International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 11 Issue 1 ǁ January 2023 ǁ PP. 408-417

An Automated System to Predict Heart Disease Using Supervised Machine Learning Techniques

Gudla Karthik
Research Scholar, Computer Science, and Engineering, Lovely Professional University, Phagwara, Punjab, India

Sandeep Kaur
Assistant Professor, Computer Science, and Engineering, Lovely Professional University, Phagwara, Punjab, India

Kadali Sai Vinesh
Research Scholar, Computer Science, and Engineering, Lovely Professional University, Phagwara, Punjab, India

Satti Vijay Bhaskar Reddy
Research Scholar, Computer Science, and Engineering, Lovely Professional University, Phagwara, Punjab, India

Abstract
There are many deadly chronic diseases in the world among them heart disease is in the top position. Due to this more people are dying which can be avoided with proper treatment in the initial stage but predicting heart disease in the early stage is difficult. In this paper, we proposed an automated system that predicts the existence of heart disease by using supervised learning techniques that can benefit by taking precautionary steps to avoid mortality due to heart diseases. In the proposed research data pre-processing and model building have been carried out using classification-based algorithms like SVM, Decision trees, and Random Forest. And an optimal model is considered based on performance metrics like Accuracy. Moreover, extensive data exploration is also carried out to find the contribution of each attribute. The optimal model for this case study is Random Forest with an accuracy of 88.26%.

KEYWORDS:*Machine Learning, Supervised Learning techniques, Support Vector Machines, Decision Tree, Random Forest.*

Date of Submission: 14-01-2023 Date of acceptance: 29-01-2023

I. INTRODUCTION

Cardiovascular heart disease is one of the major chronic diseases which causes a high death rate. It is a very difficult task to be predicted by medical practitioners as it requires higher knowledge and expertise in detecting it. Using a predictive model-based system in the diagnosis would help us to predict the disease and enhance the efficiency of the process and decrease costs [6]. The motive of the project is to produce an automated system that can identify the existence of heart disease in a person using supervised learning techniques [3]. Also, extensive data exploration is carried out to find insights from the data [2]. Our objective is to collect and explore the data and modify it to our requirements. Splitting the data into training and testing for model building. Building the classification-based model using machine learning algorithms. Finding the best fit model for the case study by following certain performance evaluation metrics. Using deployment techniques on a classified model to use it with ease.

II. LITERATURE SURVEY

[1] Yash Goyal et al. aim to predict the failure of the heart of a person in the next 10 years.10 different classification algorithms were combined to form an ensemble language. With this model, 85.2% accuracy and 87.5% test recall were achieved, and determined the model predicted that people with high bp suffer from a heart attack.

[2] Attluri Rudra et al. developed a data science framework using different classification algorithms to find the existence of a cardiovascular disease. The aim is to build an optimal classification algorithm using health records and parameters. For accurate predictions, the Data science framework is used. SVM and logistic regression have produced better performance than the other algorithms

[3]Akanksha Kumari et al. have implemented seven machine learning algorithms for predicting heart disease. AdaBoost and the voting ensemble methods were used to improve the accuracy of algorithms. Obtained results showed that using AdaBoost Algorithm has improved the accuracy of The Decision Tree classifier.

[4] Richa Sharma et al. have developed a model to predict the patient's risk level using parameters of their health. The aim of the model is to improve medical care and avoid unnecessary costs. The Experiment is based on the Cleveland dataset. To fill the missing values in the dataset AllPossible-MV algorithm is used.
[5]M.A.Jabbar et al. proposed a model using Hidden Naive Bayes for the classification of heart disease. 100% of accuracy was achieved and outperforms Naïve Bayes. The experiment has shown that the HNB model has a better performance compared with others.

[6]L. Vanitha et al. recommended a model to predict sudden cardiac arrest before 30 minutes of its occurrence based on time domain and frequency domain. Physionet dataset is used to check the validity of the work. 88% of classification efficiency was achieved using SVM.

[7]Aditi Gavhane et al. proposed a model for predicting the vulnerability of cardiovascular disease using basic parameters like sex, pulse rate, age and also using a neural network. Aim to identify heart stroke symptoms in the initial stage. The model was able to generate the result in terms of Yes or No if the patient has cardiovascular disease.

[8] Rashmi G Saboji et al proposed a framework using attributes of healthcare data to predict cardiovascular disease. Aim diagnosis of heart disease with using a small number of attributes. Random forest outperformed the Naïve-Bayes classifier. 98% accuracy was achieved by the random forest.
[9]S.Rajathi developed an application that predicts heart disease using knn algorithm. prediction is done in 2 phases first phase is done with knn algorithm and in second phase knn algorithm is combined with ant colony optimization(aco) for more accuracy. Compared with other algorithm like decision tree, svm, knn and knn with aco. knn with aco got more accuracy 70.26% and less error rate 0.526.

[10] G. Gnaneswar designed a model to predict heart disease using hybrid machine learning model on Cleveland dataset. Hybrid model is a combination of decision tree and random forest model. Experimental results shows that Hybrid machine learning model done a great performance compared with other algorithm like decision tree and random forest and 88% accuracy is achieved using hybrid model.

III. PROPOSED WORK

This is a web-based prediction project based on the heart failure prediction dataset. For this as per dataset attributes, the application takes user inputs and predicts whether a person is having heart disease or not. And gives the output along with the accuracy.

3.1. ALGORITHM

1. Input the patient details
i. Checking the format of details entered by the user
ii. And processing the details
2. Training the dataset
3. Supervise machine learning algorithms are used for predictions.
4. Displaying the result along with accuracy

3.2. ARCHITECTURE DIAGRAM

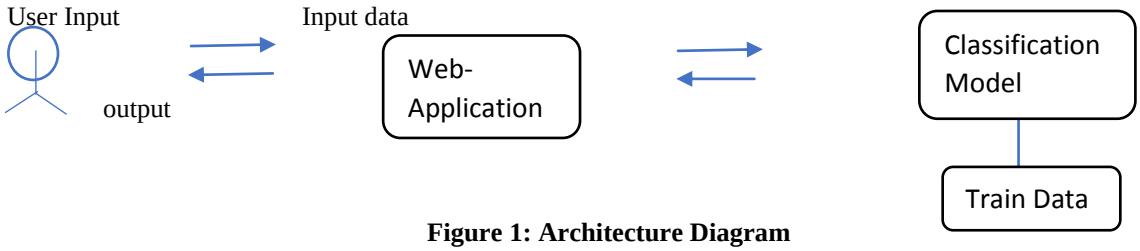

Figure 1: Architecture Diagram

IV. METHODOLOGY

4.1.1. Data Preprocessing
This is the first step in the project which involves Data Cleaning, Finding missing and null values, and factorization of attributes.

4.1.2. Data Cleaning
This dataset doesn't require data cleaning since it has no missing and null values.

4.1.3. Factorization
Here we basically assign a meaningful name for the attribute to avoid confusion for the algorithm. For example, heart disease attribute is having 0, 1 it has been changed to no, yes which signifies that no for patient with no heart disease and yes for patient having heart disease.

4.2 Analysing the dataset
The dataset has been analysed with statistical functions in R and studied about the structure of dataset, number of observations, dependent and independent variables. And a Univariate and Bi-Variate analysis has been carried out for knowing the scope of each attribute and few inferences have been found. The results of data exploration are discussed under Results and Discussion section.

4.3.1. Model Building
After a thorough analysis it is known that each feature in the dataset has important contribution for the dependent variable, heart disease. So, all the variables are considered for the predictive model building.Before building these models, for each model the dataset is spilt into 75:25 ratio as training set and testing set respectively.The proposed project is implemented with three different supervised machine-learning classification-based algorithms they are explained as follows.

4.3.2. Support Vector Machines (SVM)
A Support Vector Machine (SVM) is a supervised machine learning technique which is used in classification and regression problems but mostly used for classification problems. The SVM algorithm creates a line called hyperplane which is used to separate data and classify the data points in correct categories. The SVM algorithm considers the extreme points for creating a hyperplane are called as support vectors. And based on these extreme data points, support vectors the data is classified.

4.3.3. Decision Tree Algorithm
The Decision Tree algorithm is also a supervised learning technique. With the help of the decision tree algorithm, we can solve regression and classification problems but mostly decision trees are used for classification problems. The Decision Tree Classification Algorithm is tree-structuredone in which branches signifies the decision rules, nodes signify the attributes in the dataset and each leaf node signifies the outcomes. The Decision Tree model creates a tree structure to determine the dependent variable by some rules of independent variables which are generated from training data.

4.3.4. Random Forest
Random Forest is also a supervised learning technique that contains multiple decision trees. Random Forest algorithm is used for implementing both classification and regression problems. Random Forest algorithm first constructs a bootstrapped dataset (selects random samples with replacements).and then constructs decision tree using bootstrapped dataset these steps are repeated to get more number of decision trees. Each decision tree will generate an output final decision taken based on majority decision trees output for classification and averaging for regression. Compared to other supervised machine learning algorithms Random Forest is more accurate and gives better performance.

4.4. Model Selection

After building and finding predictions from the model, each model is evaluated with a performance metric like Accuracy. Highest accuracy model is considered as accurate model and selected for deployment.

4.5. Deploying for web-application

The web application takes dataset attributes as input values by the user and gives the result.

a) Algorithm of Methodology work:

1. Data Pre-processing:

Data cleaning
Factorization

2. Analysing the dataset and its attributes.
3. Building a predictive model using machine learning algorithms.
4. Selection of model built based on accuracy.
5. And finally developing a web-based application through the selected model.

V. RESULTS AND DISCUSSION

5.1 Exploratory Data Analysis Results

Even though the dataset has no invalid values (null or missing data) There are two attributes in the dataset Resting BP and Cholesterol which are having minimum value as zero which is ideally not true. In these dependent variables are analysed with independent variable, heart disease. Here are the results of it.

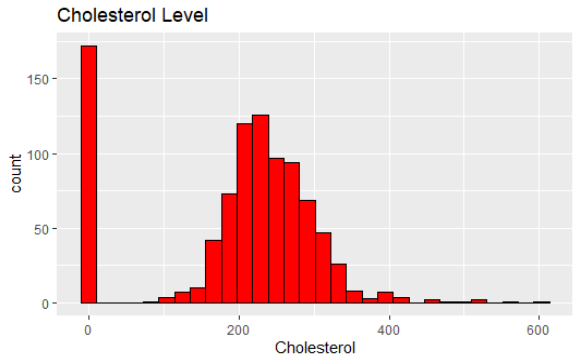

Figure 2: Cholesterol Level Distribution

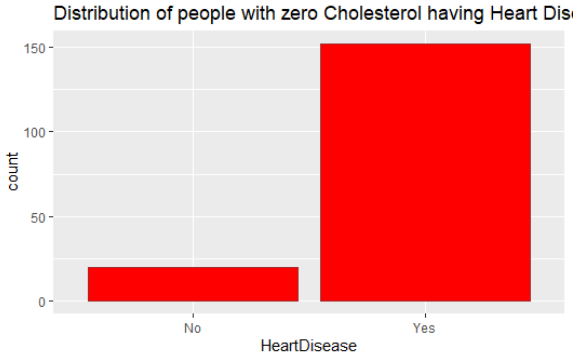

Figure 3: Cholesterol Distribution wrtHeart Disease

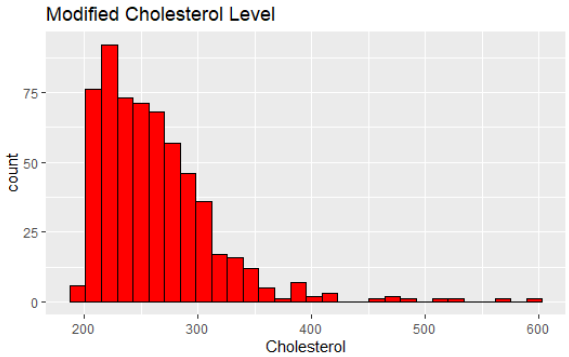

Figure 4: Modified Cholesterol Level Distribution

Figure 5: People with >200 cholesterol levels having Heart Disease

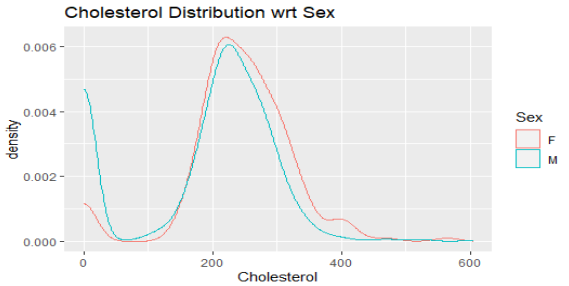

Figure 6: Cholesterol Distribution Vs Sex

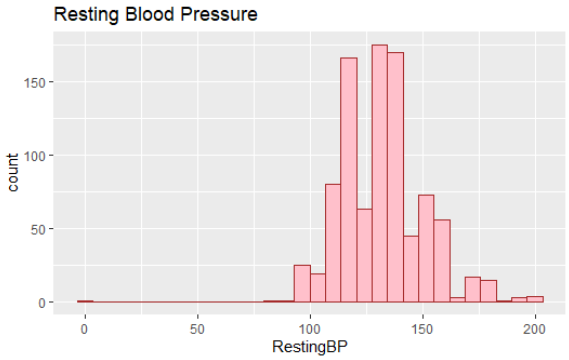

Figure 7: Resting BP Distribution

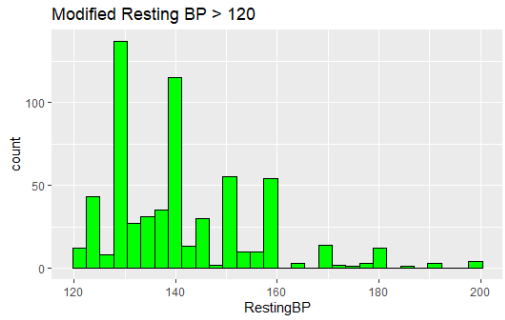

Figure 8: Modified BP Level > 120

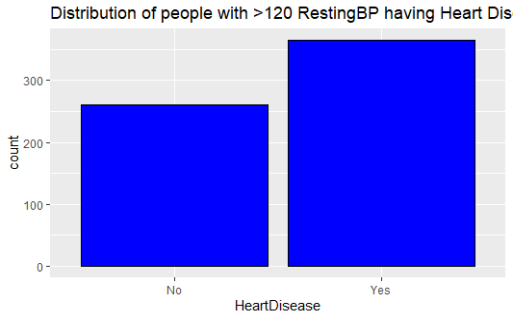

Figure 9: People >120 having Heart Disease

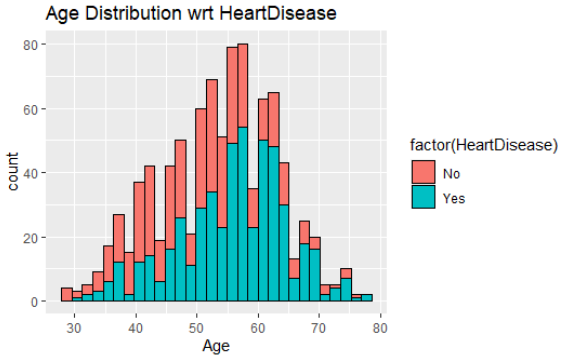

Figure 10: Heart Diseasewrt Age

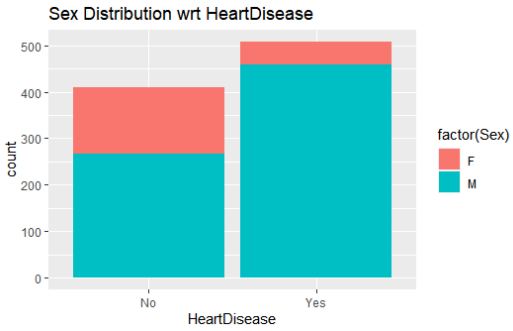

Figure 11: Heart Diseasewrt Gender

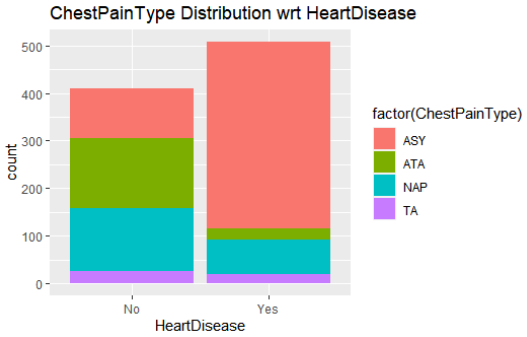

Figure 12: Heart Diseasewrt Chest Pain Type

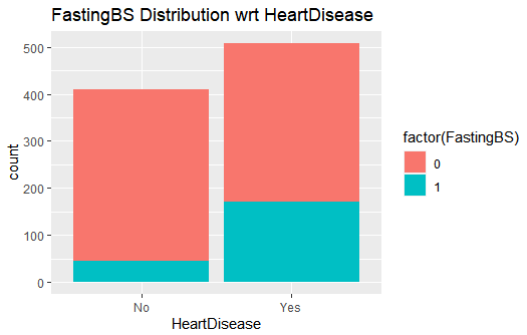

Figure 13: Heart Diseasewrt BS

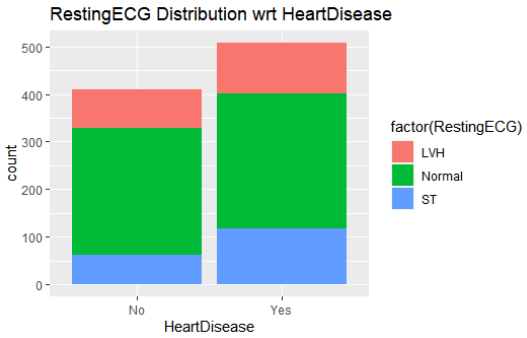

Figure 14: Heart Diseasewrt ECG

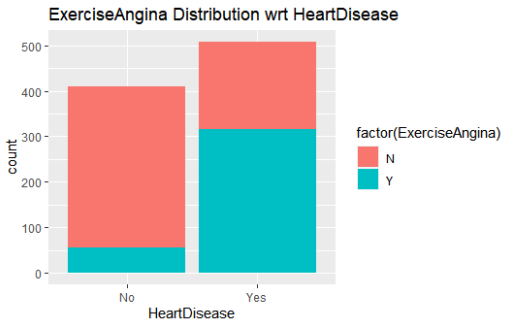

Figure 15: Heart Diseasewrt Angina

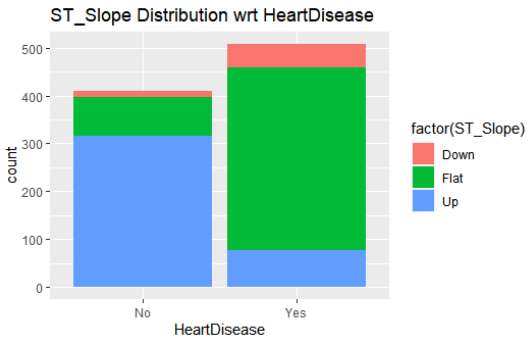

Figure 16: Heart DiseasewrtST_Slope

The average number of cholesterol levels are more for Female than Male. Initially the plot looks like as cholesterol levels starts increasing it gradually decreases and coincide at one point for both genders and began risen and slowly from one point again started decreasing and gets coincide [Figure 6]. Since there are cholesterol values 0 [Figure 2], The insights for it shows 88.4% people have heart disease while 11.6 don't moreover it says more percentage of males prone to have heart disease. It has been modified [Figure 4] and taken analysis for cholesterol levels against heart disease it showsa slight difference as 50.2% of people don't have heart disease and 49.8% are having heart disease [Figure 5].

SimilarlyRestingBP has don't have ideal values [Figure 6] hence it is modified [Figure 7] further it is modified as BP>120 and the results looks like count of patients is high in the range of 125-130 BP values [Figure 8]. And 41.6% don't have heart disease if BP >120 while 58.4% is having heart disease. Whereas the count of patients having BP > 120 is more for Male gender i.e., 80% and 20% females. [Figure 9].

Age group of 55 to 60 is having more number of heart disease also the count of not having heart disease is also more in the same age group [Figure 10]. Heart disease for males is high compared to female [Figure 11]. And ASY Chest Pain type is causing high for more number of heart failure cases, the count of not having heart failure is high for ATA chest pain type [Figure 12]. Most of the patients having Fasting Blood Sugar don't have heart disease [Figure 13]. The greater number of counts for having heart disease is high for Normal Resting ECG and for the same type count of not having heart failure is also high [Figure 14]. The count of having heart disease is more for Exercise Angina with Yes and count not having heart disease is high in Exercise Angina with No [Figure 15]. The patients with Flat ST_Slopeis having high count of heart disease and Up ST_Slope has high count of not having heart disease [Figure 16].

5.2 Classification Model Evaluation

After Building a classification model, performance metrics like Accuracy is used for selecting a model to implement web-based application. The accuracy of a model is derived by using confusion matrix. Confusion matrix (cm) is defined as a matrix that shows predicted values against actual values.

Actual

	Positive	Negative
Positive	TP	FP
Negative	FN	TN

Predicted

TN = cm [0] [0]: TN True Negative means correctly classified as not class of interest; The model finds that person has heart disease and he is not having.
FN = cm[1] [0]: FN False Negative means wrongly classified as not class of interest; The model finds that person don't have heart disease but he is having.
TP = cm[1] [1]: TP True Positive means correctly classified as class of interest; The model finds that person is has heart disease and he is having.
FP = cm [0] [1]: FP False Positive wrongly classified as class of interest; The model finds that person is has heart disease but he is not having.

And from these values we can derive Accuracy formula i.e.,

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

Accuracies of three classification models are shown in the following table

Table 1:Accuracies of Classification Models.

Algorithm	Accuracy
SVM	86.02
Decision Trees	82.09
Random Forest	88.26

Figure 17: Accuracy Vs Classification Models

Since Random Forest is having highest accuracy, it is taken for developing web-based application using R Shiny.

VI. CONCLUSION

Taking the dataset from open source Kaggle, after performing pre-processing and analyzing steps, three algorithms were implemented namely SVM, Decision trees and Random Forest. Dataset is split into training and testing sets and they trained with respective classification-based algorithm. As per our proposed work the more efficient algorithm is to be selected for developing UI based web application based on the performance metrics like Accuracy. We found the accuracies of algorithms like SVM algorithm with 86.02%, Decision tree with 82.09%, Random Forest with 88.26%. So, the high accurate model is Random Forest with 88.23% accuracy. Hence, it is used for web-based application which is built using R Shiny web tool. And few major insights have been discovered from data exploration they are ASY chest pain type, Normal Resting ECG Levels and Flat ST_Slope are causes for high heart disease cases. Also, People having Exercise Angina and BP > 120 are prone to have heart diseases. This will help the users to know their heart status in advance. Since early prediction of such major chronic diseases in world using machine learning techniques will impact the society in a good way.

REFERENCES

[1]. Harsh Agrawal, JankiChandiwala, Sarvesh Agrawal, and Yash Goyal, "Heart Failure Prediction using Machine Learning with Exploratory Data Analysis," 2021 International Conference on Intelligent Technologies (CONIT) Karnataka, India. June 25-27, 2021, IEEE, DOI: 10.1109/CONIT51480.2021.9498561.

[2]. Chittampalli Sai Prakash, Myneni Madhu Bala, Attluri Rudra, "Data Science Framework - Heart Disease Predictions, Variant Models and Visualizations," 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA), IEEE, DOI: 10.1109/ICCSEA49143.2020.9132920.

[3]. Akanksha Kumari, Ashok Kumar Mehta, "A Novel Approach for Prediction of Heart Disease using Machine Learning Algorithms," 2021 Asian Conference on Innovation in Technology (ASIANCON) Pune, India, IEEE, DOI: 10.1109/ASIANCON51346.2021.9544544.

[4]. Purushottam, Kanak Saxena, Richa Sharma, "Efficient heart disease prediction system using a decision tree," International Conference on Computing, Communication and Automation 2015, IEEE, DOI: 10.1109/CCAA.2015.7148346.

[5]. M. A. Jabbar; Shirina Samreen, "heart disease prediction system based on hidden Naïve Bayes classifier," 2016 International Conference on Circuits, Controls, Communications and Computing (I4C), IEEE, DOI: 10.1109/CIMCA.2016.8053261.

[6]. C. Jenefar Sheela, L. Vanitha, "Prediction of Sudden Cardiac Death using support vector machine," 2014 International Conference on Circuits, Power, and Computing Technologies [ICCPCT-2014], IEEE, DOI: 10.1109/ICCPCT.2014.7054771.

[7]. Aditi Gavhane, GouthamiKokkula; Isha Pandya; Kailas Devadkar, "Prediction of Heart Disease Using Machine Learning," 2018 Second International Conference on Electronics, Communication, and Aerospace Technology (ICECA), IEEE, DOI: 10.1109/ICECA.2018.8474922.

[8]. Rashmi G Saboji, Prem Kumar Ramesh, "A Scalable Solution for Heart Disease Prediction using Classification Mining Technique," 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing, IEEE, DOI: 10.1109/ICECDS.2017.8389755.

[9]. S. Rajathi, G. Radhamani, "Prediction and analysis of Rheumatic heart disease using KNN classification with ACO," 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE), IEEE, DOI: 10.1109/SAPIENCE.2016.7684132.

[10]. M. Kavitha, G. Gnaneswar, R. Dinesh, Y. Rohith Sai, R. Sai Suraj, "Heart Disease Prediction using Hybrid machine Learning Model," 2021 6th International Conference on Inventive Computation Technologies (ICICT), IEEE, DOI: 10.1109/ICICT50816.2021.9358597.