Available Online at www.ijcsmc.com/journal

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 7.056

IJCSMC, Vol. 11, Issue. 6, June 2022, pg.129 – 136

Heart Disease Prediction Using Machine Learning Algorithms: A Systematic Survey

Pavan Kumar Tadiparthi[1]; Vennelarani Kuna[2]
[1]Department of Information Technology & MVGR College of Engineering, Vizianagaram India
[2]Department of Information Technology & MVGR College of Engineering, Vizianagaram India
[1] pavank3400@gmail.com; [2] vennelaranikuna@gmail.com
DOI: https://doi.org/10.47760/ijcsmc.2022.v11i06.010

Abstract*— *The heart is the one of the most typical and important organ in our human body. Over few decades Cardiovascular Diseases became one of the most frequent reasons of deaths. This threatening not only in India but also the whole world. The heart was attacked by so many factors like age, sex, diet, stress, smoking etc. So there is a need to early diagnosing the disease accurately so that immediate treatment can be provided and saves millions of lives .The incorrect prediction may also cause side effects or loss of life. In the last few decades eminent researchers are proposed many approaches to predict the heart diseases. In this article, we are reviewed different types of efficient machine learning algorithms for heart disease prediction with correlation matrices; visualize the features and performance metrics like precision, recall, accuracy. In our survey the logistic regression approach gives the best accuracy result which is 81.9%.

Keywords*— *cardiovascular diseases, machine learning, correlation matrices, metrics.

I. INTRODUCTION

Heart disease is one of the primary reasons of deaths in adults. It is said that health is more important than wealth. According to Word health organization (who) guidelines, a good health is the fundamental right for each and every individuals. That was discovered that almost 40% of deaths in the world caused by heart diseases. The blood flow reduces in heart causes the heart attack. The risk factors like diet, blood pressure, diabetics, depression, stress, genetics features cause the heart diseases. Heart attack also depends on one of the family history like siblings, parents or ancestors have suffered early attacks. It is also known as myocardial infraction (MI) cardiac or myco cordial infraction and coronary thrombosis. Generally in our blood there is two types of cholesterols. One is good cholesterol and second one is bad cholesterol .It can be detected by lip protein test .huge amount of triglycerides raises your risk of heart attack.
Heart diseases describe a range of conditions that can majorly affect the heart and they are:

1. Blood vessels diseases
2. Congential heart affect

3. Arrhythmias
4. Heart valves diseases
5. Heart muscle diseases
6. Infection in heart

The symptom of heart disease depends on the type of heart problem we have. We might not be diagnosed with coronary artery disease until we get heart attack, stroke, heart failure or angina.
The supply of oxygen and blood flow to the heart associated with one type of heart disease which refers to coronary artery disease (CAD).The angina and heart attacks causes due to low blood flow to the heart. The heart diseases also huge impact on two types of factors .one is changed factors like smoking, high blood pressure, high cholesterol, obesity, unhealthy diet, diabetics, depression and stress. The second one is unchanged risk factors like age, gender, genetic factors, and race. Based on the different survey reports specified that the heart diseases cannot be noticed from the symptoms. However, the existing techniques like ECG, EEG, EMG, and BP, heart rate to track and trace the early detection of different heart disease abnormalities.
The huge amount of data is being generated and processed from the various hospitals and medical centres and an organization. The data cannot be used in certain way and the critical data can be future processed to the clinical decision support system and management. The doctors have a chance of leaving the hidden features from the data for analysing .It leads to wrong classification of disease and unwanted biases. This might affect the medical treatment expenditure and quality of services provided to the patients.so we need to develop an efficient system for decreasing human errors and improve patient quality services. This can be achieved by the combining medical decision support system with computer decision support system.
The eminent researchers are working under this direction and study about various effective machine learning classification algorithms for prediction of heart diseases. It also depends on very important key factor is feature selection attributes in the data set. In this article, we are study about four efficient machine learning classifiers and its performance measured by metrics like accuracy.
The rest of the article is organized as follows
The section II described in related work. Machine learning Algorithms are presented in section III. The data set mentioned in section IV. The proposed methodology explained in section V of this article. The results are described in section VI. Section VII of this article explained conclusion and future work.

II. RELATED WORK

Kannel et al [1] presented the study of various heart disease and its risks. Fizar ahmed [2] proposed K nearest neighbour algorithm (KNN) to detect the heart diseases with better accuracy. Prince akngal et al[3,4] introduced different data mining algorithms or techniques to predict heart disease and the attributes used in their data set like sex ,age, blood pleasure and blood sugar. Rajkumar et al [5] proposed data mining techniques like naïve bays algorithm. It was compared with manual diagnosed time. Naive bays took 609 ms to detect the disease. Sarman issam [6] presented different types of classifiers like NB-Tree, bayesan network etc. to detect the heart diseases. Jesse davis et al [7] focused on the risk of heart attacks diagnosed through eMedical. Himanshu sharma et al [8] introduced Machine learning and deep learning approaches and techniques for prediction of heart diseases. Beant kaur et al [9] described various neural networks techniques for prediction of heart diseases. Gayathri et al [10] focused on soft computing techniques and machine learning algorithms to detect heart diseases.

III. ALGORITHMS DESCRIPTION

The efficient machine learning algorithms are

1. Logistic regression
2. KNN algorithm
3. Decision tree
4. Random forest

The main objective of our article is to check whether a patient is likely to be diagnosed with any Cardiovascular diseases based on medical attributes like Age, sex, chest pain, testing bp, serum cholesterol, fasting blood sugar, resting electrocardiographic results, max heart rate, exercise induced angina, ST depression induced by exercise relative to rest, slope of the peak exercise ST segment, major number of vessels, thallium scan.

1) Logistic regression

Logistic regression is a part of classification and it describes the relationship between one dependent variable and one or more independent variable and the result will be 0 or 1

For example the below graph is of heart disease prediction in which Y axis is of occurrence of heart disease and X axis is the human age.

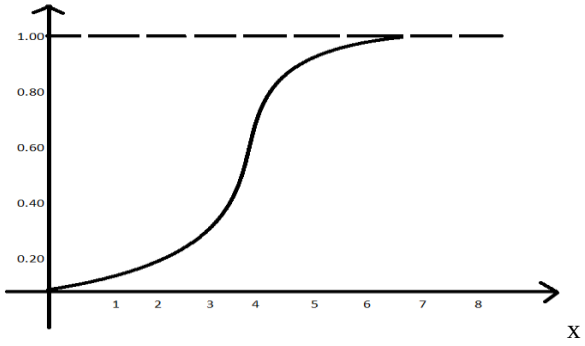

Fig: heart disease prediction

Let us consider a straight line:

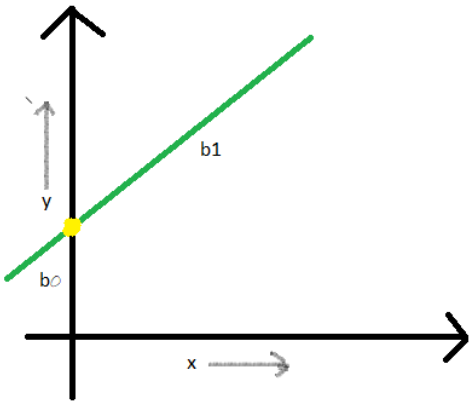

Fig: heart disease example with x and y axis

$Y=b_0 + b_1 * x$ (1)

Where b_0 is y intercept and b_1 is the slope of the line

x- is the value of the x coordinate,y- is the value of the prediction

Formula to predict odd of success is

$\log(p(x)/(1-p(x)) = b_0 + b_1x$ (2)

Exponentiation both the sides

$e^{\ln}(p(x)/(1-p(x)) = e^{(b_0 + b_1 x)}$ (3)

$p(x)/(1-p(x) = e^{(b_0 + b_1 x)}$ (4)

let $Y = e^{(b_0 + b_1 x)}$ then $p(x)/1-p(x) = Y$ (5)

$Y(1-p(x))=p(x)$, (6)

$P(x)=Y-Y(p(x))$, (7)

$p(x) + Y(p(x)) = Y$, (8)

$p(x)(1+Y) = Y$, (9)

$p(x) = Y / 1+Y$ (10)

$p(x) = e^{(b_0 +b_1 x)} / (1+e^{(b_0 +b_1 x)})$ (11)

Therefor the equation of sigmoid function is

$p(x) = 1 / (1+e^{-(b_0 +b_1 x)})$ (12)

This sigmoid function is used to map the predicate value to the probabilities and convert the value to exact 0 or 1.

2) K-Nearest Neighbour (KNN)

This is very easy to understand and implement.This algorithm works on the bases of similarity. By the help of this algorithm the data is first divided into parts on based on similarities. So whenever a dataset is introduced to the model the dataset can be easily classified. This algorithm doesn't make any assumption on fundamental data. It only used for classification purpose only.This algorthim cannot handle noise data.

The algorithm first select some n number of neighbours.
Then the algorithm calculates the euclidean distance of n neighbour

$$d(\mathbf{p},\mathbf{q}) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \qquad (13)$$

Where
p,q = two points in euclidean n-space
p_i,q_i = euclidean vectors starting from the origin ie initial point
n = n-space
now we have to take the nearest neighbour that is calculated by euclidean formula. By this way the data is categorized and the model is build up.

3) Decision tree

Decision tree use multiple algorithm to create its sub-nodes. Creation of sub-nodes increases homogeneity. Decision tree splits the nodes on all available variables and then selects the split which results in most homogeneous sub-node.
The selection of algorithm is also based on target variables.

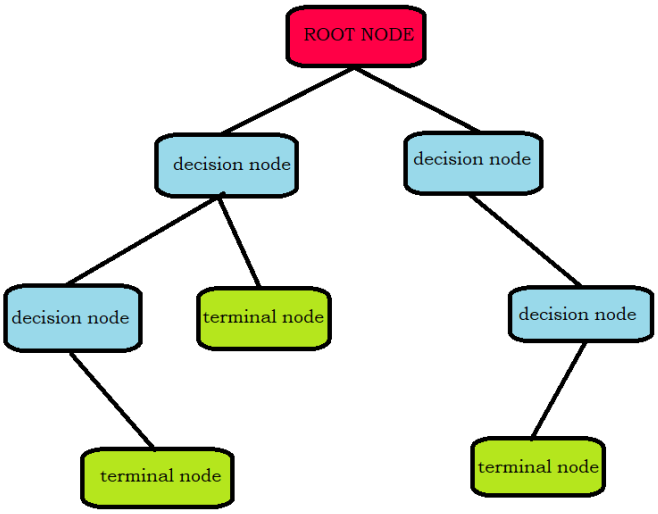

Fig:Decision tree classification process

The decision for node which to be root node,decision node and terminal node is classified using the metric called entropy.
The formula of entropy is

$$E(S) = -p_{(+)}\ \log p_{(+)} - p_{(-)}\ \log p_{(-)} \qquad (14)$$

Where
$p_{(+)}$ = the probability of positive class
$p_{(-)}$ = the probability of negative class
S is the subset of the traning set

4) Random forest

Random forest is a classifier that consists of number of decision trees on various subsets of the dataset and average the available decision tree and predict the accurate value. Greater the trees higher the accuracy.-

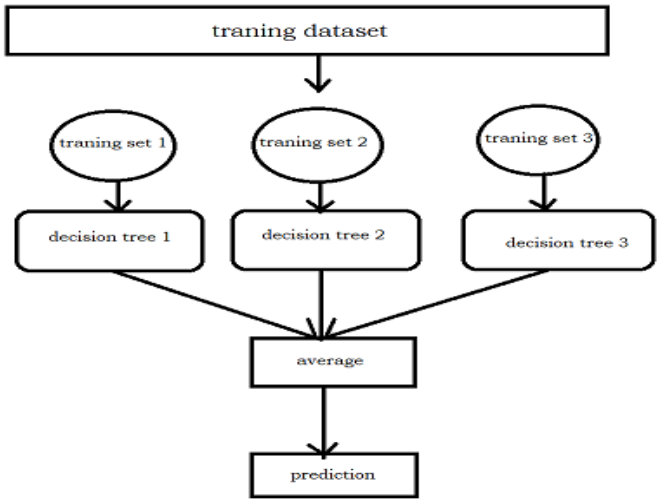

Fig: random forest classifier process

The algorithm first selects K data points from the traning set then decision tree is applied to all datapoints .The step is repeated till the k datapoints and then the average is taken to get accurate value.

IV. DATA SET CONSIDERED

For our experimentation we have considered the dataset provided by kaggle.com webpage.
The dataset consist of 500 instances with 14 attributes. The dataset consists of 14 attributes are mentioned in the below table

Features	type	Description, value
Sex	Discrete	1=male, 0=female
chest pain	Discrete	1= typical angina 2=atypical angina 3= non-angina pain 4= asymptomatic
fasting blood sugar	Discrete	fasting blood sugar 0=false 1=true
resting electrocardiographic results	Discrete	0=normal 1=abnormal
exercise induced angina	Continuous	exercise induced angina 1=yes 0=no
slope of the peak exercise ST segment	Continuous	slope of the peak exercise ST segment 2=upsloping 1=down sloping 0=flat
ST depression induced by exercise relative to rest	Continuous	number of major vessels colored by fluoroscopy (0-3)
thallium scan	Discrete	Patient heart rate, 3 = normal; 6=fixed defect; 7 = reversible defect
Age	Continuous	Patients age , 28 to 77
tresting bp	Continuous	Resting blood pressures of patients measured in mm Hg on admission to the hospital 80-200
Cholesterol	Continuous	Patient serum cholesterol measured in mg/dl. 85, 100 - 200 - 394, 400- 603

max heart rate	Continuous	Patient maximum heart rate achieved.60-202, Low: below 50, Normal:51-119, High: 120-202 [6] [7]
Serum	Continuous	ST depression made by exercise relative to rest -2.6 to -0.1, 0, 0.1 to 6.2, 120
major number of vessels	Continuous	0=low 1=medium 3=high

Table: Description of Attributes in the Data Set

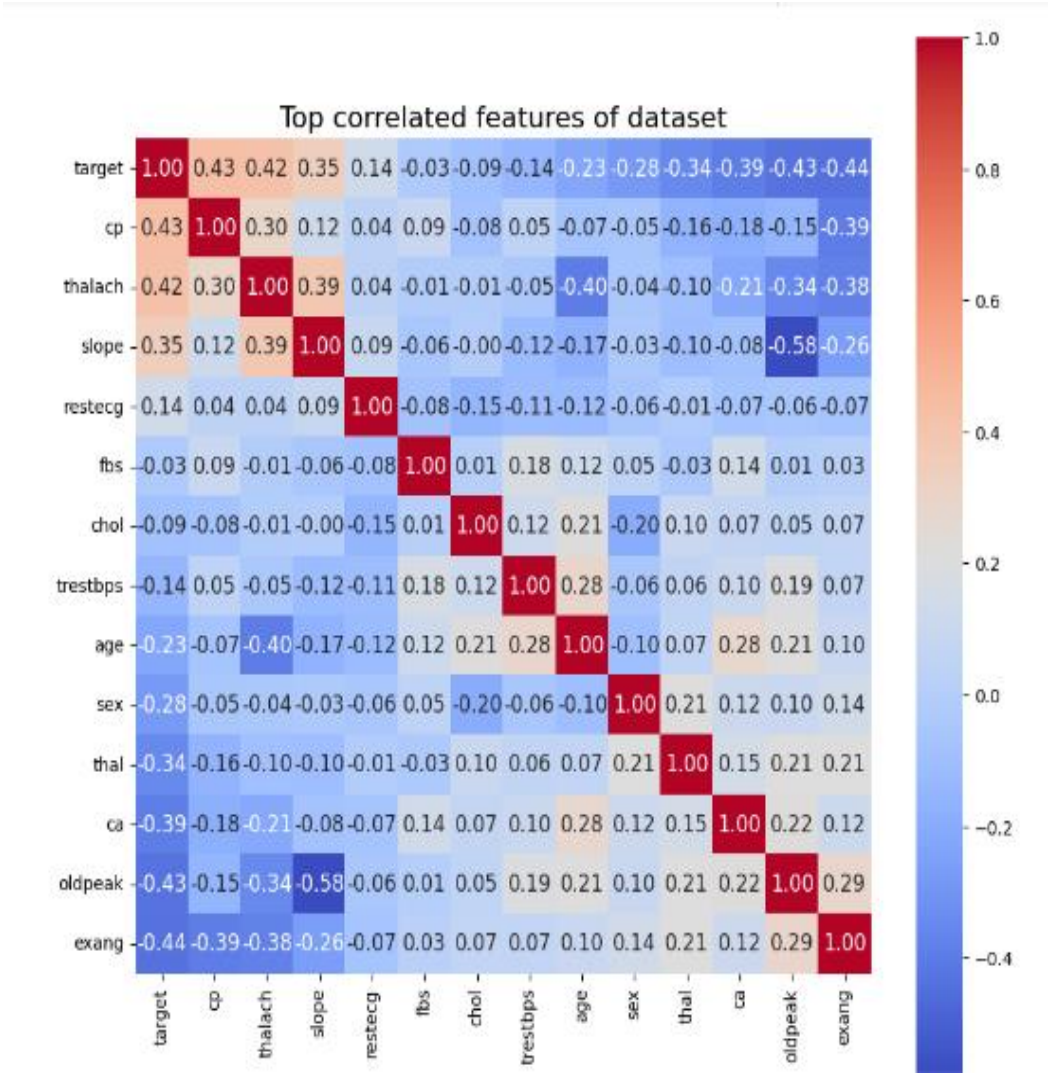

Fig: Correlation feature metrics of the data set

V. PROPOSED METHODOLOGY

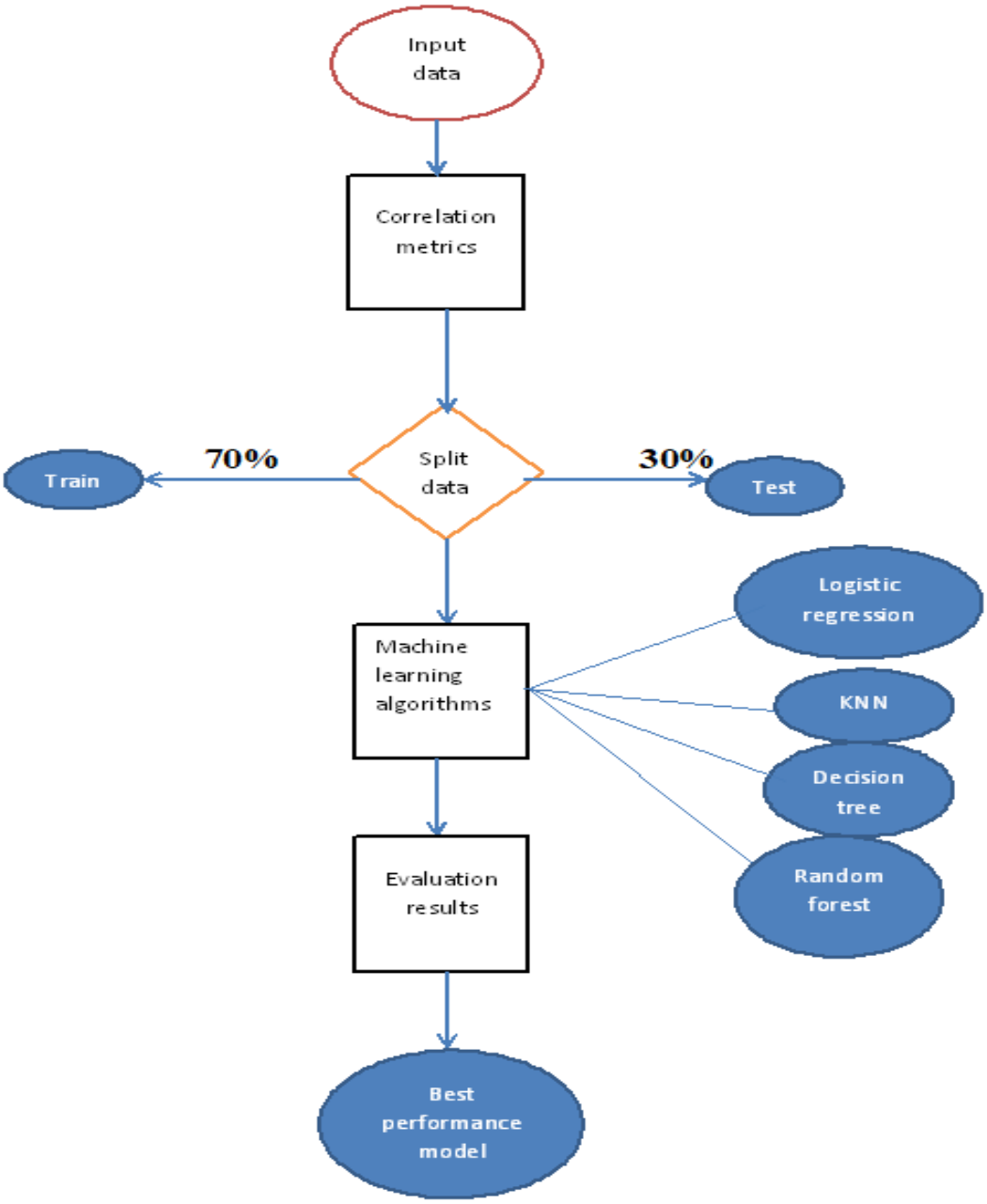

Fig: Flow chart for machine learning algorithms

A. Data pre-processing :

a. In this step the raw data is prepared such that it is efficiently used for the model.

B. Fitting four models to the trained data

a. The dataset is now trained in these models and this is used with the help of sklearn library

C. Predicting the test result

a. After dataset is well trained, we can now predict the result by using it. When the probability is greater than 0.50 then the value will be round off to 1 and if the probability is lesser than 0.50 then the value will be round off to 0.

D. Test accuracy of data

a. Now confusion matrix is created to check the accuracy of the classification.

VI. EXPERIMENTATION AND RESULTS

The experimentation was carried out in python environment. The peformace table was presented in below table and the accuracy 81.9% achieved by logistic regression.The classification model gives the better perfomance on data set with 14 features.The accuracy can be computed by below formula

Accuracy (all correct / all) = TP + TN / TP + TN + FP + FN (15)

Model Name	Accuracy (%)
Logistic regression	81.9
KNN	62.2
Decision tree	60.6
Random forest	59.0

Table :Classifier result analysis

VII. CONCLUSION AND FUTURE SCOPE

In this article, we study about four efficient machine learning algorithms for heart disease prediction. The performance was computed by accuracy metric. The Logistic regression algorithm gives better result with accuracy 81.9% from our data set with 14 features. In future, for the heart diseases prediction using deep learning and artificial intelligence techniques.

REFERENCES

[1] Kannel, William B., "Contribution of the Framingham Study to preventive cardiology", Journal of the American College of Cardiology 15.1 (1990): 206-211.

[2] Mahmooda, S. S., Levy, D., Vasan, R. S., Wang, T. J. (2014). The Framingham Heart Study and the epidemiology of cardiovascular diseases: a historical perspective. Lancet, 383(9921), 999-1008.

[3] Ahmed, Fizar. "An Internet of Things (IoT) application for predicting the quantity of future heart attack patients." International Journal of Computer Applications 164.6 (2017).

[4] Methaila, A., Kansal, P., Arya, H., Kumar, P. (2014). Early heart disease prediction using data mining techniques. Computer Science Information Technology Journal, 53-59.

[5] Raj Kumar, Asha, and G. Sophia Reena. "Diagnosis of heart disease using data mining algorithm." Global journal of computer science and technology 10.10 (2010): 38-43.

[6] Salman, Issam. "Heart attack mortality prediction: an application of machine learning methods." Turkish Journal of Electrical Engineering Computer Sciences 27.6 (2019): 4378-4389.

[7] Davis, Jesse, et al. "Machine learning for personalized medicine: Will this drug give me a heart attack." The Proceedings of International Conference on Machine Learning (ICML). 2008.

[8] Sharma, Himanshu, and M. A. Rizvi. "Prediction of heart disease using machine learning algorithms: A survey." International Journal on Recent and Innovation Trends in Computing and Communication 5.8 (2017): 99-104.

[9] Kaur, Beant, and Williamjeet Singh. "Review on heart disease prediction system using data mining techniques." International journal on recent and innovation trends in computing and communication 2.10 2014): 3003-3008.

[10] Dangare, Chaitrali, and Sulabha Apte. "A data mining approach for prediction of heart disease using neural networks." International Journal of Computer Engineering and Technology (IJCET) 3.3 (2012).

[11] Gayathri, P., and N. Jaisankar. "Comprehensive study of heart disease diagnosis using data mining and soft computing techniques." (2013).

[12] Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., Klein, M. (2002). Logistic regression. New York: Springer-Verlag.

[13] Buhlmann, P., Yu, B. (2002). Analyzing bagging. The Annals of ¨ Statistics, 30(4), 927-961.

[14] Pal, M. (2005). Random forest classifier for remote sensing classification. International journal of remote sensing, 26(1), 217-222.

[15] Rokach, L., Maimon, O. (2005). Decision trees. In Data mining and knowledge discovery handbook (pp. 165-192). Springer, Boston, MA.

