



IJIRCCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 5, May 2023

ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 8.379

9940 572 462

6381 907 438

ijircce@gmail.com

www.ijircce.com

Effective Prediction of Cardiovascular Disease

¹Dr. Jyothi K, ²Spoorthi J, ³Srusti D R, ⁴Sushma B S, ⁵Yashaswini J

¹Professor, Dept. of Information Science and Engineering, Jawaharlal Nehru New College of Engineering,
Shivamogga, India

²Student, Dept. of Information Science and Engineering, Jawaharlal Nehru New College of Engineering,
Shivamogga, India

³Student, Dept. of Information Science and Engineering, Jawaharlal Nehru New College of Engineering,
Shivamogga, India

⁴Student, Dept. of Information Science and Engineering, Jawaharlal Nehru New College of Engineering,
Shivamogga, India

⁵Student, Dept. of Information Science and Engineering, Jawaharlal Nehru New College of Engineering,
Shivamogga, India

ABSTRACT : In recent years, the cases of heart diseases are increasing at a rapid rate. The diagnosis of cardiovascular disease is a difficult task i.e., it should be performed precisely and efficiently. This paper mainly focuses on which patient is more likely to have a heart disease based on various medical attributes. The heart disease prediction system is used to predict whether the patient is likely to be diagnosed with a heart disease or not, using the medical history of the patient. Different algorithms of machine learning is used to predict and classify the patient with heart disease. A quite helpful approach was used to regulate how the model can improve the accuracy of prediction of heart disease in any individual. For the prediction of cardiovascular disease we have used random forest algorithm which gives the accuracy of 86.67%.

KEYWORDS: Cardiovascular disease, Heart Disease, Random forest.

I. INTRODUCTION

The heart is one of the main parts of the human body after the brain. The primary function of the heart is to pump blood to the whole body parts. Any disorder that can lead to disturbing the functionality of the heart is called heart disease. Several types of heart disease are there in the world. Coronary artery disease (CAD), and Heart failure (HF) are the most common heart diseases that are present. The data generated by the health or the survey are getting wasted. But as the data analytics come into existence, the hospitals and NGOs are making use of the data to generate the useful information from the data. The modern world has cardiovascular disease as its deadliest enemy. The Cardiovascular disease affects a person in such a way so that the patients can't be cured as easily as possible. So, diagnosing patients at the right time is the toughest work in medical field. The proposed model makes use of the Random Forest algorithm to effectively predict the cardiovascular disease.

II. THEORETICAL BACKGROUND

Yuepeng Liu et.al., [1] Cardiovascular diseases are among the most common serious illness affecting human health. CVDs may be prevented by early diagnosis. This paper discusses the problem of feature selection based on random forest and the prediction model of heart disease based on LSTM, CNN, DNN and KNN. The influence of different types of machine learning algorithms on accuracy was discussed, and the prediction model of heart disease was constructed. After the collection of dataset, the data pre-processing is done to transform nonstandard data and remove null values in order to make the data more suitable for analysis. The processed data features include 16 attributes. Then the classification algorithm like Random Forest is applied. It can be applied to classification problem, regression and

feature selection problems. In this paper, random forest is used for prediction they have used random tree- LSTM, DNN and KNN for comparison. The application scenario of this algorithm model is to predict the samples based on these parameters to determine whether patients are at risk of coronary heart disease proposed paper execution time is more as compared with the other algorithms.

Pranav Motarwar et.al., [2] A machine learning framework to predict the possibility of having heart disease using various algorithms. The framework is executed using five algorithms Random Forest, Naïve Bayes, Support Vector Machine, Hoeffding Decision Tree, and Logistic Model Tree (LMT). Cleveland dataset is used for training and testing the model. The dataset is pre-processed followed by feature selection to select most prominent features. The resultant dataset is then used for training the framework. The results are combined and show that Random forest gives maximum accuracy a weak classifiers are collected within the dataset in Hoeffding tree classifier which reduces the accuracy.

N. Komal Kumar et.al., [3] Hybrid approach is proposed for coronary illness forecast utilizing arbitrary random forest classifier and simple k means algorithm in machine learning. The experimental analysis takes place in two levels, in the first level the dataset is cleaned using the pandas tool and in the second level, the data is subjected to Machine learning tree classifiers such as Random Forest, Decision Tree, Logistic Regression, Support vector machine (SVM), K-nearest neighbours (KNN). Pre-processing techniques like data integration, data transformation, data reduction, and data cleaning using pandas tool are used. In this investigation of foreseeing Cardiovascular Disease, based on the precision and AUC ROC scores, Random Forest seems top be more effective classifier but the proposed paper takes more execution time for bigger data set and accuracy will be less.

Pronab Ghosh et.al., [4] For the proposed model they have used Data collection, Data pre-processing and Data Transformation methods. And features such as Relief and LASSO techniques are used. New hybrid classifier like DTBM, RFBM, KNNBM, ABBM, and GBBM are developed and classifiers with bagging and boosting methods, machine learning algorithms to calculate Accuracy, Sensitivity, Error rate, Precision and F1 score of our model Based on the result analysis, The proposed model produced highest accuracy while using RFBM and Relief feature selection methods. In the proposed paper the high level of missing values in the dataset can have adverse effect.

Vijeta Sharma et.al., [5] The objective of the paper is to build a ML model for heart disease prediction based on the related parameters. We have used a benchmark dataset of UCI Heart disease prediction for this research work, which consist of 14 different parameters related to Heart Disease. Machine Learning algorithms such as Random Forest, Support Vector Machine (SVM), Naive Bayes and Decision tree have been used for the development of model. Methodology includes Data collection, Data pre-processing, Building model and Accuracy measurement model. In this Random Forest is giving maximum accuracy. This Random forest algorithm will take more time to process and the system will be slower and will be more perplexing and will be handling more data.

III. DESIGN AND IMPLEMENTATION

The working of the system starts with the collection of data and selecting the important attributes. Then the required data is preprocessed into the required format. The data is then divided into two parts training and testing data. The Random Forest classifier applied and the model is trained using the training data. The accuracy of the system is obtained by testing the system using the testing data. This system is implemented using the following modules.

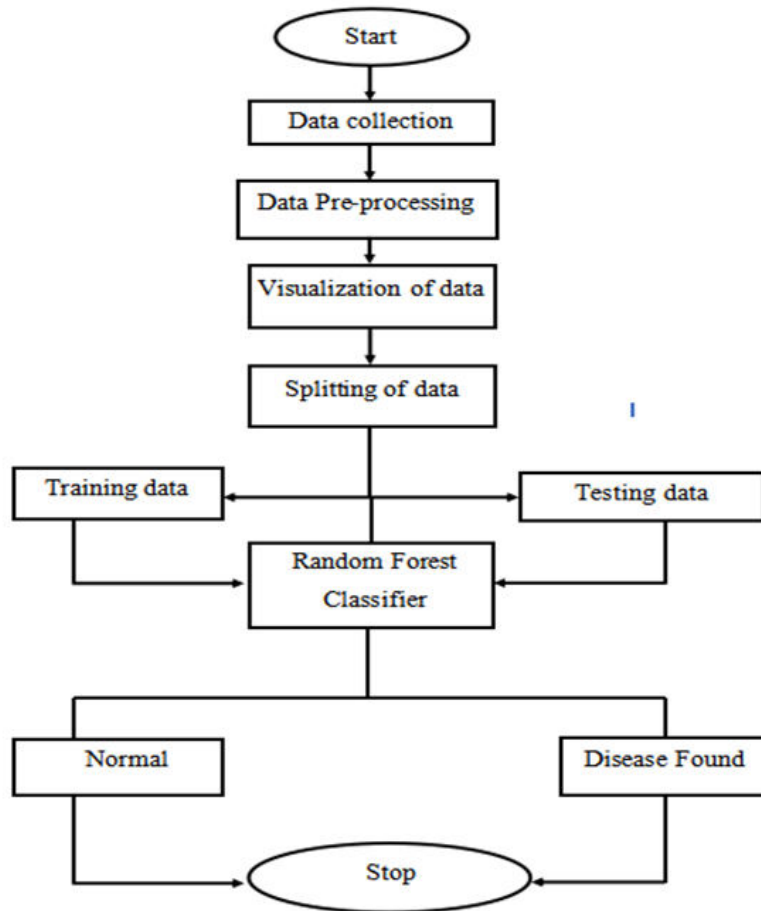


Figure3.1:Flow chat of the system

The Figure 3.1 shows the flowchart of the proposed methodology. Data required for the prediction is collected using open resources.

Data collection is an important step as the quality and quantity of the data that we gather for the proposed system will directly determine how good output that predictive model can be. The general approach is that we can collect data from open sources like Kaggle (<https://www.kaggle.com/datasets>).

Dataset Details

- The dataset used for this project is Heart Disease UCI.
- The dataset consists 14 attributes which are considered for the prediction of the output. Figure 3.2 shows the attributes of dataset.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	condition
2	69	1	0	160	234	1	2	131	0	0.1	1	1	0	0
3	69	0	0	140	239	0	0	151	0	1.8	0	2	0	0
4	66	0	0	150	226	0	0	114	0	2.6	2	0	0	0
5	65	1	0	138	282	1	2	174	0	1.4	1	1	0	1
6	64	1	0	110	211	0	2	144	1	1.8	1	0	0	0
7	64	1	0	170	227	0	2	155	0	0.6	1	0	2	0
8	63	1	0	145	233	1	2	150	0	2.3	2	0	1	0
9	61	1	0	134	234	0	0	145	0	2.6	1	2	0	1
10	60	0	0	150	240	0	0	171	0	0.9	0	0	0	0
11	59	1	0	178	270	0	2	145	0	4.2	2	0	2	0
12	59	1	0	170	288	0	2	159	0	0.2	1	0	2	1
13	59	1	0	160	273	0	2	125	0	0	0	0	0	1
14	59	1	0	134	204	0	0	162	0	0.8	0	2	0	1
15	58	0	0	150	283	1	2	162	0	1	0	0	0	0
16	56	1	0	120	193	0	2	162	0	1.9	1	0	2	0
17	52	1	0	118	186	0	2	190	0	0	1	0	1	0
18	52	1	0	152	298	1	0	178	0	1.2	1	0	2	0
19	51	1	0	125	213	0	2	125	1	1.4	0	1	0	0
20	45	1	0	110	264	0	0	132	0	1.2	1	0	2	1
21	42	1	0	148	244	0	2	178	0	0.8	0	2	0	0
22	40	1	0	140	199	0	0	178	1	1.4	0	0	2	0
23	38	1	0	120	231	0	0	182	1	3.8	1	0	2	1
24	34	1	0	118	182	0	2	174	0	0	0	0	0	0
25	74	0	1	120	269	0	2	121	1	0.2	0	1	0	0
26	71	0	1	160	302	0	0	162	0	0.4	0	2	0	0
27	70	1	1	156	245	0	2	143	0	0	0	0	0	0
28	66	1	1	160	246	0	0	120	1	0	1	3	1	1
29	63	0	1	140	195	0	0	179	0	0	0	2	0	0
30	62	1	1	120	281	0	2	103	0	1.4	1	1	2	1
31	62	1	1	128	208	1	2	140	0	0	0	0	0	0
32	59	1	1	140	221	0	0	164	1	0	0	0	0	0
33	58	1	1	120	284	0	2	160	0	1.8	1	0	0	0

Figure 3.2 : The attributes of dataset

Data preprocessing main steps that included in data preprocessing are as follows: Renaming of columns, splitting of data into feature(x) and target(y), Feature Scaling.

Data visualization refers to the representation of data or information in a visual or graphical format. It involves creating visual images or graphics to convey complex data or information in a simple and easy-to-understand format. Visualization can be used to represent various types of data, including numerical, textual, spatial, and temporal data. The goal of visualization is to make data more accessible and understandable by representing it in a format that is easy to interpret and analyze.

Splitting of data is the process to divide the dataset into **training** and **testing** 75% of the data is given to training and remaining 25% of the data is given to testing. Testing the data is used to evaluate the performance of the model using ML algorithm. Based on the training data and testing data the best model is selected. The training data is different from testing data. The obtained data is applied to the Random Forest algorithm The splitting of data into two parts is shown in the figure 3.3.

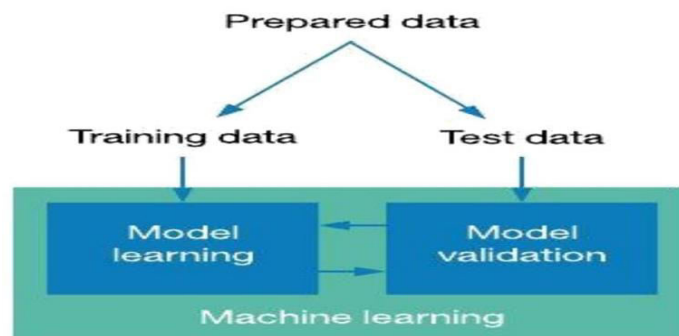


Figure3.3 Splitting of data

RANDOM FOREST ALGORITHM an extension of machine learning classifiers which include the bagging to improve the performance of Decision Tree. It combines tree predictors, and trees are dependent on a random vector which is independently sampled. The distribution of all trees are the same. Random Forests splits nodes using the best among of a predictor subset that are randomly chosen from the node itself, instead of splitting nodes based on the variables.

Algorithm Steps:

It works in four steps:

- Select random samples from a given dataset.
- Construct a Decision Tree for each sample and get a prediction result from each Decision Tree.
- Perform a vote for each predicted result.
- Select the prediction result with the most votes as the final prediction.

IV. RESULT AND DISCUSSION

```

Classification Report
precision    recall  f1-score   support

   0       0.88    0.88    0.88         42
   1       0.85    0.85    0.85         33

 accuracy          0.87         75
  macro avg          0.86         75
 weighted avg          0.87         75

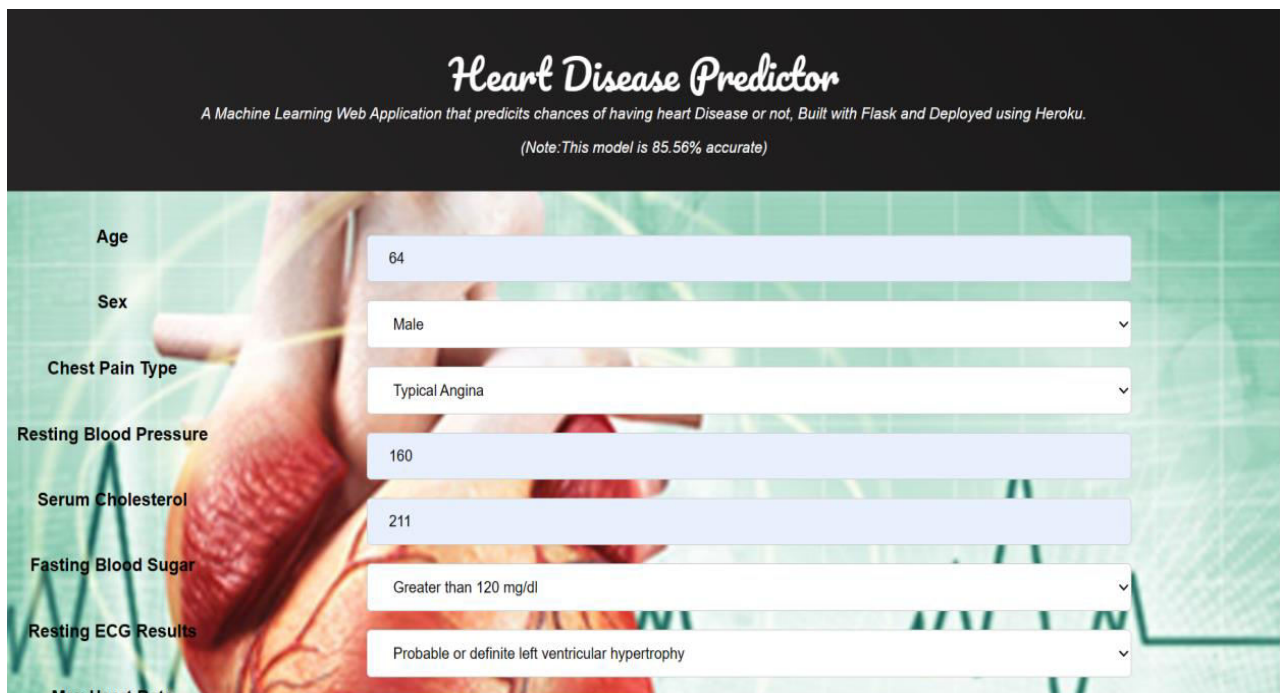
Accuracy: 86.67%

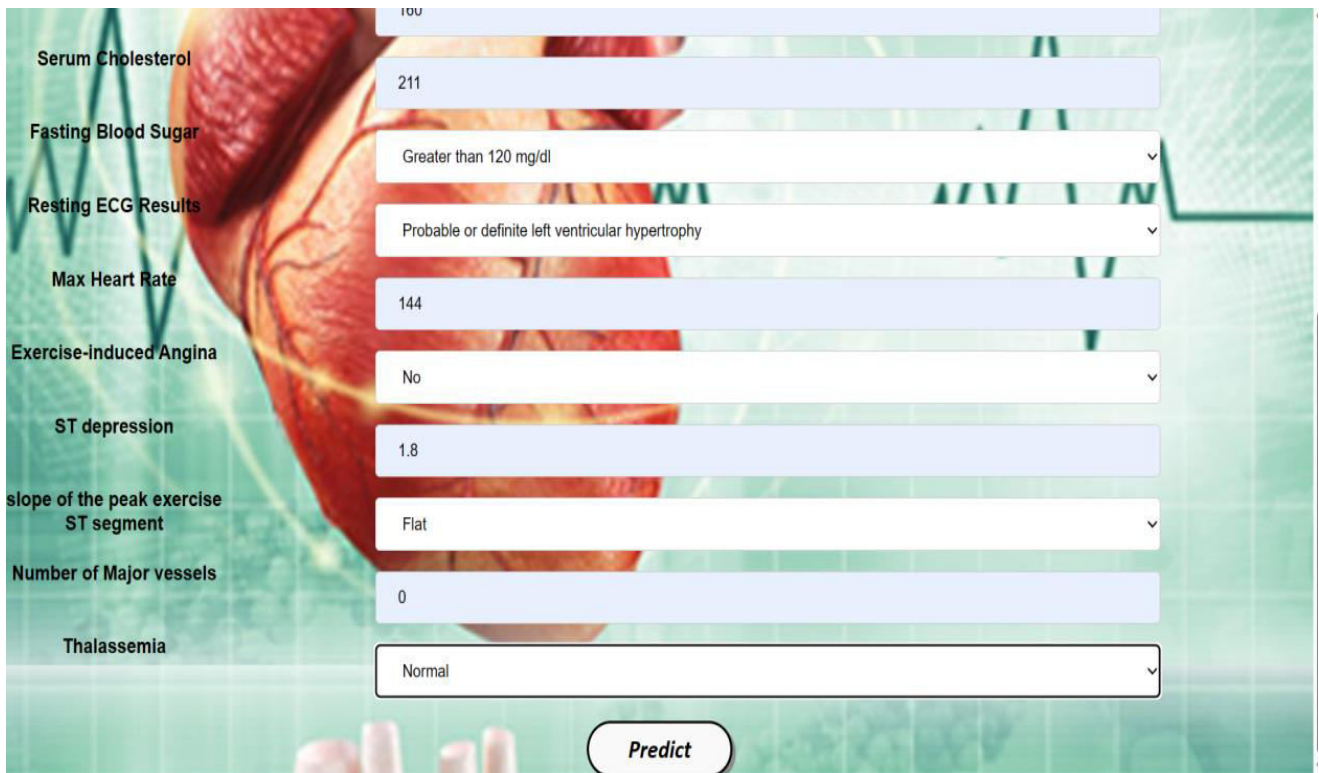
[[37  5]
 [ 5 28]]
    
```

Figure4.1 Result of RFM with accuracy

Figure4.1 depicts the performance of the random forest classifier in predicting the presence or absence of cardiovascular disease, the confusion matrix can be used to calculate metrics such as accuracy, precision, recall, and F1 score. These metrics provide information about the overall performance of the classifier and can be used to compare different models or to evaluate the impact of different parameters on model performance.

Results of User Interface





The image shows a web application interface for a heart disease predictor. It features a background image of a human heart and an ECG line. The interface consists of several input fields and a 'Predict' button. The data entered in the fields is as follows:

Parameter	Value
Serum Cholesterol	211
Fasting Blood Sugar	Greater than 120 mg/dl
Resting ECG Results	Probable or definite left ventricular hypertrophy
Max Heart Rate	144
Exercise-induced Angina	No
ST depression	1.8
slope of the peak exercise ST segment	Flat
Number of Major vessels	0
Thalassemia	Normal

A 'Predict' button is located at the bottom center of the form.

Figure.4.2 User Interface with entered details

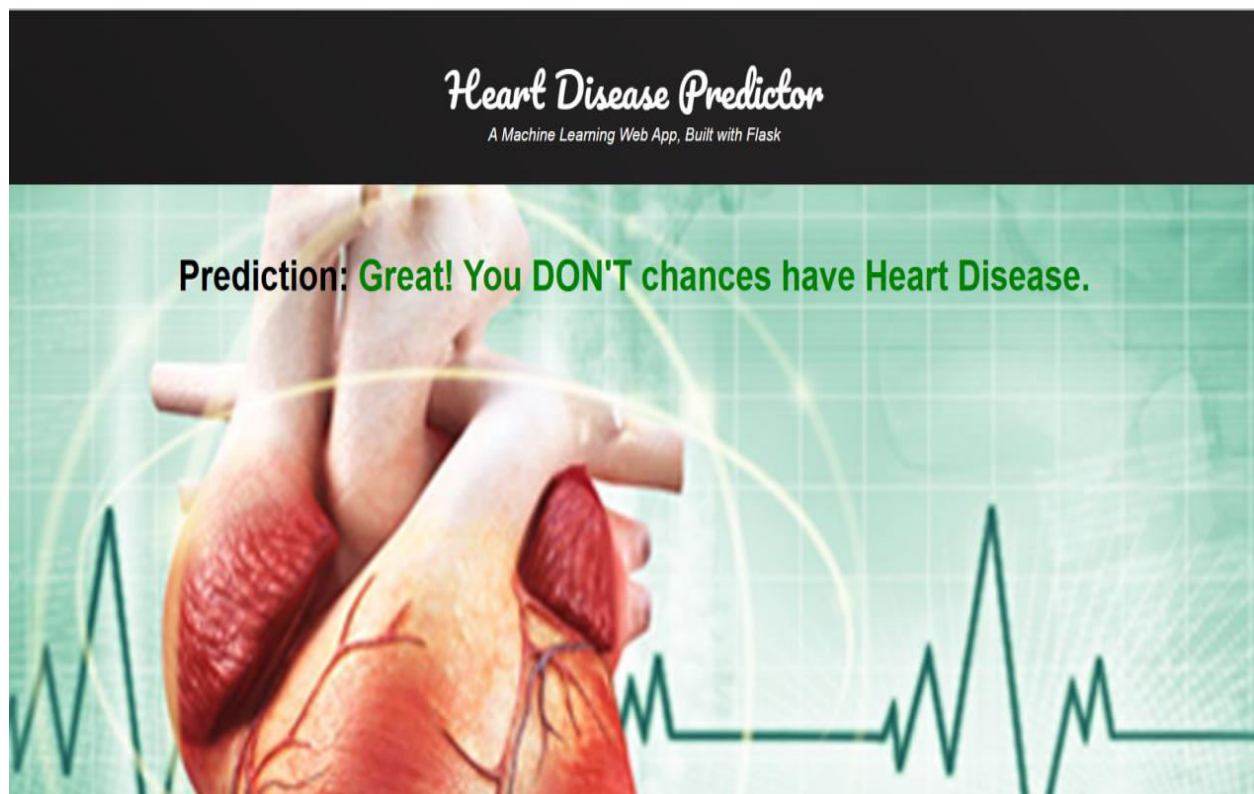
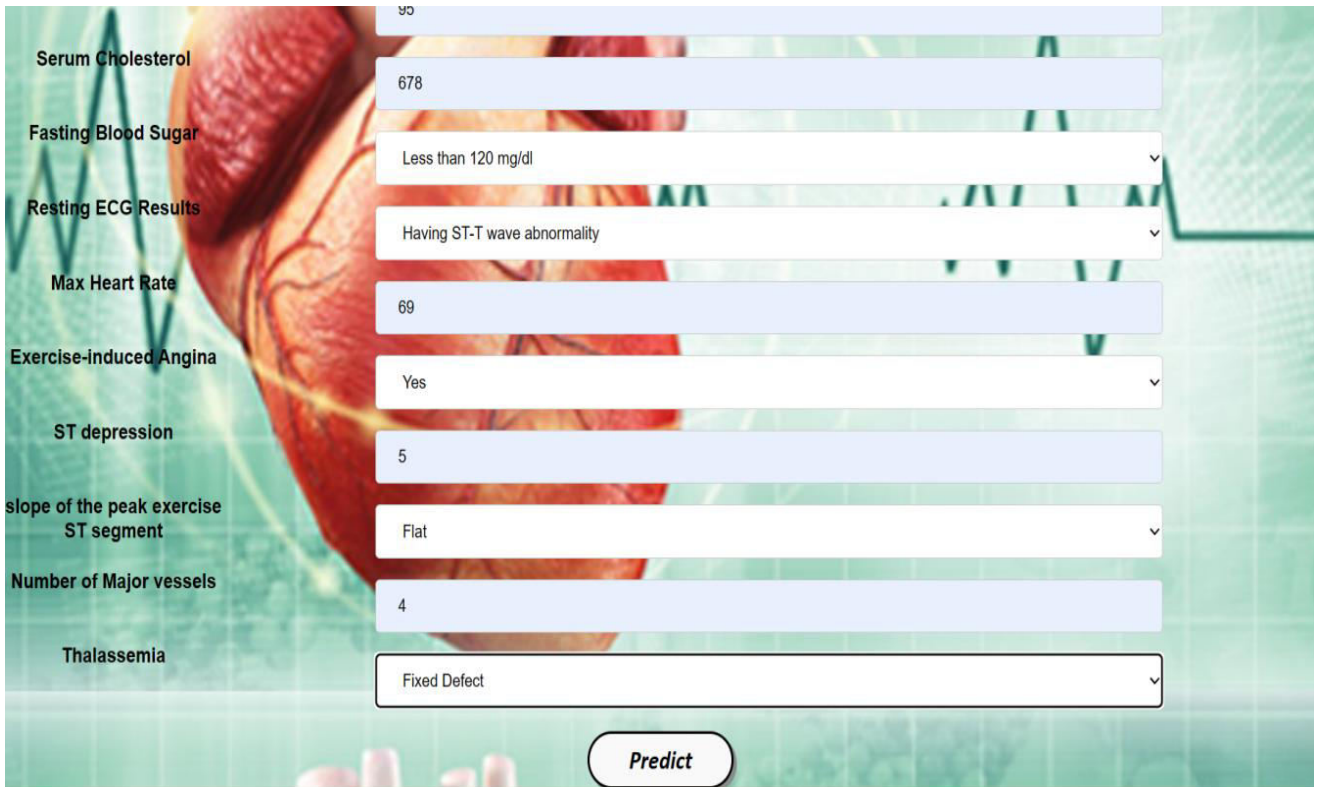


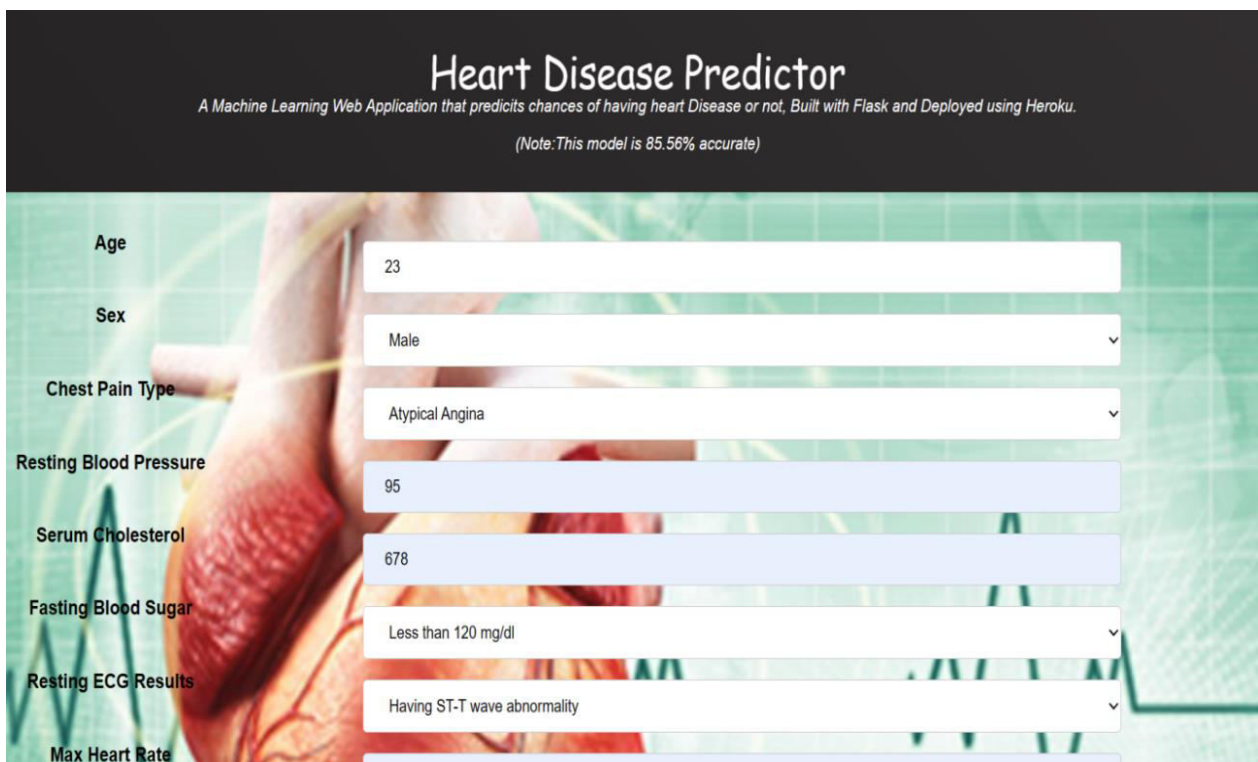
Figure.4.3 Prediction without Heart Disease

This is the result showing the patient without heart disease for the entered details.



Serum Cholesterol	678
Fasting Blood Sugar	Less than 120 mg/dl
Resting ECG Results	Having ST-T wave abnormality
Max Heart Rate	69
Exercise-induced Angina	Yes
ST depression	5
slope of the peak exercise ST segment	Flat
Number of Major vessels	4
Thalassemia	Fixed Defect

Predict



Heart Disease Predictor

A Machine Learning Web Application that predicts chances of having heart Disease or not, Built with Flask and Deployed using Heroku.
(Note: This model is 85.56% accurate)

Age	23
Sex	Male
Chest Pain Type	Atypical Angina
Resting Blood Pressure	95
Serum Cholesterol	678
Fasting Blood Sugar	Less than 120 mg/dl
Resting ECG Results	Having ST-T wave abnormality
Max Heart Rate	

Figure 4.4 User interface with entered details



Figure 4.5 Prediction with Heart Disease
Here it will predict the patient with heart disease for the entered patient's details.

The Accuracy score achieved using Random Forest is : 86.67%

```
[[38  4]
 [ 6 27]]
```

Comparitive Analysis

The accuracy score achieved using Decision Tree is: 73.33 %

Figure 4.6 The result of comparitive analysis of Random forest algorithm and decision tree

Accuracy Table

After performing the machine learning approach for training and testing we find that accuracy of the Random Forest is better compared to Decision Tree. Accuracy is calculated with the support of the confusion matrix of each algorithm, here the number count of TP, TN, FP, FN is given and using the equation of accuracy, value has been calculated and it is concluded that Random Forest is best with 86.67% accuracy and the comparison is shown below.

TABLE: Accuracy comparison of Algorithms

ALGORITHM	ACCURACY
Random Forest	86.67%
Decision Tree	73.33%

V. CONCLUSION

This paper predicts people with cardiovascular disease by extracting the patient medical history that leads to a fatal heart disease from a dataset that includes patient's medical history such as chest pain, sugar level, blood pressure, etc. This Heart Disease detection system assists a patient based on his/her clinical information of them been diagnosed with a previous heart disease. Heart diseases are a major killer in India and throughout the world, application of promising technology like machine learning to the initial prediction of heart diseases will have a profound impact on society. The early prognosis of heart disease can aid in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications, which can be a great milestone in the field of medicine. The number of people facing heart diseases is on a raise each year. This prompts for its early diagnosis and treatment. The utilization of suitable technology support in this regard can prove to be highly beneficial to the medical fraternity and patient. Therefore, in conclusion this project helps to predict the patients who are diagnosed with heart diseases by cleaning the dataset and applying classification algorithm.

REFERENCES

- [1] Yuepeng Liu, Mengfei Zhang, Zezhong Fan, Yinghan Chen "Heart disease prediction based on randomforest and LSTM" In Proceeding of Second International Conference on Information Technology and Computer Application , IEEE, pp. 630-635, 2020.
- [2] Pranav Motarwar, Ankita Duraphe, G Suganya M Premalatha" Cognitive Approach for Heart Disease Prediction using Machine Learning" In proceeding of International Conference on Emerging Trends in Information Technology and Engineering, IEEE, pp. 1-5, 2020.
- [3] N. Komal Kumar ,G.Sarika Sindhu , D.KrishnaPrashanthi, A.ShaeenSulthana. "Analysis and Prediction of Cardio Vascular Disease using Machine Learning Classifiers". In proceeding of sixth International Conference on Advanced Computing & Communication Systems, IEEE, pp.15-21, 2020.
- [4] Pronab Ghosh, Sami Azam, MirjamJonkman, (Member, IEEE), Asif Karim, F.M.Javed Mehedi Shamrat, Eva Ignatious, Shahana Shultana, Abhijith Reddy Beeravolu and Friso De Boer, "Efficient prediction of cardiovascular disease using machine learning algorithms with Relief and LASSO feature selection techniques", vol.9, pp.19304-19326, IEEE, 2021.
- [5] Vijeta Sharma, Shrinkhala Yadav, Manjari Gupta "Heart Disease Prediction using Machine Learning Techniques" In proceeding of Second International Conference on Advances in Computing, Communication Control and Networking, IEEE, pp.7281 - 8337, 2020.



Impact Factor: 8.379



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

 9940 572 462  6381 907 438  ijircce@gmail.com



www.ijircce.com

Scan to save the contact details