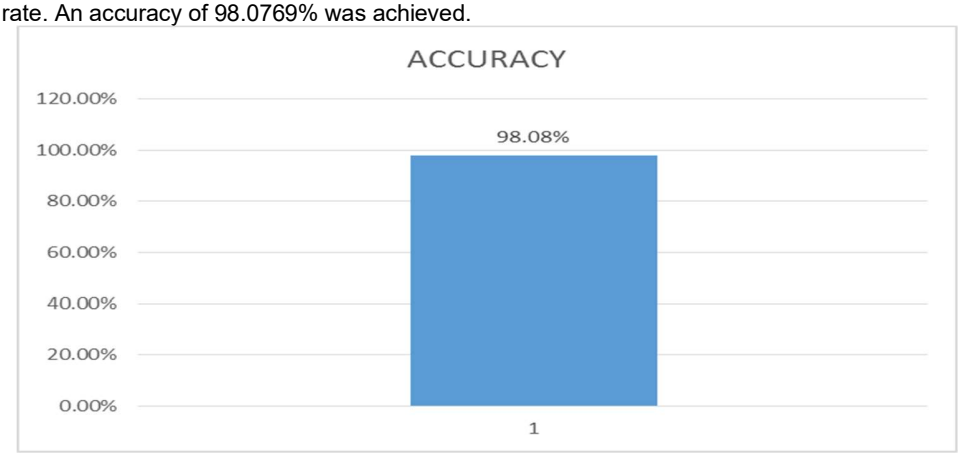

Figure 5.0: Classification Accuracy

5. CONCLUSION

Heart diseases are the main causes of death over the world. They represent 7.2 million death i.e. 12.8% of fatalities on the world .This system is expected to help doctors to early analyze such an ailment and evaluate coronary illness hazards. This research work considered heart disease prediction on male dataset using Probabilistic Principal Component Analysis and K-nearest neighborhood. The probabilistic principal components extracted components score, from the dataset for the purpose of data scaling and the extracted component was passed into the K-NN Euclidean classifier for prediction. The results obtained justifies effectiveness of the combination of these two data mining techniques as this was able to help achieve an accuracy of 98.0769% , thereby showing a high detection rate of heart conditions.

REFRENCES

1. Ali A. and Mehdi N. (2010),"A Fuzzy Expert System for Heart Disease Diagnosis" A Proceeding of the International Multi-Conference of Engineers and Computer Scientists, vol.1 .
2. Allahverdi, N., Torun, S., Sarıtaş, (2013) İ. Design of a Fuzzy Expert System for Determination of Coronary Heart Disease Risk, International Conference on Computer Systems and Technologies, CompSysTech'07.
3. Nazmy, T.M., El-Messiry, H., Al-Bokhity, B., (2015) .Classification of Cardiac Arrhythmia based on Hybrid System, International Journal of Computer Applications, 2(4).
4. Sikchi, S.S, Sikchi, S., Ali, M.S., (2012) Design of Fuzzy Expert System for Diagnosis of Cardiac Diseases, International Journal of Medical Science and Public Health, 2(1).
5. Wei, M., Bai, B., Sung, A., Liu, Q., Wang, J. & Cather (2007), M. Predicting injection profiles using ANFIS, Information Sciences Journal.
6. World Health Organization, 2014. http://www.who.org. 9.2.
7. ZAPTRON Systems, Inc. Neurofuzzy (1999),"A Different Type of Neural Nets" Zaptron's High Order Nonlinear Neural Networks.|Fuzzylogic|Neural.
8. M.A .Nishara Banu and B.Gomathy, (2014) "Disease Forecasting System Using Data Mining Methods".
9. Abhishek Taneja, (2013) "Heart Disease Prediction System Using Data Mining Techniques", Oriental Scientific Publishing Co., India.
10. Ms. Ishtake S.H, Prof. Sanap S.A., (2013) "Intelligent Heart Disease Prediction System Using Data Mining Techniques", International J. of Healthcare & Biomedical Research.
11. Chitra Jegan., (2013) "Review of heart disease prediction system using data mining and hybrid intelligent techniques, Noorul Islam University
12. Nidhi Bhatla, Kiran Jyoti, (2012), "An Analysis of Heart Disease Prediction using Different Data Mining Techniques", International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 8.
13. Tipping ME (2009), Bishop CM. Probabilistic principal component analysis. Journal of the Royal Statistical Society, Series B.
14. Saed Sayad, University of Toronto http://chem-eng.utoronto.ca/~data mining
15. Nitin Bhatia,Candana"Survey on nearest neighbor techniques"IJCSIS,Vol 80,no 2 (2010)
16. Tipping ME, Bishop CM: Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B* 1999, 61(3):611–622. 10.1111/1467-9868.00196
17. Chang D: On using principal components before separating a mixture of two multivariate normal distributions. Applied Statistics 1983, 32: 267–275. 10.2307/2347949
18. McLachlan GJ, Peel D: Finite Mixture Models. New York: Wiley; 2000.
19. Roweis S: EM Algorithms for PCA and SPCA. Neural Information Processing Systems 1998, 10: 626–632.