Diagnosis and Identification of Risk Factors for Heart Disease Patients Using Generalized Additive Model and Data Mining Techniques

Sabyasachi Mukherjee[1], Suman Kapoor[2], Proloy Banerjee[3]

[1]Department of Mathematics; NSHM Knowledge Campus, Durgapur-713212, West Bengal , INDIA.
[2]Senior Bio-statistician, Department of Bio-Statistics, Quintiles Clinical Research, Bangalore, Karnataka, INDIA.
[3]Research Scholar, Department of Statistics, Aliah University, Kolkata, West Bengal, INDIA.

ABSTRACT

Background: Yearly death rate is increasing due to heart disease. Major factors for the increasing death rate due to heart disease are (a) misdiagnosed by the medical doctors or (b) ignorance by the patients. Heart diseases can be described as any kind of disorder which affects the heart. **Methods:** The dataset of 'statlog' from the UCI Machine Learning with 270 patients related to heart disease isused in this article. The dataset comprises attributes of patients diagnosed with heart diseases. The diagnosis was used to confirm whether heart disease is present or absent in the patient. The present article aims to identify the risk factors/variables which influence this diagnosis. Classification is a very important part of the disease diagnosis but it is also relevant to identify the risk factors/variables. Two classification techniques namely Support Vector Machines (SVM), Multi-Layer Perceptrons ensembles (MLPE) and one advanced regression technique,Generalized additive model (GAM) with binomial distribution and'logit' link have been introduced for diagnosis and risk factors/variables identification. **Results:** GAM explains 65% deviance with adjusted R square value 0.70 approximately. Sensitivity analysis has been performed under SVM, which is the best model for this dataset with approximately 85% classification accuracy rate. MLPE gives 82% classification accuracy rate approximately.Maximum heart rate, vessel, old peak, chest pain, thallium scan are the most important factors/variables find through both sensitivity analysis under SVM and GAM. **Conclusion:** The present article attempt to remove some new information regarding heart disease through probabilistic modeling which may provide better assistance for treatment decision making using the individual patient risk factors and the benefits of a specific treatment. These findings may help the medical practitioners for better medical treatment.

Key words: Heart disease, Data Mining, SVM, MLPE, Sensitivity analysis, GAM.

Correspondence
Sabyasachi Mukherjee
Assistant Professor in Statistics; Department of Mathematics; NSHM Knowledge Campus, Durgapur-713212, West Bengal, INDIA.

Ph.no: 919775745309
E-mail address: sabyasachi99@gmail.com

Submission Date: 02-06-2017;
Revision Date: 12-07-2017;
Accepted Date: 08-09-2017.
DOI : 10.5530/jcdr.2017.4.31

INTRODUCTION

The heart is the most essential organ of human body which also can be described as the size of a fist and a strong muscle in the body. Any disorderliness that affects the heart from infection to genetic defects and blood vessel disease is referred to as heart disease.[1] Heart disease is a serious disease and proper diagnosis of heart disease at early stage remains challenging task.[2] In fact, up to 25% of people with heart disease have no symptoms despite insufficient blood flow to the heart, a condition that is referred to as silent heart disease.[3] In the United State of America about 600,000 people die as a result of heart disease every year which is calculated to be one in every four deaths.[4] Diagnosis usually appears when a patient visits the doctor to have symptoms checked out. Patients may be met with shortness of breath, pain in the chest or back, painful, persistent coughing or any number of other symptoms, none of which immediately alert the doctor to a diagnosis of heart disease. Many studies were carried out about heart disease diagnosis in all over the world generally using by artificial intelligence techniques or data mining methods.[5-8] The use of data mining techniques in medical diagnosis has been increasing gradually. There is no doubt that evaluations of data taken from patients and decisions of experts are the most important factors in diagnosis. However, sometimes different artificial intelligence techniques or machine learning techniques are used for disease diagnosis.[5-9-11]

In health care, data mining or statistical machine learning plays a vital role in the medical applications including diagnosis, prognosis, and therapy.[12] Clinical data mining involves the conceptualization, extraction, analysis, and interpretation of the available clinical data for practical knowledge-building, clinical decision making, and partition reflection.[12] A medical diagnosis is a classification problem[13] In the predictive data mining, the data set consists of instances, each instance is characterized by attributes or features and another special attribute represents the outcome variable or the class.[14] Often, the goal of any data mining project is to build a model from the available data. Thus, data mining models are objective models rather than subjective since it is driven by the available data.

Data mining (DM) techniques[15] aim at extracting high-level knowledge from raw data. There are several DM algorithms, each one with its own advantages. DM techniques perform regression and classification tasks. In case of neural networks (NNs), the back propagation algorithm was first introduced in 1974[16] and later popularized in 1986.[17] Since then, neural networks (NNs) have become increasingly used. More recently, support vector machines (SVMs) have also been proposed.[18,19] Due to their higher exibility and nonlinear learning capabilities, both NNs and SVMs are gaining an attention within the DM field, often attaining high predictive performances.[20,21] SVMs present theoretical advantages over NNs, such as the absence of local minima in the learning phase. In effect, the SVM was recently considered one of the most influential DM algorithms.[22] Therefore in this paper, a study of SVM on heart disease diagnosis was realized.

In the statistical analysis of clinical trials and observational studies, the identification and adjustment of prognostic factors is an important activity in order to get valid outcome. The failure to consider important prognostic variables, particularly in observational studies, can lead to errors in estimating treatment differences. In addition, incorrect modeling of prognostic factors can result in the failure to identify nonlinear trends or threshold effects on survival. This article describes flexible statistical methods that may be used to identify and characterize the effect of potential prognostic factors on disease endpoints. These methods are called 'Generalized Additive Models' (GAM).[23] Many mathematical and statistical methodologies for building classification models, from the classical statistical methods to machine learning theory to classification trees, are reviewed and compared.[24-27] Many work and research has been done into better and accurate models for the Heart Disease Dataset. The work[28] gives a knowledge driven approach. Initially Logistic Regression was used by Dr. Robert Detrano for heart disease diagnosis.[29] Newton Cheung utilized C4.5, Naive Bayes, BNND and BNNF algorithms and reached the classification accuracies of 81.11%, 81.48%, 81.11% and 80.96%, respectively.[30] proposed a method that uses artificial immune system (AIS) and obtained more classification accuracy than the previous works.[31] shows comparative results of many study performed on this heart disease data.[10] In this present article 10-flod cross-validation along with 5 runs in each experiment has been performed for getting more stability in classification accuracy rate. Aim of the present article is to explore a relationship between chance of having heart disease of a patient with others biomedical parameters as a cofactors. Due to complex relationship between cofactors and response variable, GAM has been introduced here for better accuracy in prediction. The another aim of this study is to find a best classifier which gives a good performance evolution measures and also try to find the important input variables for heart disease diagnosis using strong data mining techniques. Many authors had used various classification techniques to this dataset for heart disease diagnosis.[5-11] but probably, SVM and MPLE are not been used under proper modeling scheme. This study shows high classification accuracy rate and presented a significant variable input importance chart for heart disease diagnosis.

In this research work, we used the heart disease dataset obtained from the UCI Machine Learning to develop intelligent systems using data mining and GAM for diagnosis of heart disease. The results obtained from these systems were compared and the highest recognition rate obtained was taken as the best system for diagnosis of heart disease. This system will solve the problem of misdiagnose of heart disease and also try to identify the risk or important biomedical parameters responsible for probable heart disease. This can guide the doctors about prognostic factors and patients for greater awareness regarding heart disease.

MATERIALS AND METHODS

MATERIALS

The present article is considered 270 heart disease patients with 14 factors or variables. The current secondary data set is taken from the report. The data set can be downloaded at http://archive.ics.uci.edu/ml/datasets.html. Description of the covariates, factors and their levels are described in Table 1. The summarized statistics such as the mean, standard deviation, and proportion of the levels are given in Table 1. The current data contains 5 continuous variables and 9 attribute characters. The description of each variable or attribute character, attribute levels, and how they are operationalized in the present report is displayed in Table 1. Here present or absent of heart disease in patient is playing a role of dependent variable (for regression) or output variable (for classification) and rest of the variables are playing the role of independent variables/ cofactors.

METHODS

In this present article data mining techniques with sensitivity analysis is performed for diagnosis of the heart disease and tried to find out the important factors which are most responsible in this diagnostic work respectively. Apart from this, the generalized additive logistic models are also applied to find the risk factors for heart disease. In case of data mining Multi-Layer Perceptrons ensembles (MLPE), Support vector machines (SVM) are used for classification and there after Sensitivity analysis done only upon the best model out of this classifier for this heart disease data set.[20]

Best GAM[32] model can be selected through some model checking criteria namely R square value, AIC or UBRE value and regression diagnostic plots like normal probability plot, Residuals against fitted value plot etc.[14,32] Cofactors are significant or not judged through p-value. For this heart disease data set Absence and presence of heart disease is taken as response variable (Y), and Age, Sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting ECG results, maximum heart rate achieved, exercise induced angina, oldpeak, slope of the peak exercise ST segment, number of major vessels, thal (thallium scan) are the cofactors (X_i's).

Data mining techniques want to classify the data using different classifiers whereas GAM wants to identify the risk factors for this disease. The brief descriptions of the used methods are given below.

Data Mining Techniques

DM is an iterative process that consists of several steps. The CRISP-DM,[33] a tool-neutral methodology supported by the industry (e.g. SPSS, DaimlerChryslyer) partitions a DM project into 6 phases: 1. business understanding; 2. data understanding; 3. data preparation; 4. modeling; 5. evaluation; and 6. deployment.

This work addresses steps 4 and 5, with an emphasis on the use of NNs and SVMs to solve classification and regression goals. Both tasks require a supervised learning, where a model is adjusted to a dataset of examples that map *I* inputs into a given target. In case of classification models output a probability *p(c)* for each possible class *c*, such that $\sum_{c=1}^{N_c} p_c = 1$. For assigning a target class *c*, one option is to set a decision threshold $D \in 0,1$ and then output *c* if $p(c) > D$, otherwise return *c*. This method is used to build the receiver operating characteristic (ROC) curves. Another option is to output the class with the highest probability and this method allows the definition of a multi-class confusion matrix. For more details see.[34]

To evaluate a model for classification, common metrics are.[35] ROC area (AUC), confusion matrix, accuracy (ACC), true positive/negative rates (TPR/TNR). A classifier should present high values of ACC, TPR, TNR and AUC. The model's generalization performance is often estimated by the holdout validation (i.e. train/test split) or the more robust k-fold cross-validation.[14] The latter is more robust but requires around k times more computation, since k models are fitted.

MLPE neural network model

In DM techniques, NN means the popular multilayer perceptron (MLP). A major concern in their use is the difficulty to define the proper network for a specific application, due to the sensitivity to the initial conditions and to overfitting and underfitting problems which limit their generalization capability. A very promising way to partially overcome such drawbacks is the use of MLP ensembles (MLPE); averaging and voting techniques are largely used in classical statistical pattern recognition and can be fruitfully applied to MLP classifiers. For classification problem MLPE are used, which is a combinations of MLP models. This network includes one hidden layer of *H* neurons with logistic functions (Figure 1 (a)). The overall model is given in the form:

$$y_i = f_i\left(w_{i,0} + \sum_{j=I+1}^{I+H} f_j\left(\sum_{n=1}^{I} x_n w_{m,n} + w_{m,0}\right) w_{i,n}\right) \qquad (1)$$

Where is the output of the network for node *i*, $w_{i,j}$ is the weight of the connection from node *j* to *i* and f_i is the activation function for node *j*.

For a binary classification ($N_c = 2$), there is one output neuron with a logistic function. Under multi-class tasks ($N_c > 2$), there are linear output neurons and the softmax function is used to transform these outputs into class probabilities:

$$p(i) = \frac{exp\ (y_i)}{\sum_{c=1}^{N_c} exp\ (y_c)} \qquad (2)$$

Where is the predicted probability and is the NN output for class *i*. The training (BFGS algorithm) is stopped when the error slope approaches zero or after a maximum of epochs. For classification it maximizes the likelihood.[14] Since NN training is not optimal, the final solution is dependent of the choice of starting weights. To solve this issue, the solution adopted is to train different networks and then select the NN with the lowest error or use an ensemble of all NNs and output the average of the individual predictions.[14] In general, ensembles are better than individual learners.[36] The final NN performance depends crucially on the number of hidden nodes. The simplest NN has $H = 0$, while more complex NNs use a high H value.

Support Vector Machine (SVM) model

When compared with NNs, SVMs present theoretical advantages, such as the absence of local minima in the learning phase.[14] The basic idea is transform the input $x \in \Re^I$ into a high m-dimensional feature space by using a nonlinear mapping. Then, the SVM finds the best linear separating hyperplane, related to a set of support vector points, in the feature space (Figure 1 (b)). The transformation (φ(x)) depends of a kernel function.

Here, SVM uses the sequential minimal optimization (SMO) learning algorithm adopting the popular Gaussian kernel, which presents less parameters than other kernels (e.g. polynomial): $K(X,X') = exp(-\gamma\|X - X'\|^2), \gamma > 0$. The classification performance is affected by two hyperparameters:, the parameter of the kernel, and C, a penalty parameter. The probabilistic SVM output is given by [37]

$$f(x_i) = \sum_{j=1}^{m} y_j \alpha_j\ K(x_j, x_i) + b$$

$$p(i) = 1/(1+exp\ (Af(x_i)+B)) \qquad (3)$$

Where m is the number of support vectors, $y_i \in \{-1,1\}$; is the output for a binary classification, and are coefficients of the model, and A and B are determined by solving a regularized maximum likelihood problem. When $N_c>2$, the one-against-one approach is used, which trains $N_c(N_c-1)/2$ binary classifiers and the output is given by a pairwise coupling.[37]

Sensitivity Analysis

The sensitivity analysis is a simple procedure that is applied after the training procedure and analyzes the model responses when a given input is changed. Let $y_{a,j}$ denote the output obtained by holding all input variables at their average values except x_a, which varies through its entire range ($x_{a,j}$, with j ϵ {1,2,…..L} levels). Variance (V_a) of $y_{a,j}$used as a measure of input relevance.[38] If $N_c>2$ (multi-class), it sets as the sum of the variances for each output class probability ($(p(c)_{a,j})$). A high variance (V_a) suggests a high x_a relevance, thus the input relative importance (R_a) is given by:

$$R_a = \frac{V_a}{\sum_{i=1}^{I} V_i \times 100(\%)} \qquad (4)$$

For a more detailed analysis, the variable effect characteristic (VEC) curve, Cortez *et al.* has been proposed, which plots the $x_{a,j}$ values (x-axis) versus the $y_{a,j}$ predictions (y-axis).[39]

Generalized Additive Model (GAM)

GAM[32,-40] is an extension of the Generalized Linear Model (GLM)[41] where the modeling of the mean functions relaxes the assumption of linearity, albeit additivity of the mean function pertaining to the covariates is assumed. Whilst the mean functions of some covariates may be assumed to be linear, the non-linear mean functions are modeled using smoothing methods, such as kernel smoothers, lowess, smoothing splines or regression splines. In general, the model has the following structure

$$g(\mu) = \alpha_0 + \sum_{j=1}^{p} f_j\,(X_j) \qquad (5)$$

where, $\mu=E(Y)$ for a response variable with some exponential family distribution, g is the *link* function and f_i are some smooth functions of the covariates X_i for each j=1,2,…..,p.

GAMs provide more flexibility than do GLMs, as they relax the hypothesis of linear dependence between the covariates and the expected value of the response variable. The main drawback of GAMs lies in the estimation of the smooth functions f_i, and there are different ways to address this. One of the most common alternatives is based on splines, which allow the GAM estimation to be reduced to the GLM context.[42] Smoothing splines,[43] use as many knots as unique values of the covariate X_i and control the model's smoothness by adding a penalty to the least squares fitting objective.[44,45]

Generalized additive models can be used in virtually any setting where linear models are used. For a single observation (i^{th})the basic idea is to replace $\sum_{j=1}^{p} x_{ij}\beta_j$, the linear component of the model with an additive component $\sum_{j=1}^{p} f_j\,(x_{ij})$.

In the logistic regression model the outcome y_i is '0' or '1' with '1' indicating an event and '0' indicates no event. (In this article '1' indicates absence of heart disease and '0'indicates presence of the heart disease in patient). Then the generalized additive logistic model assumes the log-odds are given below

$$log \frac{p(y_i\ |x_{i1},\ldots\ldots\ldots,x_{ip})}{1-p(y_i\ |x_{i1},\ldots\ldots\ldots,x_{ip})} = \beta_0 + f_1(x_{i1}) + \cdots + f_p(x_{ip}) \qquad (6)$$

Where $f_1,f_2,\ldots.f_p$ are the smooth functions which are estimated by splines algorithm. For more details see these references.[23-32]

Performance Evolution Measures

Classification Accuracy (ACC)

Classification accuracy refers to the ability of the model to correctly predict the class level of new or previous unseen data. Classification Accuracy is the percentage (%) of testing set examples correctly classified by the classifier. The quality of classification can be assessed through overall accuracy. That is

$$\text{Accuracy}(T) = \frac{\sum_{i=1}^{|T|} assess(t_i)}{|T|}, t_i \in T \qquad (7)$$

$$assess(t) = \left\{ \begin{array}{l} 1 iff classify\ (t) \equiv t.c \\ 0 otherwise \end{array} \right. \qquad (8)$$

Where T is the set data items to be classified (the test set in this case), t∈T,t.c is the class of item t, and (t) returns the classification of by the used classifier (here, SVM and MLPE). For more details see.[46]

Area under Curve (AUC)

AUC is a common evaluation metric for binary classification problems. Consider a plot of the true positive rate vs. the false positive rate as the threshold value for classifying an item as 0 or is increased from 0 to 1 and if the classifier is very good, the true positive rate will increase quickly and the area under the curve will be close to 1. One characteristic of the AUC is that it is independent of the fraction of the test population which is class 0 or class 1; this makes the AUC useful for evaluating the performance of classifiers on unbalanced data sets.

k-fold Cross Validation

k-fold cross validation is a common technique for estimating the performance of a classifier. Given a set of *m* training examples, a single run of *k*-fold cross validation proceeds as follows:

1. Arrange the training examples in a random order.
2. Divide the training examples into *k*-folds. (*k* chunks of approximately *m*/*k* examples each.)
3. For *i*=1,2,…..*k*:
   (i) Train the classifier using all the examples that do not belong to fold.
   (ii) Test the classifier on all the examples in fold.
   (iii) Compute, the number of examples in fold that were wrongly classified.
4. Return the following estimate to the classifier error:

$$E = \frac{\sum_{i=1}^{k} n_i}{m} \qquad (9)$$

To obtain an accurate estimate to the accuracy of a classifier, *k*-fold cross validation is run several times, each with a different random arrangement in Step- 1. After performing these steps several numbers of times takes an average of each run result to produced final classification accuracy. For more details see.[14]

All GAM regression and data mining works are performed in R statistical software with proper library packages.[40-47] (http://www3.dsi.uminho.pt/pcortez/rminer.html),[34]

RESULTS AND DISCUSSIONS

Table 2 presents the summarized results of Generalized Additive Model used for heart disease diagnosis. Here response variable is whether a patient has heart disease or not? Rest of the variables is cofactors. GAM has two parts of estimation methods; one is parametric estimation for those cofactors which entered in model parametrically and non-parametric estimation used for smoothing cofactors. In this present article only Age is the smoothing cofactors and rest are under parametric estimation method. The detailed results and interpretations of Table 2 (Binomial with logit link fitted model) are described as follows. The GAM regression coefficients give the change in the log odds of the Heart disease (response) for a one unit increase in the cofactors (predictor). Here we have considered the P-values up to approximately 10% level as significant, and more than 10% to approximately 20% as partially significant.[40,41-49,50]

Results of Estimation of Parametric coefficients

Heart disease (HD) is very high positively significantly associated with chest pain of a patient. Out of four types of chest pain, asymptomatic chest pain changes the log odds of HD by 2.7777 with p-value 0.0008.

Table 1: Operationalization of variables with the analysis & summarized statistics

Variable name	Operationalization	Mean	Standard deviation	Proportion of levels of Attributes
Age (Year)	Age at study	54.43	9.10	---
Sex	Gender : (Female = 1 ; Male = 2)	---	---	1= 32.22% ; 2= 67.78%
Chest Pain	Chest pain type (1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 4 = asymptomatic)	---	---	1= 7.41% ; 2=15.56% ; 3=29.26% ; 4=47.78%
Resting BP	Resting blood pressure (in mm Hg on admission to the hospital)	131.34	17.86	---
Cholesterol	Serum cholesterol in mg/dl	249.66	51.69	---
Fasting BS	Fasting blood sugar > 120 mg/dl (1 = False; 2 = True)	---	---	1= 85.19% ; 2=14.81%
Resting ECG	Resting electrocardiographic results (1 = Normal; 2 = Having ST-T; 3 = Hypertrophy)	---	---	1=48.52% ; 2=0.74% ; 3=50.74%
Max HR	Maximum heart rate achieved	149.68	23.17	---
Exercise Ang	Exercise induced angina (1 = No; 2 = Yes)	---	---	1=67.04% ; 2=32.96%
Oldpeak	ST depression induced by exercise relative to rest	1.05	1.14	---
Slope	The slope of the peak exercise ST segment (1 = Up sloping; 2 = Flat; 3 = Down sloping)	---	---	1=48.15% ; 2=45.19% ; 3=6.67%
Vessel	Number of major vessels (0-3) colored by fluoroscopy. (Treated as a discrete variable)	---	---	0=59.26% ; 1=21.48% ; 2=12.22%; 3=7.04%
Thal	Thallium heart Scan (1 = Normal; 2 = Fixed defect; 3 = Reversible defect)	---	---	1=56.30% ; 2=5.19% ; 3=38.52%
Heart disease	Diagnosis of heart disease (1= Absence; 2= Presence)	---	---	1=55.56% ; 2=44.44%

Table 2: Results for GAM of Heart disease data analysis using Binomial distribution with 'logit' link

		Estimation of Parametric coefficients		
Covariates	**Estimate**	**Standard Error**	**Z value**	**p-value**
Intercept	-6.644423	2.600914	-2.555	0.010629 *
Chest Pain 2	1.498281	0.963307	1.555	0.119862
Chest Pain 3	0.662778	0.824066	0.804	0.421237
Chest Pain 4	2.777748	0.829641	3.348	0.000814 ***
Cholesterol	0.009850	0.004513	2.183	0.029053 *
Max. HR	-0.032619	0.011325	-2.880	0.003974 **
Old peak	0.515073	0.223007	2.310	0.020906 *
Resting BP	0.024378	0.011871	2.053	0.040025 *
Resting ECG 2	2.187153	3.543705	0.617	0.537107
Resting ECG 3	0.768672	0.439692	1.748	0.080429.
Sex 2	2.080282	0.624856	3.329	0.000871 ***
Thal 2	0.063903	0.845742	0.076	0.939771
Thal 3	1.693988	0.477088	3.551	0.000384 ***
Vessel	1.263642	0.285799	4.421	<0.0001***
		Approximate Significance of smooth terms (Non-parametric)		
Smooth Covariate	**Edf**	**Ref. df**	**Chi.sq**	**p-value**
Age	8.1	8.593	14.18	0.0957.

Edf: Estimated degrees of freedom; **Ref.df:** Degrees of freedom before smoothing; **Chi. Sq:** Chi square value. **Significance Level:**'***' 0.001; '**' 0.01; '*' 0.05; '.' 0.1. **R-sq.(adj)** =0.697 ;**Deviance explained** = 64.3% ; **UBRE (Un biased risk estimator)** = -0.24238

Therefore, patient having higher chance of HD if he/she has asymptomatic chest pain.

In the GAM fitted model, for every one unit change in Cholesterol the log odds of HD increased by 0.0098 with p-value 0.029. Cholesterol has a positive significant association with HD which indicates that patients with high Cholesterol having a higher chance of HD.

HD is high negatively significantly associated with the Maximum Heart rate (Max.HR) of a patient. For every one unit change in Max.HR the log odds of HD decreased by 0.0326 with p-value 0.003. That means patients with maximum heart rate having lower risk of HD.

For one unit change in Old peak the log odds of HD increased 0.5150 with p-value 0.020.The HD is positively significantly associated with Old peak. Therefore patients with high Old peak value having higher risk of HD.

In this GAM fitted model, for every one unit change in Resting BP the log odds of HD increased by 0.0243 with p-value 0.040. Resting BP has a positive significant association with HD which indicates that patients with high Resting BP having a higher chance of HD.

Heart disease (HD) is positively significantly associated with Resting ECG of a patient. Out of three types of Resting ECG, Hypertrophy Resting ECG changes the log odds of HD by 0.7686 with p-value 0.080. Therefor patients having higher chance of HD if they have Hypertrophy Resting ECG result than others.

Sex (Gender) of a patient has a very positive significant association with HD. Male patient changes the log odds of HD by 2.0802 with p-value <0.001than a female patient. This indicates male patients having a higher chance of HD.

HD is very high positive significant association with Thallium heart scan (Thal) result. A patient with Reversible defect in his/her thallium heart scan report changes the log odd of HD by 1.6939 with p-value <0.001. It means patient has higher chance of HD if his/her thallium heart scan report shows Reversible defect than others.

Numbers of major vessels (Vessel) treated as a discrete variable in this GAM fitted model has a very high positive significant association with HD. For every one number increase in Vessel causes 1.2636 increment in log odds of HD with p-value <0.001.

Results of Non-parametric estimation for approximate significance of Smooth term

In this GAM fitted model only one cofactor namely Age, used as smoothing factor. As it is a nonparametric method of estimation so Chi-square test statistic has been used for testing the hypothesis. From table 2 it is observed that smoothness of the cofactor Age is partially significance with p-value 0.0957.

It also noticed from Table 2 that, the GAM fitted model has an Adjusted R-square value 0.70 with 65% of its deviance explained. UBRE (Un biased risk estimator) score is -0.2423 which is also very low compare to other models.

From Table 2, the final selected GAM fitted binary logistic model of the Heart disease (y) is shown below

log odds(HD)

=**-6.64**+**1.49** Chest Pain 2+**0.66** Chest pain 3+**2.77** Chest pain4

+**0.0098** Cholesterol-**0.03** *Max.HR*+**0.51***Old peak*+**0.02** *Resting BP*

+**2.18** *Resting ECG*2+**0.76** *Resting ECG*3+**2.08** *Sex* 2+**0.06** *Thal* 2

+**1.69** *Thal* 3+**1.26***Vessel*+*f*(*Age*)

In the above predictive formula, except Age all the cofactors entered in this additive model parametrically. Age is the only smoothing term here whose approximate significance has been judged through non-parametrical methods (Chi-Square test).

In Figure 2 and 3, the GAM diagnostic plots have been examined for binomial logit model. Figure 2(a) shows the histogram of the residuals for binomial logit GAM, which indicates that the residuals are normally distributed. Figure 2(b) represents the plot of the smooth terms for cofactor Age with confidence belt. It shows that the non-linearity with respect to its smoothness.

In Figure 3(a), the absolute residual values are plotted against the fitted values of GAM. This residual plot is completely a flat diagram indicating that the variance is constant with the respective means. Figure 3(b) reveals the normal probability plot for the fitted model, which shows no systematic departure or lack of fit, or response distribution, or variables or outliers with respect to the fitted GAM model.

Results of Data Mining Techniques

Table 3 presents the results of Data Mining Techniques for heart disease diagnosis. Mainly two classification methods SVM and MLPE are introduced for diagnosis. Two performance measures namely Classification accuracy rate (ACC) and Area under curve (AUC) are checked here using 10-flods cross validation with 5 runs in each experiment. It observed from Table 3 that for both of these two performance measures SVM is superior to MLPE. After 10-flods cross validation with 5runs the average ACC value for SVM is almost 85% whereas MLPE shows 82% accuracy rate. In case of AUC value SVM and MLPE show almost 0.90 and 0.86 respectively.

In Figure 4, the plots from sensitivity analysis under SVM are shown. Figure 4(a) shows the Input importance bar charts for heart disease diagnosis. Maximum heart rate is most important input variables for heart disease diagnosis under SVM (best classifier out of all data mining techniques). Figure 4(b) shows the variable effective curve (VEC) for Max HR and it is decreasing, results form Table-2 also suggests this.

CONCLUSION

The current article is considered the Heart Disease/HD (whether a patient has a heart disease or not) as the response variable. It is a binary variable with values ‘1’ and ‘2’ which stand for absent and present of the heart disease respectively. This HD has been modeled based on general-

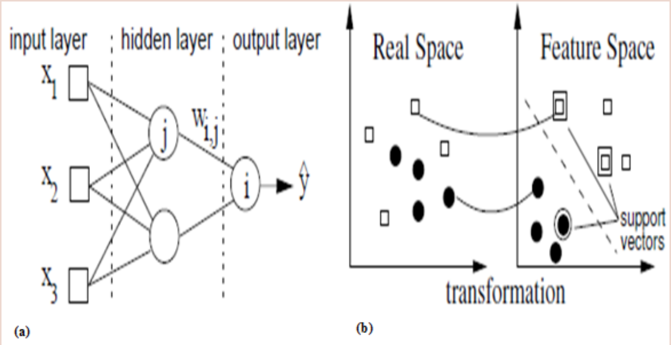

Figure 1: Data Mining Techniques (a) Multi-Layer Perceptron Neural Network (MLPE)(b) Support Vector machine (SVM)

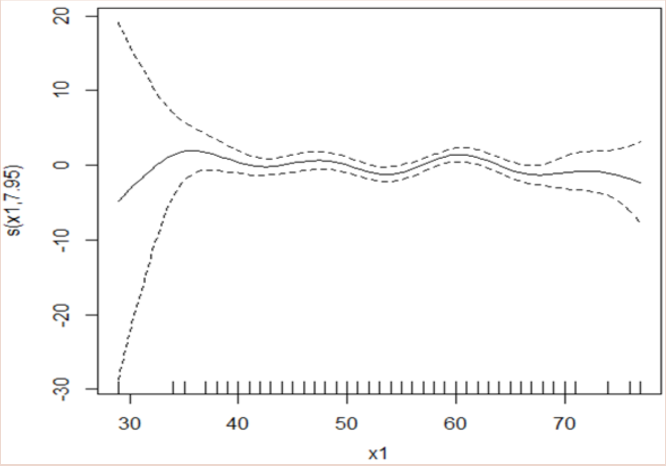

Figure 2(b): Smoothing term (Age) plot with confidence belt.

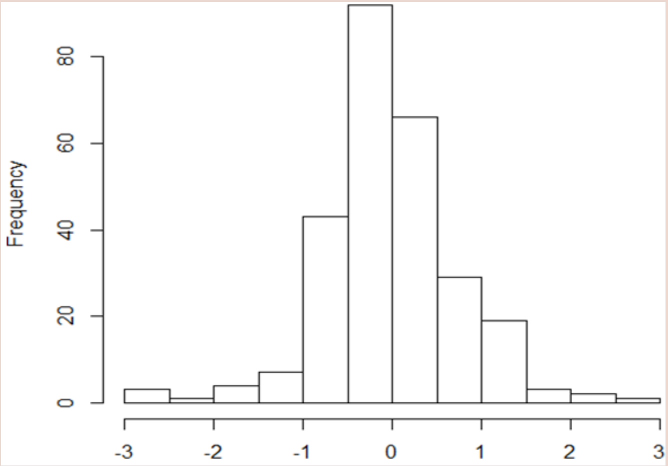

Figure 2(a): Histogram of residuals.

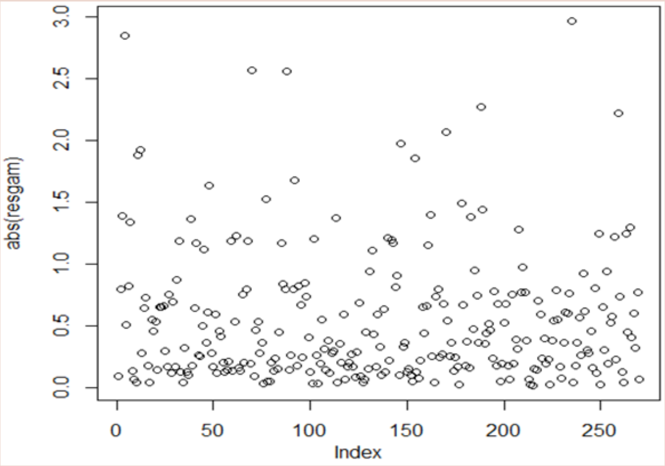

Figure 3(a): Absolute residual plot.

Table 3: Results of ACC and AUC heart disease dataset by 10 folds cross validation in 5 runs

	ACC (Classification Accuracy Rate in %)						AUC (Area Under Curve in 0-1)					
Run **Method**	1st	2nd	3rd	4th	5th	Average	1st	2nd	3rd	4th	5th	Average
SVM	84.45	85.45	84.75	84.75	84.45	**84.77**	0.8968	0.9023	0.8955	0.9028	0.8968	**0.8985**
MLPE	82.20	80.74	82.22	81.85	82.22	**81.82**	0.8724	0.8545	0.8622	0.8566	0.8594	**0.8610**

SVM: Support vector machine; **MLPE:** Multilayer perceptron ensembles.

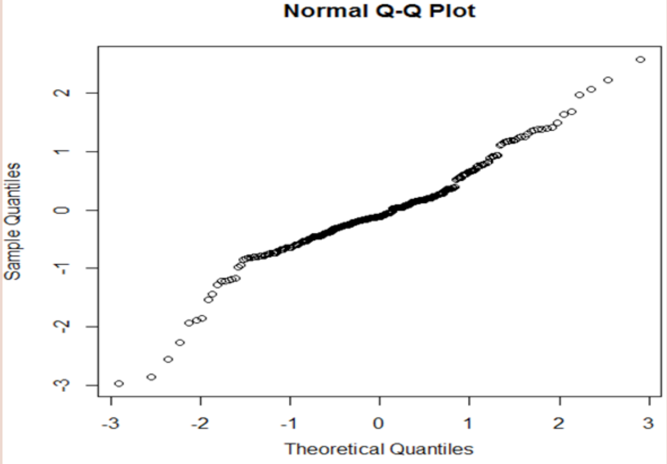

Figure 3(b): Normal probability plots of residuals.

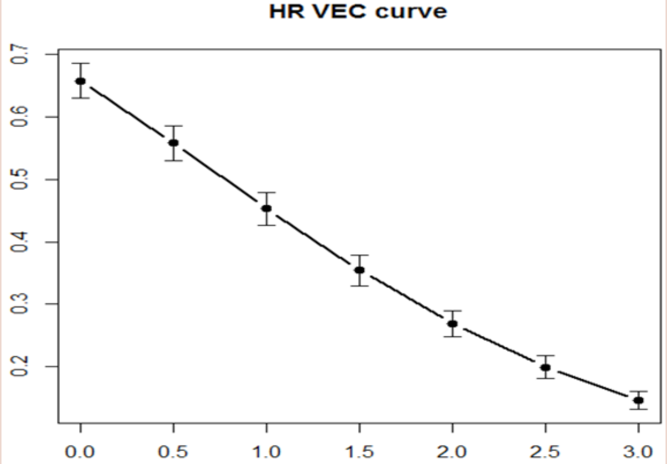

Figure 4(b): Variable effective Curve for Max. HR(most important input variable).

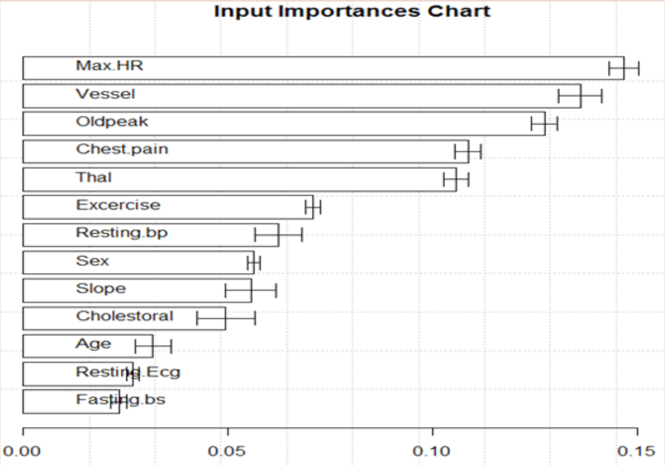

Figure 4(a): Input Importance Chart.

ized additive model. The GAM fitted model results are displayed in Table 2.

The current reported results (Table 2), though not completely conclusive, are revealing. The determinants of HD are derived satisfying the following regression analysis criteria. First, the determinants are selected based on GAM fitted model analyses. Second, the final model is selected based on UBRE.[40-47] Third, final model is justified based on GAM diagnostic plots. Fourth, the standard error of the estimates is very small, indicating that the estimates are stable 48

Fifth, the final model of the HD is selected based on locating the appropriate statistical distribution. The HD distribution is identified herein as the binomial distribution. For more extension regarding this please follow the references.[49,50]

To the best of our knowledge, the present models (Results & interpretation section) can be considered as one of the best first building block of a regression analysis. The current models may provide better assistance for treatment decision making using the individual patient risk factors and the benefits of a specific treatment. The current results have focused many interesting conclusions. These findings may help the medical practitioners for better medical treatment. Thallium scan report, Chest pain type are highly important for identification of a heart disease patients. Especially for male patient, it is recommended that they must take care about their heart during their older age.

ACKNOWLEDGEMENT

We would like to acknowledge all the previous authors who had work on this data set and also the UCI Machine Learning Repository for making available of this dataset. Finally, we are very much thankful to the reviewers for their valuable comments for betterment of this article.

CONFLICT OF INTEREST

None declared.

ABBREVIATION USED

SVM: Support vector machine; **MLPE:** Multi layer perceptron ensemble; **MLP:** Multi layer perceptron; **GAM:** Generalized additive model; **HD:** Heart disease; **DM:** Data mining; **VEC:** Variable effective curve.

REFERENCE

4. Pampel FC, Pauley S. Progress against heart disease. Greenwood Publishing Group; 2004.
5. Lahsasna A, Ainon RN, Zainuddin R, Bulgiba AM. A transparent fuzzy rule-based clinical decision support system for heart disease diagnosis. Knowledge Technology. 2012;295(2):62-71.
6. Yaron G. Symptoms and Complications of Heart_Disease. [Online]: www.itamar-medical.comPatient_Information/Cardio_101/.
7. Bhasin M, Raghava GP. Analysis and prediction of affinity of TAP binding peptides using cascade SVM. Protein Science. 2004;13(3):596-607.
8. Parthiban G, Srivatsa SK. Applying machine learning methods in diagnosing heart disease for diabetic patients. International Journal of Applied Information Systems (IJAIS). 2012;3:2249-0868.
9. Prerana THM, Shivaprakash NC, Swetha N. Prediction of Heart Disease Using Machine Learning Algorithms- Naïve Bayes, Introduction to PAC Algorithm, Comparison of Algorithms and HDPS. International Journal of Science and Engineering. 2015;3(2):90-9.
10. Ghumbre S, Patil C, Ghatol A. Heart Disease Diagnosis using Support Vector Machine. International Conference on Computer Science and Information Technology (ICCSIT'2011) Pattaya Dec. 2011
11. Rajkumar A, Reena GS. Diagnosis of heart disease using datamining algorithm. Global journal of computer science and technology. 2010;10(10):38-43.
12. Olaniyi EO, Oyedotun OK, Adnan K. Heart diseases diagnosis using neural networks arbitration. International Journal of Intelligent Systems and Applications. 2015;7(12):72.
13. Khanna D, Sahu R, Baths V, Deshpande B. Comparative Study of Classification Techniques (SVM, Logistic Regression and Neural Networks) to Predict the Prevalence of Heart Disease. International Journal of Machine Learning and Computing. 2015;5(5):414.
14. Mythili T, Mukherji D, Padalia N, Naidu A. A heart disease prediction model using SVM-Decision Trees-Logistic Regression (SDL). International Journal of Computer Applications. 2013;68(16).
15. Shomona GJ, Ramani GR. Data Mining in Clinical Data Sets: A Review. International Journal of Applied Information Systems. 2012;4(6):15-26.
16. Saidi M, Chikh MA, Settouti N. Automatic identification of diabetes diseases using a modified artificial immune recognition system2 (MAIRS2). InProceedings of 3ème conference internationale sur l 'informatique et ses applications 2011.
17. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. International journal of medical informatics. 2008;77(2):81-97.
18. Witten IH, Frank E. Data mining: practical machine learning tools and techniques. Morgan Kaufmann; 2005.
19. Werbos PJ. Beyond regression: New tools for prediction and analysis in the behavioral sciences. Doctoral Dissertation, Applied Mathematics, Harvard University, MA. 1974.
20. Rumelhart D, Hinton G, Williams R. Learning Internal Representations by Error Propagation. (Book chapter -8) Parallel Distributed Processing: Explorations in the Microstructures of Cognition. 1986; 1:318-362, MIT Press, Cambridge.
21. Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. InProceedings of the fifth annual workshop on Computational learning theory 1992 Jul 1 (pp. 144-152). ACM.
22. Smola AJ, Schölkopf B. A tutorial on support vector regression. Statistics and computing. 2004;14(3):199-222.
23. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference and Prediction. 2nd edition, 2008; Springer-Verlag, NY, USA.
24. Huang Z, Chen H, Hsu CJ, Chen WH, Wu S. Credit rating analysis with support vector machines and neural networks: a market comparative study. Decision support systems. 2004;37(4):543-58.
25. Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Philip SY, Zhou ZH. Top 10 algorithms in data mining. Knowledge and information systems. 2008 ;14(1):1-37.
26. Hastie T, Tibshirani R. Generalized additive models for medical research. Statistical Methods in Medical Research. 1995;4:187-196.
27. Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American statistical association. 2002;97(457):77-87.
28. Lee JW, Lee JB, Park M, Song SH. An extensive comparison of recent classification tools applied to microarray data. Computational Statistics & Data Analysis. 2005;48(4):869-85.
29. Li T, Zhang C, Ogihara M. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics. 2004;20(15):2429-37.
30. Liao JG, Chin KV. Logistic regression for disease classification using microarray data: model selection in a large p and small n case. Bioinformatics. 2007;23(15):1945-51.
31. Nahar J, Imam T, Tickle KS, Chen YP. Computational intelligence for heart disease diagnosis: A medical knowledge driven approach. Expert Systems with Applications. 2013;40(1):96-104.
32. Detrano R, Janosi A, Steinbrunn W, Pfisterer M, Schmid JJ, Sandhu S, et al. International application of a new probability algorithm for the diagnosis of coronary artery disease. The American journal of cardiology. 1989;64(5):304-10.
33. Cheung N. "Machine learning techniques for medical analysis," B.Sc.Thesis, School of Information Technology and Electrical Engineering, University of Queenland, 2001.
34. Polat K, Sahan S, Kodaz H, Günes S. A new classification method to diagnosis heart disease: Supervised artificial immune system (AIRS). Inproceedings of the turkish symposium on artificial intelligence and neural networks (TAINN) 2005.
35. Hastie T, Tibshirani R. Generalized additive models. John Wiley & Sons, Inc.; 1990.
36. Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer C, Wirth R. CRISP-DM 1.0: Step-by-step data mining guide. CRISP-DM consortium; 2000.
37. Cortez P. A tutorial on the rminer R package for data mining tasks, Teaching Report, Department of Information Systems, ALGORITMI Research Centre, Engineering School, University of Minho, Guimarães, Portugal; 2015.
38. Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.2nd edition, Morgan Kaufmann, San Francisco, CA; 2005.
39. Rocha M, Cortez P, Neves J. Evolution of Neural Networks for Classification and Regression. Neurocomputing. 2007;70:2809-16.
40. Wu TF, Lin CJ, Weng RC. Probability estimates for multi-class classification by pairwise coupling. Journal of Machine Learning Research. 2004;5:975-1005.
41. Kewley RH, Embrechts MJ, Breneman C. Data strip mining for the virtual design of pharmaceuticals with neural networks. IEEE Transactions on Neural Networks. 2000;11(3):668-79.
42. Cortez P, Teixeira J, Cerdeira A, Almeida F, Matos T, Reis J. Using Data Mining for Wine Quality Assessment. InDiscovery Science 2009(Vol. 5808, pp. 66-79).
43. Wood SN. Generalized Additive Models: An Introduction with R. London: Chapman and Hall; 2006.
44. Myers RH, Montgomery DC, Vining GG, Robinson TJ. Generalized linear models: with applications in engineering and the sciences. John Wiley & Sons; 2012.
45. Currie ID, Durban M, Eilers PH. Generalized linear array models with applications to multidimensional smoothing. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2006;68(2):259-80.
46. Green PJ, Silverman BW. Nonparametric regression and generalized linear models: a roughness penalty approach. CRC Press; 1993.
47. Ruppert D. Selecting the number of knots for penalized splines. Journal of computational and graphical statistics. 2002;11(4):735-57.
48. Eilers PH, Marx BD. Flexible smoothing with B-splines and penalties. Statistical science. 1996;1:89-102.
49. Watkins A. AIRS: a resource limited artificial immune classifier. Master thesis. MississippiState University; 2001.
50. Ruppert D, Wand MP, Carroll RJ. Semi parametric Regression, first ed. Cambridge University Press New York; 2003.
51. Chatterjee S, Hadi AS. Regression Analysis by Example, fifth ed. John Wiley & Sons, New Jersey; 2006.
52. Das RN, Mukherjee S, Panda RN. Association between Body Mass Index and Cardiac Parameters of Worcester Heart Attack Study. BAOJ Cell Mol Cardio. 2016; 2:006.
53. Das RN, Mukherjee S. Joint Mean-Variance Overall Survival Time Fitted Models from Stage III Non-Small Cell Lung Cancer. Epidemiology (Sunnyvale). 2017;7:296

Cite this article : Mukherjee S, Kapoor S, Banerjee P. Diagnosis and Identification of Risk Factors for Heart Disease Patients Using Generalized Additive Model and Data Mining Techniques. J Cardiovasc Disease Res. 2017;8(4):137-44.