Heart Disease Prediction using PCA-KNN in Data Mining

Shweta Gupta*, Prof. Aditi Nema** and Prof. Kiran Agrawal***
*-***Department of Computer Science& Engineering, Bansal Institute of Research & Technology
Bhopal, India
shwetaguptacs03@gmail.com, aadi.nema@gmail.com

Abstract: Heart disease is considered as one of the major causes of death throughout the world. It cannot be easily predicted by the medical practitioners as it is a difficult task which demands expertise and higher knowledge for prediction. This paper addresses the issue of prediction of heart disease according to input attributes on the basis of data mining techniques. We have investigated the heart disease prediction using PCA-kNN technique through MATLAB2012a software. The performance of these data mining techniques is measured by combining the results of predictive accuracy, specificity and senstivity using a standard data set as well as a collected data set. Based on performance factor PCA-kNN techniques show optimum performances than the performances of SVM techniques.

Keywords: Data Mining, Heart Disease, PCA, k-NN, MATLAB.

Introduction

In today's opportunity at numerous spots clinical test outcomes are regularly made in light of specialists' instinct and experience as opposed to on the rich data accessible in numerous expansive databases. Numerous a times this procedure prompts inadvertent predispositions, lapses and a tremendous medicinal expense which influences the nature of administration gave to patients. Today numerous doctor's facilities introduced some kind of quiet's data frameworks to manage their social insurance or patient information. These data frameworks commonly produce a lot of information which can be in distinctive organization like numbers, content, diagrams and pictures yet sadly, this database that contains rich data is once in a while utilized for clinical choice making.

Like business knowledge and examination, the term information mining can mean diverse things to distinctive individuals. In exceptionally straightforward way we can characterize information mining as this is the investigation of substantial information sets to discover examples and utilize those examples to foresee or fore-cast the probability of future occasions. The motivation to do this problem comes from World Health Organization estimation. According to the World Health Organization estimation till 2030, very nearly 23.6 million individuals will pass on because of Heart malady. So to minimize the danger, expectation of coronary illness ought to be finished. Analysis of coronary illness is typically in view of signs, manifestations and physical examination of a patient. The most troublesome and complex assignment in medicinal services area is finding of right ailment. This colossal entirety huge of rough data is the rule resource that can be capably pre-taken care of and inspected for key information extraction that direct or by suggestion influences the remedial society for cost sufficiency and reinforce decision making. Authentic determination of coronary sickness can't be possible by using simply human understanding. There are heaps of parameters that can impacts the accurate conclusion like less exact results, less experience, time subordinate execution, data up degree and whatnot. Packs of headway and examination happened in this field using multi-parametric qualities with nonlinear and direct parts of Heart Rate Variability (HRV).A novel framework was proposed by Heon Gyu Lee et al. [1]. To fulfill this, various experts have used various classifiers e.g. CMAR (Classification Multiple Association Rules), SVM (Support Vector Machine), Bayesian Classifiers and C4.5). A latest's rate techniques in this field depicted in [2].Some plausible strategies and technique we recommended incorporates the clinical information institutionalization, examination and the information sharing over the related industries to improve the precision & viability of information mining applications in social insurance. [3] It is likewise prudent to investigate the utilization of content digging and picture digging for extension the nature and extent of information mining applications in medicinal services part. Information mining application can likewise be investigated on computerized indicative pictures for application viability. Some advancement has been made in these areas. [4][5]. There is a lot of data put away in stores that can be utilized viably to guide medical practitioners in decision making in human services. This brings up an essential issue: "By what means would we be able to transform information into helpful data that can empower medicinal services practitioners to settle on viable clinical decision?" This is the primary goal of this research. In this paper, we present data mining algorithm PCA-KNN to diagnosis the heart disease and its simulation is done in MATLAB2012a and comparative analysis is done using accuracy, specificity and sensitivity performance parameters.

Heart Disease

The heart is important organ or part of our body. Life is itself dependent on efficient working of heart. If operation of heart is not proper, it will affect the other body parts of human such as brain, kidney etc. It is nothing more than a pump, which pumps blood through the body. If circulation of blood in body is inefficient the organs like brain suffer and if heart stops working altogether, death occurs within minutes. Life is completely dependent on efficient working of the heart. The term Heart disease refers to disease of heart & blood vessel system within it. There are number of factors which increase the risk of Heart disease:

Family history of heart disease

- Smoking
- Cholesterol
- Poor diet
- High blood pressure
- High blood cholesterol
- Obesity
- Physical inactivity
- Hyper tension

Symptoms of a Heart Attack

- Discomfort, pressure, heaviness, or pain in the chest, arm, or below the breastbone.
- Discomfort radiating to the back, jaw, throat, or arm.
- Fullness, indigestion, or choking feeling (may feel like heartburn).
- Sweating, nausea, vomiting, or dizziness.
- Extreme weakness, anxiety, or shortness of breath.
- Rapid or irregular heartbeats

Related Work

Keerthana T K [6] proposed a HDPS based on three dissimilar data mining techniques. The different data mining methods used are Naive Bayes, Decision tree (J48), Random Forest and WEKA API. The system can envisage the likelihood of patients getting a heart disease by using medical profiles such as cholesterol, age, sex, blood pressure and blood sugar. Also, the performance will be compared by computation of confusion matrix. This can help to determine accuracy, precision, and recall. The overall system offers high performance and better accuracy. Rupali and R. Patil [7] developed an intellectual Heart Disease Prediction System. This system is constructed using Naive Bayes algorithm and it also uses a smoothing technique (Jelinek mercer smoothing) to improves the performance. Here the system is proposed using Cleveland heart disease database as the input dataset. Each attribute of the dataset were fed to the Naive Bayesian classifier and it produces the prediction results based on the classification process. We can conclude that efficiency can be improved with the use of smoothing technique. This model could answer complex queries which traditional decision support systems cannot. Ankita Dewan, Meghna Sharma [8] developed a prototype which can find out and mine unknown knowledge (patterns and relations) related with heart disease from a precedent heart disease database record. It can resolve convoluted queries for sensing heart disease and therefore support medical practitioners to make elegant clinical decisions which conventional decision support systems were not competent to. By providing proficient treatments, it can help to decrease costs of treatment. B. Venkatalakshmi, M.V Shivsankar [9] proposed and extend diagnosis and forecast system for heart diseases using predictive mining. Different experiment has been conducted to estimate the performance of dissimilar predictive data mining methodology comprising a Decision tree and Naïve Bayes algorithms. In this anticipated work, a 13 attribute ordered medical database from UCI Machine Learning Repository has been used as a resource data. Decision tree and Naive Bayes have been implemented and their performance on identification has been estimated. Naïve Bayes outperforms when compared to Decision tree.

S. U. Amin, K. Agarwal, and R. Beg, [10] applied a hybrid system that utilizes global optimization assistance of genetic algorithm for commencing of neural network weights. The prophecy of the heart disease is based on vulnerability factors such as family history, age, diabetes, high cholesterol, smoking, hypertension, alcohol intake and obesity.

Data Mining Techniques

Dissimilar Data Mining algorithms and techniques such as Decision Trees, Genetic Algorithm, Regression, Artificial Intelligence, Neural Networks, Nearest Neighbor method, Classification, Clustering, Association Rules, etc., are used for knowledge discovery from databases. Some data mining techniques are describing below [11]:

Decision Tree

The management of the Decision Tree technique [12, 13] in the cure of heart disease have been scrutinized by the researchers with significant success. Decision hierarchy is a tree-like organization, which consists of internal nodes, branches and leaf nodes, in which every branch denotes an attribute value, each interior node denoted a test on an attribute which is used for and a leaf node represents the predicted classes or class allocations. The classification starts from the root node, then traverses the tree based on the predictive attribute value. The methodology involves data partitioning, data classification, decision tree category selection, and the request of reduction of fault trimming to create trimmed decision trees. Classification methods are labeled as supervised and unsupervised approaches. The supervised classification approaches contain chi merge and entropy while the unsupervised methods include identical width and identical frequency. The data partitioning entails testing with or lacking of voting. Three Decision Tree types are tested: Gini Index, Information Improvement, and Gain Ratio. Finally, reduced error trimming is useful to provide more closed decision rules. The Figure 1 demonstrated the accomplishment of ID3 algorithm on patient data.

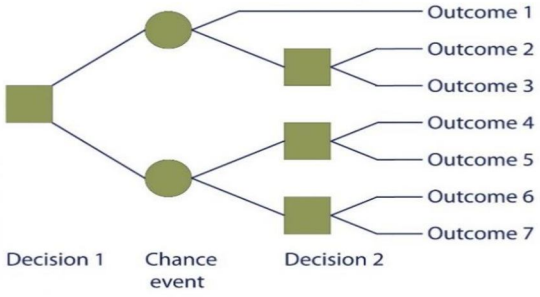

Fig. 1: Decision Trees

Bayesian Classifier

Using Bayesian classifiers, the system can ascertain the concealed knowledge associated with diseases from historical records of the patients having heart disease. Bayesian classifiers envisages the rank membership probabilities, in a manner that the prospect of a specified sample belongs to a fastidious class statistically. Bayesian classifier is based on Bayes' theorem. We can employ Bayes theorem to resolve the possibility that a proposed diagnosis is acceptable, given the scrutiny. A simple probabilistic, the naive Bayes classifier is utilized for categorization based on which is based on Bayes' theorem. According to naïve Bayesian classifier the occurrence (or incurrence of a meticulous feature of a class is considered as independent to the presence (or absence) of any other feature. When the dimension of the inputs is high and more efficient result is expected, the chief Naïve Bayes Classifier technique [14, 15] is applicable. Naïve Bayes model identifies the physical characteristics and features of patients suffering from heart disease. For each input it gives the prospect of attribute for the expectable state. The Figure 3.2 shows the accomplishment of Naive Bayes algorithm on patient data.

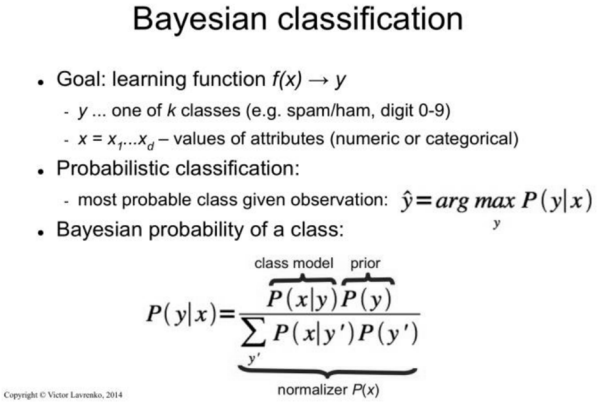

Fig. 2: Naïve Bayes algorithm on patient data

Neural Network

In practical applications, neural networks are well known to generate highly accurate results. By using nourish forward neural network (NN) model [16], inconsistent learning rate and backpropagation learning algorithm with momentum, the neural multifaceted is trained with Heart Diseases database.

The design of the model is as follows: It starts with the input of clinical data and progresses to develop ANN algorithm. After training model, it can produce the prediction results. The computational steps of neural network algorithm begin with the classification of clinical data into two equal parts randomly. One is used for testing and additional is used for training. An initial weight is assigned to each feature randomly. The calculated errors are used to adjust the weight of all features. Every feature's final weight is found out when the errors meet with the termination conditions. The process is repetitive for number of times. After constructing the training models, we can calculate the performance results from the testing data. The Figure 3 shows the accomplishment of neural network algorithm on clinical data.

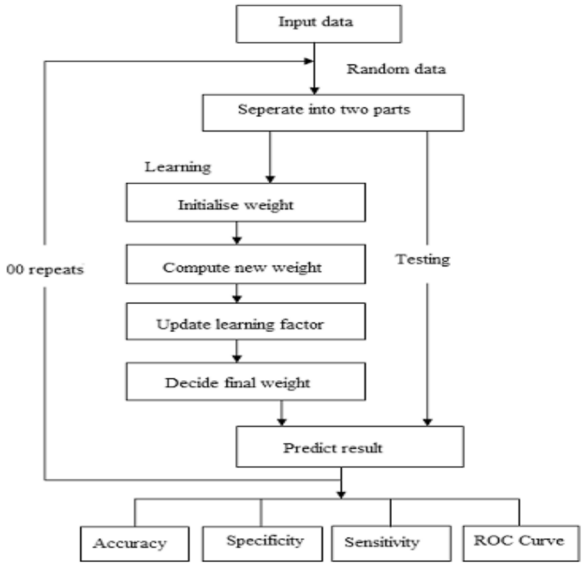

Fig.3: Implementation of neural network (NN) algorithm on clinical data

2.1 Clustering

Clustering is a process of partitioning a set of data (or objects) into a set of meaningful sub-classes, called clusters. It helps users to understand the natural grouping or structure in a data set. Clustering is an unsupervised classification and has no predefined classes. They are used either as a stand-alone tool to get insight into data distribution or as a pre-processing step for other algorithms. Moreover, they are used for data compression, outlier detection, understand human concept formation. Some of the applications are Image processing, spatial data analysis and pattern recognition. Classification via Clustering is not performing well when compared to other two algorithms.[9]

Support Vector Machine

Support Vector Machine (SVM) is a category of universal feed forward networks like Radial-basis function networks, pioneered by Vapnik. SVM can be used for pattern classification and nonlinear regression. More precisely, the support vector machine is an approximate implementation of the method of structural risk minimization. This principle is based on the fact the error rate of a learning machine on test data is bounded by the sum of the training-error rate and term that depends on the Vapnik-Chervonenkis (VC) dimension. V The support vector machine can provide good generalization performance on pattern classification problem

[17]. Optimal Hyperplane for patterns: Consider the training sample $\{(x_i, y_i)\}_{i=1}^{N}$ where x_i is the input pattern for the i^{th} instance and yi is the corresponding target output. With pattern represented by the subset yi= +1 and the pattern represented by the subset yi= -1 are linearly separable. The equation in the form of a hyperplane that does the separation is

$$w^T x + b = 0 \quad (1)$$

Where x is an input vector, w is an adjustable weight vector, and b is a bias. Thus,

$$w^T x_i + b \geq 0 \text{ for } y_i = +1 \quad (2)$$

$$w^T xi + b < 0 \text{ for } y_i = -1 \quad (3)$$

For a given weight vector w and a bias b, the separation between the hyperplane defined in eq. 1 and closest data point is called the margin of separation, denoted by ρ as shown in figure 4, the geometric construction of an optimal hyperplane for a two- dimensional input space.

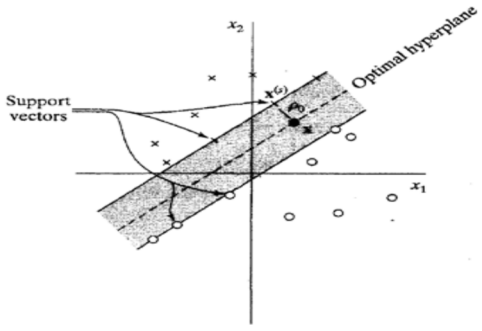

Fig.4: Optimal Hyperplane for a two tier input space

The discriminant function gives an algebraic measure of the distance from x to the optimal hyperplane for the optimum values of the weight vector and bias, respectively.

$$g(x) = w_o T_x + b_o \qquad (4)$$

Proposed Methdology

To assess the severity of heart disease risk using the data mining models, this thesis proposes the following architecture specified in figure 4.2 which include pre-processing, preparing, training and testing with individual models, evaluation of results and the prediction of heart disease risk. The proposed work is implemented in MATLAB2012a.

Heart Disease Dataset

This record comprises 76 attributes, except all published experiments refer to using a subset of 14 of them. In meticulous, the Cleveland record is the simply one that has been used by ML researchers to this date. [38] The "end" field signifies to the occurrence of heart illness in the patient. It is integer valued from 0 (no presence) to 4. Experiments with the Cleveland record have concentrated on simply attempting to discriminate presence (values 1, 2, 3, 4) from absence (value 0).

Only 14 attributes used:

1. #3 (age)	2. #4 (sex)	3. #9 (cp)
4. #10 (trestbps)	5. #12 (chol)	6. #16 (fbs)
7. #19 (restecg)	8. #32 (thalach)	9. #38 (exang)
10. #40 (oldpeak)	11. #41 (slope)	12. #44 (ca)
13. #51 (thal)	14. #58 (num) (the predicted attribute)	

Data Pre-processing

This step incorporates taking out data from the UCI repository in a recognized format. At this point data is transformed by involving the evacuation of missing fields, evacuation of outliers and normalization of data. The missing attributes are handled by entering mean.

Training and Testing the Models

Principal Component Analysis

Principal component analysis (PCA) is a standard tool in modern data analysis. It is an uncomplicated non parametric method for extracting pertinent information from confusing data sets. Principal components analysis (PCA) technique is used for accomplishing the generalization and produces a novel set of variables called principal components. Each principal constituent is a linear arrangement of the original variables. All the principal components are orthogonal to each supplementary so there is no superfluous information. The principal components as a whole form an orthogonal basis for the space of the data. The procedure can be followed in many ways i.e. a) Using curious value decomposition method (SVD) b) using the covariance matrix method. In this work they have used MATLAB software for executing the principal components [33].

k-Nearest Neighbor Classifier

K nearest neighbor (KNN) is an uncomplicated algorithm, which supplies each case and classifies novel cases based on comparison measure. KNN algorithm also known as 1) case based reasoning 2) k nearest neighbor (knn) 3) example based reasoning 4) instance based learning 5) memory based reasoning 6) sluggish learning [4]. KNN algorithms have been utilized since 1970 in diverse applications analogous to statistical estimation and pattern recognition etc. KNN is a non parametric classification method which is commonly classified into two types:

1) NN techniques of Structure less
2) NN techniques based on Structure.

In construction less NN methods/techniques whole data is categorized into training and test assessment data. From training point to model point distance is anticipated, and the point with modest distance is called nearest neighbor. Structure based NN methods are based on structures of records similar to orthogonal structure tree (OST), nearest future line k-d tree, ball tree, axis tree and central line [12]. Nearest neighbor classification is used chiefly when all the attributes are continuous. Simple K-nearest neighbor (k-NN) algorithm is exposed in figure 5.

Steps 1) find the K training instances which are closest to unknown instance
Step2) pick the most commonly occurring classification for these K instances

Fig 5: K nearest neighbor algorithm

In entropy technique, the attribute which minimizes entropy and maximizes information gain is selected as the tree root. To select tree root, it is first necessary to calculate the information gain of each attribute. Then, the attribute maximizing information gain should be selected. Information gain, or entropy, is derived from eqn.1.

Entropy H(X) = $-p_1\log_2 p_1 - p_2\log_2 p_2 - p_3\log_2 p_3 - \ldots\ldots\ldots\ldots - p_n\log_2 p_n$ ***eq(1)***

$= -\sum_{i=1}^{m} pi\ log2\ pi$

In the equation 1, the class-wise probability has been settled then entropy has been calculated of each individual attributes.

Then gain was calculated as follows:

Gain = Entropy(X) - Entropy (X|Y) ……eq (2)

So as per the above process feature reduction has been done, where gain was higher than that attribute has been qualified for the process and less gain was reduced from dataset.

ALGORITHM STEPS

Read a heart dataset and stored into separate variable *x1*
Make X1 reduced datasets from a database.
Set a learning algorithm to individual pattern into selected dataset.
Suppose $x_1, x_2, \ldots x_M$ are Nx1vectors
Step 1: $\overline{x} = \frac{1}{M}\sum_{i=1}^{M} x_i$
Step 2: Subtract the mean: $\Phi_i = x_i - \overline{x}$
Step 3: From the matrix $A = [\Phi_1 \Phi_2 \ldots \Phi_M]$(N *M matrix), then compute:

$C = \frac{1}{M}\sum_{N=1}^{M} \Phi_n \Phi_n = AA^T$

(Sample covariance matrix, N*N, characterizes the scatter of the data)

Step 4: Compute the Eigen values of C: $\lambda_1 > \lambda_2 > \cdots > \lambda_N$
Step 5: Compute the eigenvectors of C: $u_1, u_2, \ldots u_N$

Since C is symmetric, $u_1, u_2, \ldots u_N$ form a basis, (i.e., any vector x or actually$(x - \overline{x})$, can be written as a linear combination of the eigenvectors):

$$(x - \overline{x}) = b_1 u_1 + b_2 u_2 + \cdots + b_N u_N = \sum_{i=1}^{N} b_i u_i$$

Step 6: (feature selection step) keep only the terms corresponding to the *K* largest Eigen values:

$$x - \overline{x} = \sum_{i=1}^{K} b_i u_i$$

To choose *K*, use the following criterion

$\frac{\sum_{i=1}^{K} \lambda_i}{\sum_{i=1}^{N} \lambda_i}$> Threshold (e.g., 0.9 or 0.95)

Step 7: Determine parameter K = 5 number of nearest neighbors //they have assumed K=5
Step 8: Here apply Euclidean distance

Distance functions

Euclidean	$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Calculate the distance between the query record and all the training samples
Step 9: Sort determined data as per the distance and determine nearest neighbors based on the K^{th} minimum distance
Step 10: For each normalized sample the query instance is (att_i, m X n), instead of calculating the distance using step 9
Step 11: Sort the distance Step 10
Step 12: Rank minimum distance
Step 13: It will be included n-Nearest neighbors
Step 14: If the rank of K > 5 then that attribute value will not be included into the classified result.
Step 15: ELSE
Step 16: Use majority of the same category of nearest neighbors predicted value of the query instance
Step 17: We have 4 classes except normal condition, the 4 classes are namely goal1, goal2, goal3 and goal4 and on the same classes classification had done on X1 feature selected data set.
Step 18: Finally result measurement is as follows: specificity, sensitivity and accuracy.
Here block diagram shows that the working of proposed approach, where at initial state health care dataset is selected for the processing, then into next stage entire dataset is logically separate for the moment due to it is containing string fields as well as numeric fields, so in the designing approach they developed separate approach for string and numeric data.
Pre-processing: It converts the data which is more reliable for unsupervised learning by removing the labels from the dataset.
Data fraction: Preprocessed data are used to partition into training & testing sets samples.
Detection of Normal: in this step normal data is separated from the training data sample, here training process is done by training and normalize using min-max.
And if the normal class has been easily detected then its goes to the separately normal class otherwise if not detected then it will go to the KNN-PCA classifier.
And in this process each class has been accurately predicted with their own identity, after successful prediction the result analysis approach follows for the detected intrusions.

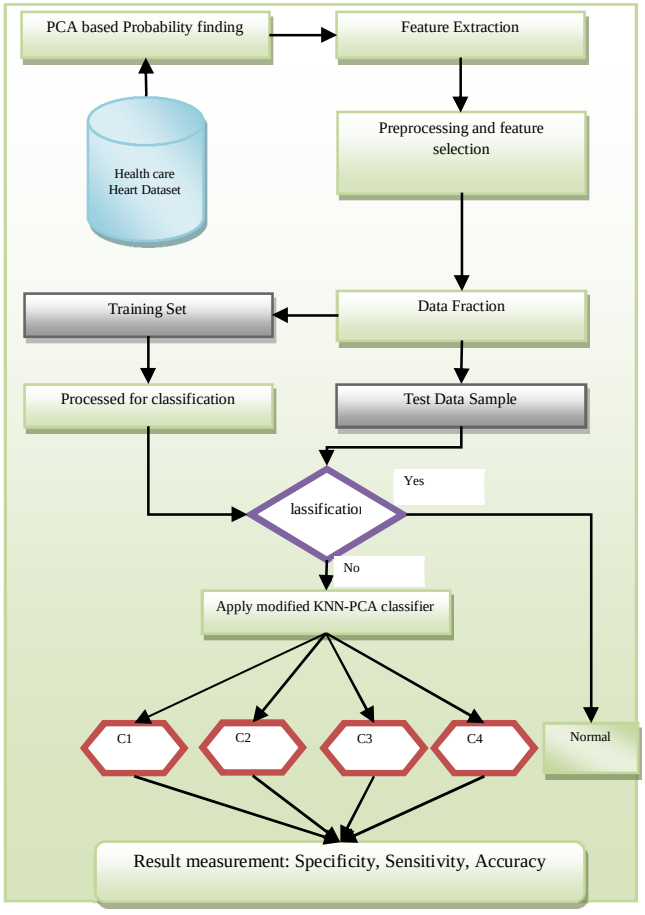

Figure 5: Block Diagram of proposed methodology

Experimental Results

In this section, we present the results from our extensive experiments to compare the performance of M.KNN, SVM and 2tier proposed method on real health care data from a Chinese city. All the experiments are conducted on the MATLAB platform, which includes three Intel 3.4 GHz machines, each running 16GB RAM.

This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1, 2, 3, 4) from absence (value 0).

Only 14 attributes used [18]

Evaluation Criteria

To evaluate the performance of the proposed technique, three metrics are used i.e., accuracy, specificity, sensitivity. Reference to the current study of heart disease diagnosis, accuracy is the calculation of the discriminating results between the patients and healthy subject classes. If the results of classification do not provide correct discrimination between alternative states of health, then the accuracy is not significant while correct discrimination provides high accuracy. The classification probability of the patient is called the sensitivity. The incorrect classification as patient class may also be viewed from this measure. The specificity shows the classification probability for the healthy subjects. It gives the information of patients misclassified as normal subject as well.

Sensitivity =TP/(TP+FN)

Specificity =TN/(TN+FP)

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

Where TP is true positive and here it represents all the stats for heart patient classified as heart patient, TN is true negative and here it is collection of classification results for normal subject classified as normal subject. False positive (FP) shows the sum for all those classification results where the subjects were normal and classified as heart patient where as false negative (FN) shows the stats about all those subjects who were heart patient and system classified them as normal. The above discussion and relations clearly demonstrate that the three evaluation measurements i.e., accuracy, sensitivity and specificity, are quite enough to show the quality of the classification results.

GUI Environment

This section shows the main GUI environment of the anticipated methodology and probability entropy gain process of it.

Result Analysis

The result analysis of the proposed work is performed using the accuracy and specificity and sensitivity here figure 6 shows that the main gui environment of the implemented system in MATLAB2012A.

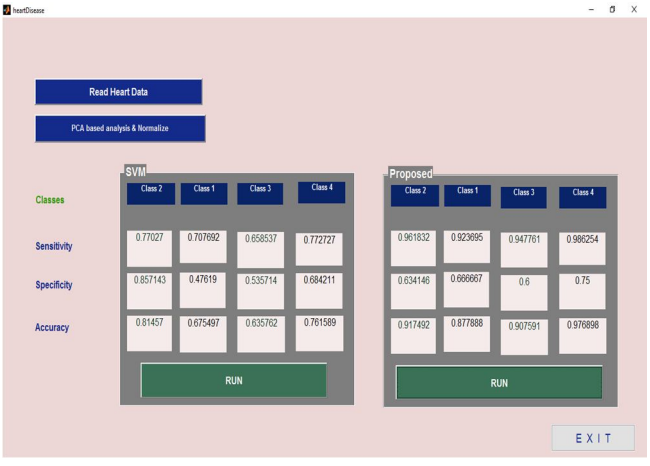

Fig. 6: Main GUI of the Proposed Methodology

Accuracy Analysis

For this parameter the comparison between SVM, projected PCA-k-NN method is perform in which it is found that the accuracy rate of SVM is about 71%, and our proposed method is about 90% which means our method generates better accuracy rate than the existing SVM method.

Table 1: Accuracy result analysis of the projected PCA-KNN method

	SVM	Proposed
class2	0.81457	0.917492
class1	0.675497	0.877888
class3	0.635762	0.907591
class4	0.761589	0.976898

Fig. 7: Accuracy graph between SVM and proposed PCA-k-NN Method

Specificity Analysis

For this parameter the comparison among SVM and anticipated method is perform in which it is found that the accuracy rate of SVM is about 61%, and our method PCA-k-NN is about 54% which means our method generates better specificity rate than the existing SVM method.

Table 2: Specificity result analysis of the SVM Proposed PCA-k-NN method

	SVM	Proposed
class2	0.857143	0.634155
class1	0.47619	0.666667
class3	0.535714	0.6
class4	0.635762	0.75

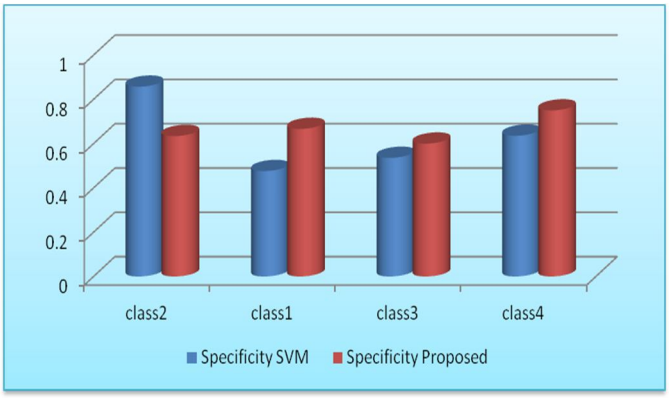

Fig. 8: Specificity graph between SVM Proposed PCA-k-NN method

Senstivity Analysis

For this parameter the comparison between SVM and proposed method is perform in which it is found that the accuracy rate of SVM is about 71%, and PCA-kNN is about 95% which means our method generates better sensitivity rate than the existing SVM method.

Table 3: Sensitivity result analysis of the SVM Proposed PCA-k-NN method

Sensitivity		
Class/Method	SVM	Proposed
class2	0.77027	0.961832
class1	0.707692	0.923695
class3	0.658537	0.947761
class4	0.772727	0.986254

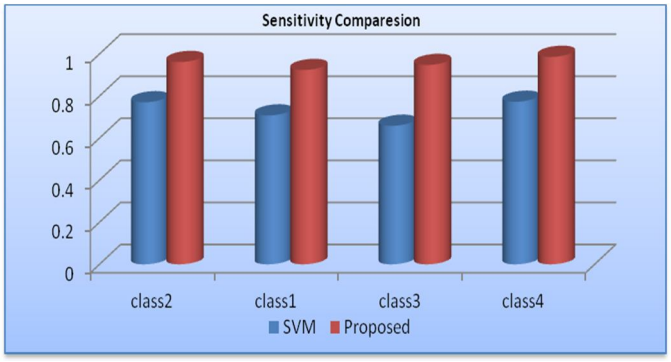

Fig. 9: Senstivity graph between SVM Proposed PCA-k-NN method

Conclusion

Heart disease is the main causes of death over the world. It represents 7.2 million death, i.e., 12.8% of fatalities on the world. Although cardiovascular disease have been recognized as the main source of death in the previous decade, they are the most preventable and controllable disease in the meantime. Deaths from cardiovascular maladies demonstrate a regularly expanding pattern. Then again, their initial determination assumes an imperative part in enhancing patients' well being status and diminishing fatalities. Thusly, this examination expected to help doctors to early analyze such ailments and evaluate coronary illness hazard factors in considered people.. This paper introduced a PCA-k-NN based approach for ordering coronary illness. As an approach to verify the anticipated strategy, this is tried on UCI repository machine learning dataset. We have quite recently chosen just 14 dataset among 76 dataset for the prescience of coronary illness quiet. This anticipated forecast display helps the specialists in capable heart disease finding or discovery process with fewer characteristics. This disease is regularly found in India and in Andhra Pradesh. The investigation of anticipated technique is performing utilizing sensitivity, specificity and accuracy or exactness. The simulation result of accuracy parameter of our anticipated technique is around 90% which an excessive amount of more noteworthy than SVM strategy. This method is relied upon to be actualized in future on a limited dataset with non aggressive indices in general. This, thusly, forces bring down expenses and difficulties on patients.

References

[1] Heon Gyu Lee, Ki Yong Noh, Keun Ho Ryu, “Mining Biosignal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV,” LNAI 4819:

[2] Chhikara, S & Sharma,P Data Mining Techniques on Medical Data for Finding Locally Frequent Diseases, I JRASET 2014,PP 396-402.

[3] Cody, W.F. Kreulen, J.T. Krishna, V. & Spangler, W.S. (2002). The integration of business intelligence and knowledge management. IBM Systems Journal, 41(4), 697-713

[4] Ceusters, W. (2001). Medical natural language understanding as a supporting technology for data mining in healthcare. In Medical Data Mining and Knowledge Discovery, Cios, K. J. (Ed.), PhysicaVerlag Heidelberg, New York, 41-69.

[5] Megalooikonomou, V. & Herskovits, E.H. (2001). Mining structure function associations in a brain image database. In Medical Data Mining and Knowledge Discovery, Cios, K. J. (Ed.), Physica-Verlag Heidelberg, New York, 153-180.

[6] Keerthana T K “ Heart Disease Prediction System using Data Mining Method”, International Journal of Engineering Trends and Technology (IJETT) – Volume 47 Number 6 May 2017.

[7] Ms. Rupali R. Patil "Heart Disease Prediction System using Naïve Bayes and Jelinek-mercer smoothing" IJARCCE 2014.

[8] Ankita Dewan, Meghna Sharma “Prediction of Heart Disease Using a Hybrid Technique in Data Mining Classification”, In proceeding of IEEE 2015.

[9] B.Venkatalakshmi, M.V Shivsankar “Heart Disease Diagnosis Using Predictive Data mining”, International Conference on Innovations in Engineering and Technology (ICIET’14) On 21st&22ndMarch, Volume 3, Special Issue 3. In proceeding of IJIRSET.

[10] S. U. Amin, K. Agarwal, and R. Beg, "Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors," in Proceedings of 2013 IEEE Conference on Information and Communication Technologies (ICT 2013), 2013, no. ICT, pp. 1227–1231.
[11] Mayuri Takore, Prof.R.R. Shelke "Data Mining Techniques to Find Out Heart Diseases: An Overview", International Journal for Research in Applied Science & Engineering Technology (IJRASET).
[12] Komal G, Vekariya V. Novel approach for heart disease prediction using decision tree algorithm. International Journal of Innovative Research in Computer and Communication Engineering. 2015; 3(11):11544–1.
[13] Shouman M, Turner T, Stocker R. Using decision tree for diagnosing heart disease patients. Proceedings of the 9th Australasian Data Mining Conference (AusDM'11); Ballarat, Australia. 2011. p. 23–30.
[14] Liu X, Lu R, Ma J, Chen L. Privacy-preserving patient-centric clinical decision support system on naïve bayesian classification. IEEE Journal of Biomedical and Health Informatics. 2016; 20(2):655–88.
[15] Pattekari SA, Parveen A. Prediction system for heart disease using naive bayes. International Journal of Advanced Computer and Mathematical Sciences. 2012; 3(3):290–4.
[16] Rani KU "Analysis of heart diseases dataset using neural network approach" IJDKP. 2011; 1(5):1–8.
[17] Christopher J.C. Burges. "A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery" Springer, 2(2), pp.121-167, 1998.
[18] Blake, C.L., Mertz, C.J.: "UCI Machine Learning Dataset", http://mlearn.ics.uci.edu/databases/heartdisease/,2004.