

Heart Disease Prediction Using Multi Feature and Hybrid Approach

Smt. Akshitha Katkeri¹, Nagashree M S², Shilpa R³, Srilakshmi N⁴, Srilalitha C S⁵

¹Assistant Professor, VTU, Department of CSE, BNM Institute of Technology, Bangalore, Karnataka, INDIA
^{2,3,4,5}VTU, Department of CSE, BNM Institute of Technology, Bangalore, Karnataka, INDIA

Abstract - Heart disease is a build-up of fatty plaques in the arteries and calcium outside the major artery. Many techniques have been used for the ailment of this problem by using various algorithms. These manual method of consultation is difficult and time consuming in severe cases. This study proposes an easy method of user-system interaction by hybrid approach of several algorithms like logistic regression, Gaussian NB, linear SVC, K-Neighbours, Decision Tree and Random Forest. In this hybrid approach the best performed algorithm is used in the final evaluation. Results: For heart disease detection, The Linear SVC model achieved best results with accuracy: 90.78%, precision: 96.87%, sensitivity: 83.78%, F1 score: 89.85%, ROC: 90.60%. Conclusion: This proposed system illustrates the use of interactive system to predict heart disease by using multi feature classification and hybrid approach which has promising results compared the previous studies and methods.

Key Words: Gaussian NB, Linear SVC, Random Forest, K-Neighbours, Decision Tree, Random Forest, Arteries.

1. INTRODUCTION

Heart disease is also known as cardiovascular disease (CVD) which remains as the number one reason for death rate globally. There are various CVD diseases, such as angina, heart failure, Coronary heart disease, congenital heart disease and so on. Nearly, 17.9 million people are losing their lives who are at the early age of 70's because of this CVD. The main risk factors of heart disease nowadays are due to unhealthy diet plans, intake of alcohol and tobacco, smoking, lack of physical activities and stress due to work. The effects of these risk factors lead to raise in blood pressure, blood lipids, overweight and so on. The other main reason for CVD is because of the building up of calcium in major artery outside the heart which is predicted as future heart attack or stroke. The more extensive the calcium in the walls of blood vessel, the greater will be the risk of future CVD.

There are several classifiers used to detect heart disease such as logistic regression, Gaussian NB, Linear SVC, Decision Tree, K-Neighbours, and Random Forest. Logistic Regression is a supervised machine learning algorithm that is used to model the probability of a certain class or an event. It is used when the data is linearly separable and its outcome is binary in nature.

Gaussian NB is a generative model. It assumes that each class follows a Gaussian distribution. It is used specifically when the features have continuous values.

Linear SVC is to fit to the data provided and resulting the best fit hyper plane which categorizes the data. After getting the hyper plane, some features can be fed to the classifier to check what the predicted class is.

Decision Tree uses various algorithms to decide to split a node into 2 or more sub-nodes. As the sub-nodes increases its purity also increases. The data is split continuously according to the specified parameters.

The main goal of this study is to develop a hybrid model of all the algorithms that best suit the prediction and make the model more accurate by people having the knowledge about their health condition much before so that they can have a proper treatment and get cured without any serious issues. Thereby, reducing the death rate globally due to heart disease.

2. METHODOLOGY

The proposed methodology aims to predict whether the patient is suffering from the heart disease or not. This automation helps doctors to analyze the critical condition of the patients. Hence it also helps in improvement of treatments. Patients can take many precautions and helps to save many lives. In this project, we are using various algorithms i.e. we are implementing by using hybrid technology with multiclass dataset. Multiclass dataset represents various levels i.e. 0, 1, 2, 3, 4. The model makes use of several machine learning techniques and algorithms in an effort to offer a more precise answer to the problem. Numerous ML techniques are used here on the data set. For instance, the K Nearest Neighbors method, Random Forest, Logistic Regression, Gaussian NB, Regression Tree, etc. A hybrid model is created employing all these techniques for increased accuracy. Additionally, the model works with practically all patient record types. Pre-processing of the dataset involves reducing noise and outliers. The dataset has now been split into train and test data. Data that is 75% trained is referred to as train data. Data deemed to be test data make up 25% of the total. The figure-based methods below are used to generate the ML models.



Fig -1: Flow Chart

The below figure shows the data flow of the proposed model:

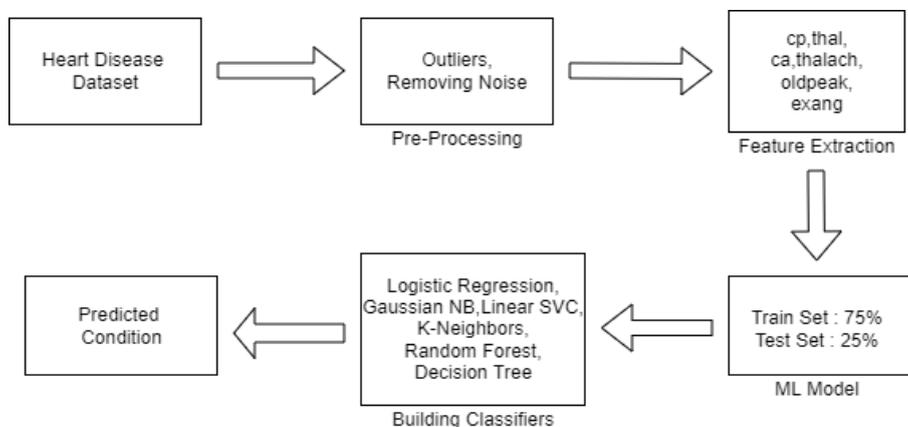


Fig -2: Proposed System

2.1 Data source and Dataset Description

In this project, the dataset is taken from UCI Repository of Machine Learning Databases. This dataset contains a total of 303 records with 14 medical features. The original values 1, 2, 3, 4 were transformed in one that is the presence of heart disease. All features have some values in the dataset. It is explained in the below table.

Sr.no	Attribute	Attribute representation	Description
1	Age	Age	Age of patients in years.
2	Sex	Sex	0 is for females, 1 is for males.
3	Chest Pain	Cp	1=typical angina 2=atypical angina 3=non-angina pain 4=asymptomatic
4	Resting bloodpressure	Trestbps	Blood pressure
5	Serum cholesterol	Chol	Minimum Cholesterol:126 Maximum Cholesterol:564
6	Fasting blood sugar	Fbs	0=false 1=true
7	Rest electrocardiograph	Restecg	0=normal 1=abnormality of ST 2=left ventricular hypertrophy
8	Max Heart rate	Thalach	Maximum heart rate achieved
9	Exercise-induced angina	Exang	0=no 1=yes
10	ST depression	Oldpeak	Exercise induced angina 0=no 1=yes
11	Slope	Slope	Slope of peak exercise 1=unsloping 2=flat 3=down sloping
12	No of vessels	Ca	Major vessels colored (0-3) by fluoroscopy
13	Thalassemia	Thal	3=normal 6=fixed defect 7=reversible defect

Table -1: Detailed Heart disease dataset attributes with the description

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slop	ca	thal	pred_attribute
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	299.000000	301.000000	303.000000
mean	54.438944	0.679868	3.158416	131.689769	246.693069	0.148515	0.990099	149.607261	0.326733	1.039604	1.600660	0.672241	4.734219	0.937294
std	9.038662	0.467299	0.960126	17.599748	51.776918	0.356198	0.994971	22.875003	0.469794	1.161075	0.616226	0.937438	1.939706	1.228536
min	29.000000	0.000000	1.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	1.000000	0.000000	3.000000	0.000000
25%	48.000000	0.000000	3.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000	0.000000	1.000000	0.000000	3.000000	0.000000
50%	56.000000	1.000000	3.000000	130.000000	241.000000	0.000000	1.000000	153.000000	0.000000	0.800000	2.000000	0.000000	3.000000	0.000000
75%	61.000000	1.000000	4.000000	140.000000	275.000000	0.000000	2.000000	166.000000	1.000000	1.600000	2.000000	1.000000	7.000000	2.000000
max	77.000000	1.000000	4.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	3.000000	3.000000	7.000000	4.000000

Table -2: Static study of the data related to heart disease

2.2 Data pre-processing

The procedure used to efficiently prepare a dataset for categorization is known as data pre-processing. There might be some missing values in the real-world data that has been gathered and saved in the database. This is the most typical issue because every patient would have entered their information incorrectly. The normalization of the attribute data fills in the missing values.

$$x_i = \frac{x_i - \mu_i}{\sigma_j}$$

Where μ = mean, σ = standard deviation, x = single value feature. Utilizing a unit mean and zero variance, the data characteristics are standardized.

2.3 Feature extraction

The goal of feature extraction is to achieve the aim by extracting a subset of new features from the original set using some functional mapping. The extremely significant characteristics are chosen for prediction once the feature significance graph is plotted for feature extraction.

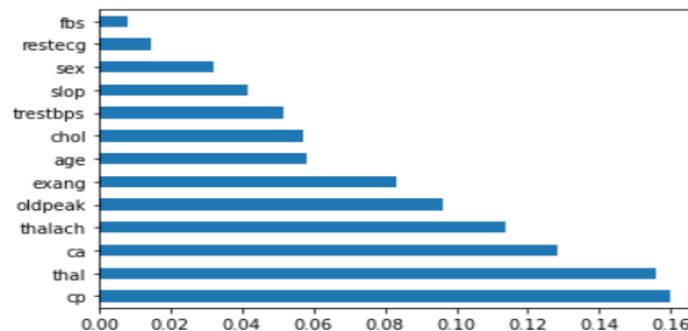


Chart 1: Feature Importance Plot

The below figure shows the attribute distribution.

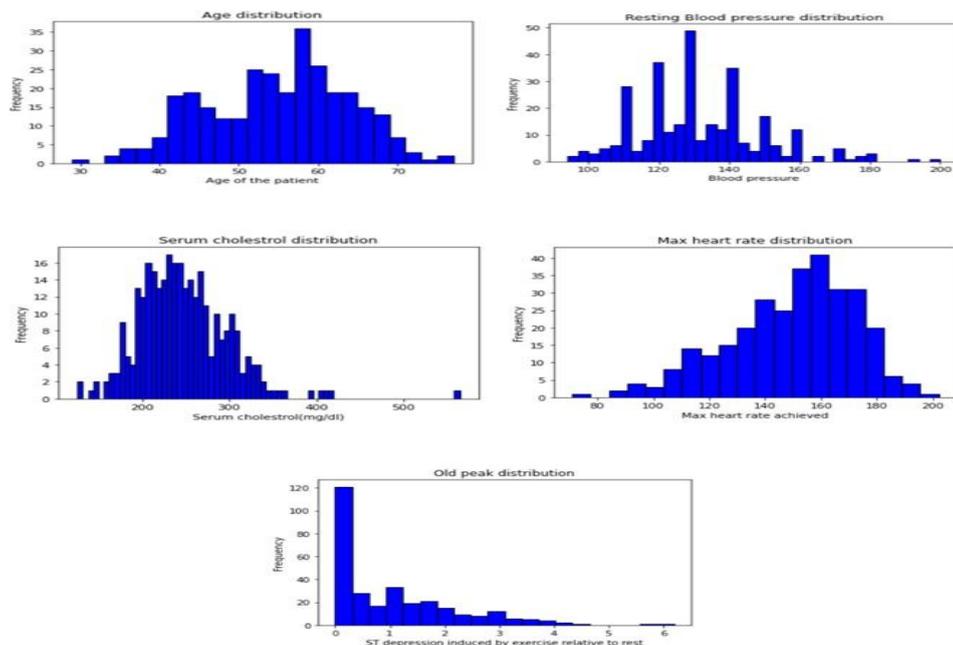


Chart 2 : Visualization of the Data

3. MODELING AND ANALYSIS

3.1 Classification

Following pre-processing, the data is separated into train data and test data. In this study, the hybrid model is proposed. On the train data, a variety of classification techniques are used to train the model. Gaussian NB, Linear SVC, LogisticRegression, Decision Tree Classifier, Random Forest Classifier, KNN, and SVM are the algorithms employed in the suggested model.

3.2 Confusion matrix

A table called a confusion matrix is used to describe how well a classification system performs. A confusion matrix shows and sums up a classification algorithm's performance. The confusion matrix for each classifier is shown below in the figures. The following is a definition of each entry in the confusion matrix:

- The total number of accurate findings or hypotheses where the real class was positive is known as the true positive rate (TP).
- The total number of inaccurate findings or forecasts made while the actual class was positive is known as the false positive rate (FP).
- The total number of accurate findings or hypotheses where the actual class was negative is known as the true negative rate (TN).
- The amount of incorrect outcomes or predictions made when the actual class was negative is known as the false negative rate (FN).

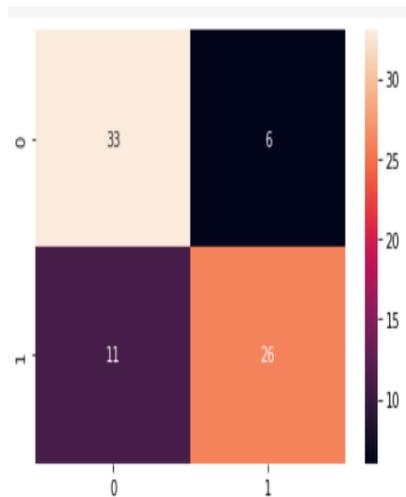


Fig 3. Random Forest classifier

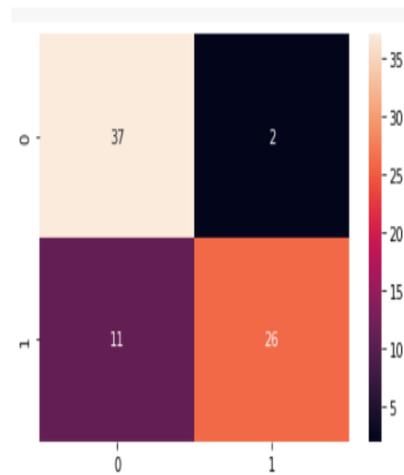


Fig 4. Logistic Regression

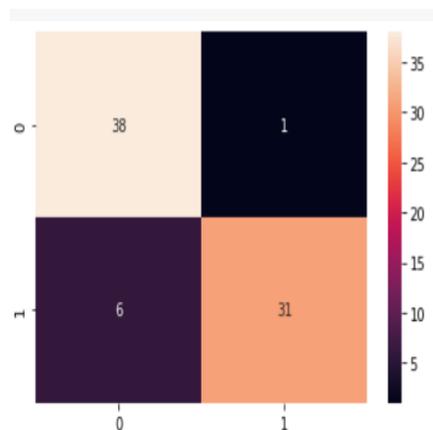


Fig 5. Gaussian NB

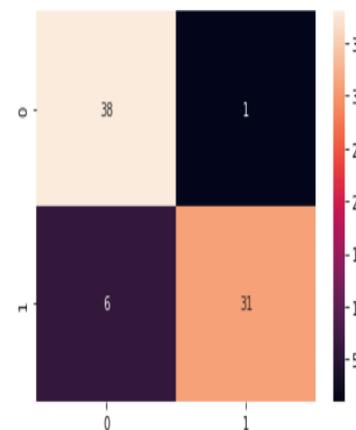


Fig 6. Linear SVC

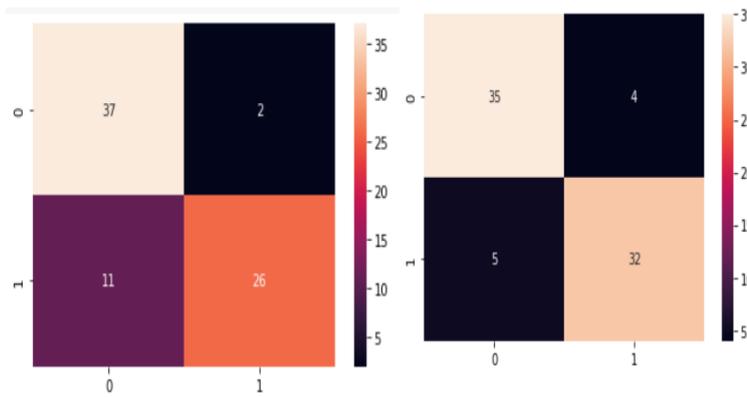


Fig 7. Decision Tree

Fig 8. KNN

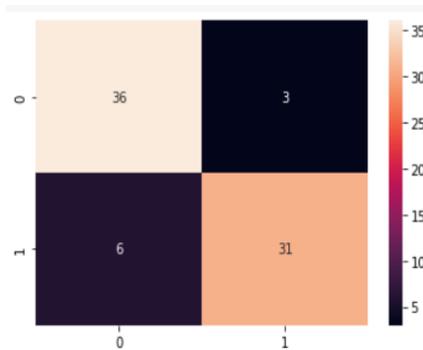


Fig 9. SVM

Table 3. Performance table of classifier algorithms

	Model	Accuracy	Precision	Sensitivity	Specificity	F1 Score	ROC	Log_Loss	mathew_corrcoef
0	Random Forest	0.828947	0.928571	0.702703	0.948718	0.800000	0.825710	5.907970	0.674986
1	Logistic Regretion	0.776316	0.812500	0.702703	0.846154	0.753623	0.774428	7.725842	0.555636
2	Gaussian NB	0.907895	0.968750	0.837838	0.974359	0.898551	0.906098	3.181214	0.822230
3	Linear SVC	0.907895	0.968750	0.837838	0.974359	0.898551	0.906098	3.181214	0.822230
4	Decision tree	0.828947	0.928571	0.702703	0.948718	0.800000	0.825710	5.907970	0.674986
5	KNN	0.881579	0.888889	0.864865	0.897436	0.876712	0.881150	4.090160	0.763094
6	SVM	0.881579	0.911765	0.837838	0.923077	0.873239	0.880457	4.090150	0.764901

The algorithm with the greatest accuracy score is taken into consideration for the prediction of the patient's heart illness based on the performance table of classifier algorithms. The classifier used for heart disease prediction uses the Linear SVC model since it provides the best accuracy compared to all other classifiers.

4. RESULTS

4.1 UI Design

A result is the final consequence of actions or events expressed qualitatively or quantitatively. Performance analysis is an operational analysis, is a set of basic quantitative relationship between the performance quantities.

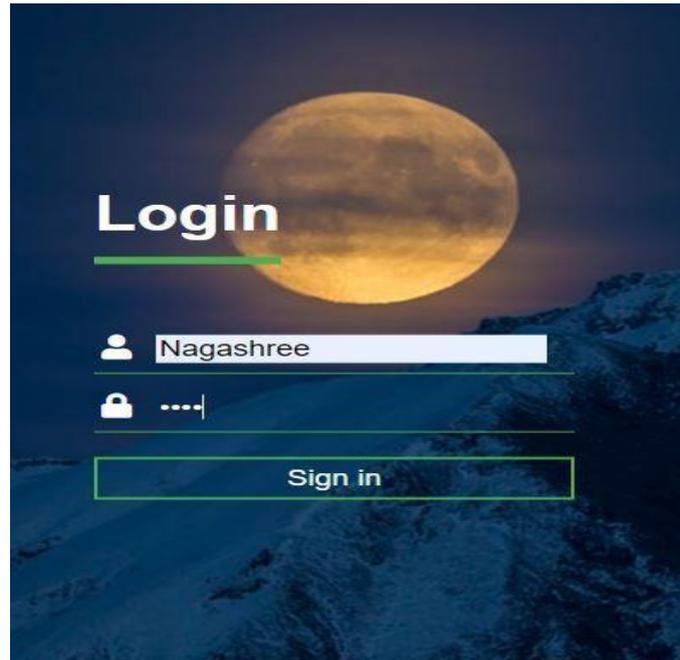


Figure 10: Login Page

This is the login page where user can enter into the website by entering name and the password set by user once registered to the website. Once the details are entered, it goes back checks whether the details given while registering match. If it matches then it allows to login into website.

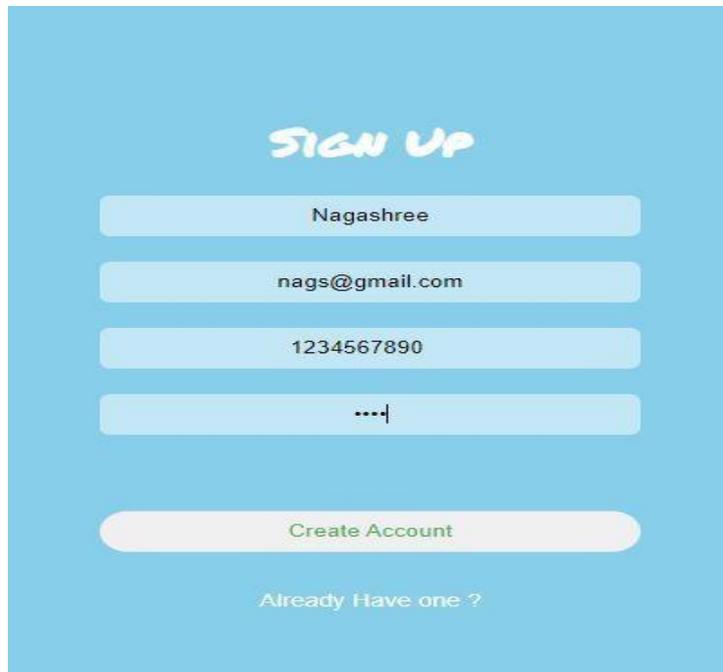


Figure 11: Sign Up Page

The above figure shows the sign-up page. If the user is visiting the website for the first time, then the user should register to the site by providing the personal information asked in the form. The user after filling the details should click on Create Account button. If the user is already registered, then the alert is shown that the user is already registered. If the user is registered, then already have one? Option is clicked. The button takes the user to login page.

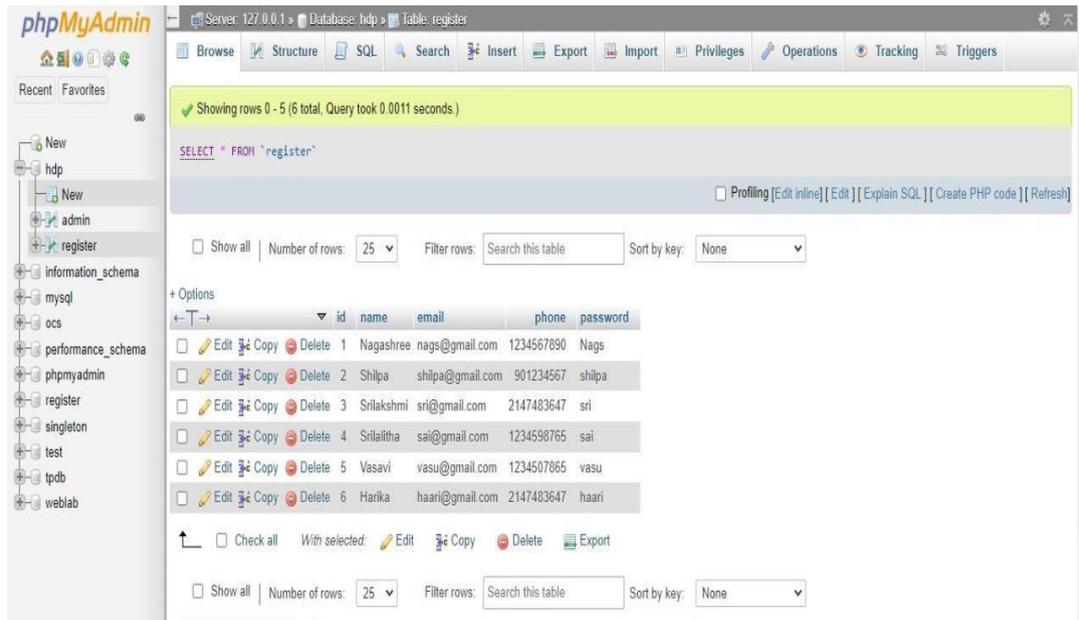


Figure 12: Database Page

The above figure shows the database page. The data provided by the user during registration is stored in the database. Each user has unique username. If the new user registers by providing the username already existing, it shows the user already exists prompting user to input the other username. The user while logging-in, has to enter the correct username and password. If the username and password provided by the user is matched with the database, then the user is successfully.

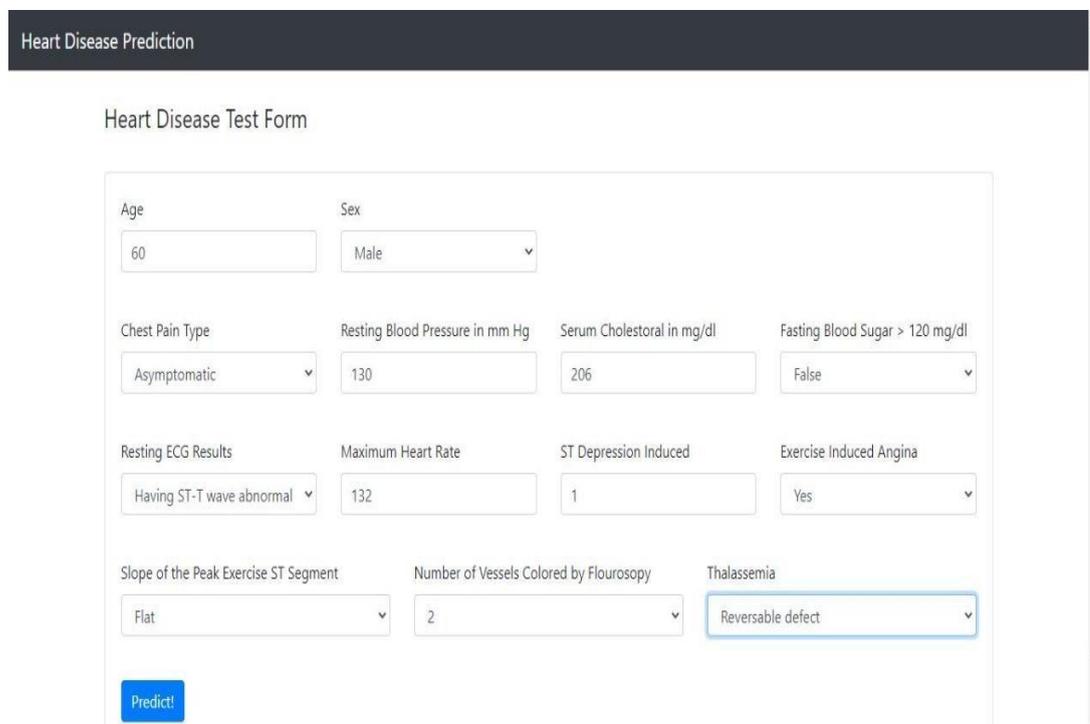
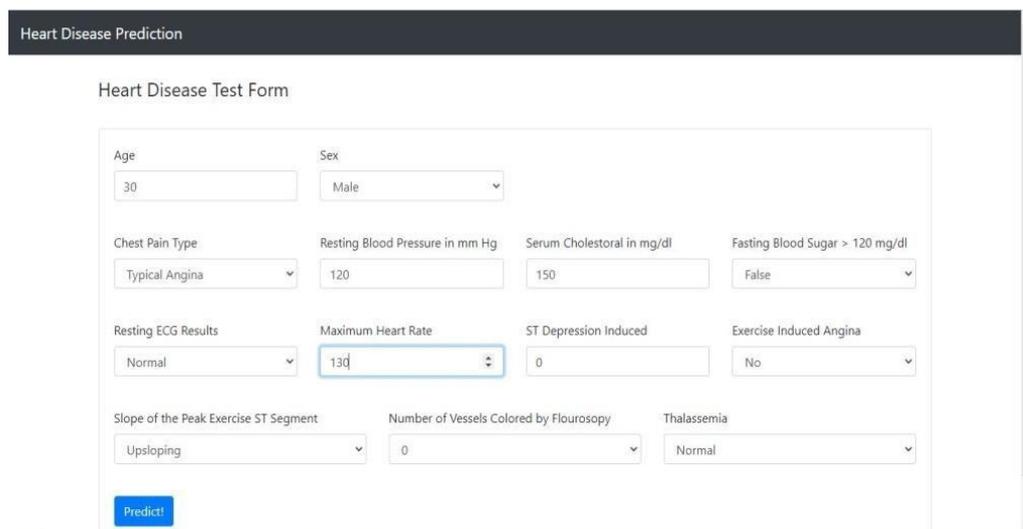


Figure 13: Input Page

The above figure is the user input page. Once the user logs-in into the website, then he is directed to this page. User has to input the data according to their health conditions by selecting from the dropdown options. Age is of integer type; the user has to input the age in numbers. The Sex field have the options male and female, the user has to choose one. The Resting Blood Pressure, Serum Cholesterol in mg/dl, Maximum Heart Rate, ST Depression induced are the integer type; the user has to input data from the medical record provided. The other fields like Chest pain type, Fasting Blood Sugar, Resting ECG Results, Exercise Induced Angina; the user select one of the options from the dropdown menu.



Figure 14: Result for Test Cases



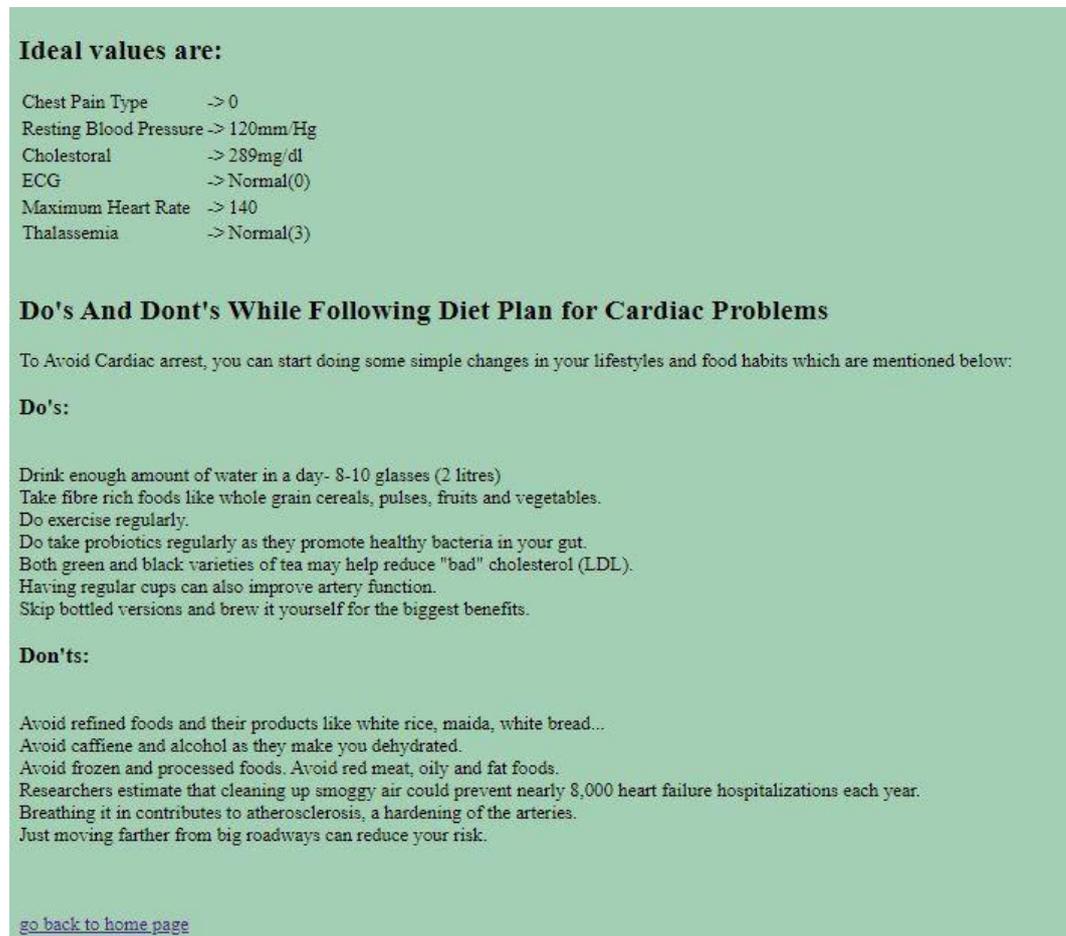
The image shows a web form titled "Heart Disease Prediction" with a sub-header "Heart Disease Test Form". The form contains several input fields: "Age" (text box with "30"), "Sex" (dropdown menu with "Male"), "Chest Pain Type" (dropdown menu with "Typical Angina"), "Resting Blood Pressure in mm Hg" (text box with "120"), "Serum Cholesterol in mg/dl" (text box with "150"), "Fasting Blood Sugar > 120 mg/dl" (dropdown menu with "False"), "Resting ECG Results" (dropdown menu with "Normal"), "Maximum Heart Rate" (text box with "130"), "ST Depression Induced" (text box with "0"), "Exercise Induced Angina" (dropdown menu with "No"), "Slope of the Peak Exercise ST Segment" (dropdown menu with "Upsloping"), "Number of Vessels Colored by Fluoroscopy" (text box with "0"), and "Thalassemia" (dropdown menu with "Normal"). A blue "Predict!" button is located at the bottom left of the form.

Fig 15. Prediction Form



Fig 16. Result for Test Cases

The above figure shows the output page. Once user gives the input and select to predict, it then redirects to the result page and after processing it gives the result whether heart disease is present or not. The page also contains the option "go back to home page" which takes the user to home page.



Ideal values are:

Chest Pain Type	-> 0
Resting Blood Pressure	-> 120mm/Hg
Cholesterol	-> 289mg/dl
ECG	-> Normal(0)
Maximum Heart Rate	-> 140
Thalassemia	-> Normal(3)

Do's And Dont's While Following Diet Plan for Cardiac Problems

To Avoid Cardiac arrest, you can start doing some simple changes in your lifestyles and food habits which are mentioned below:

Do's:

- Drink enough amount of water in a day- 8-10 glasses (2 litres)
- Take fibre rich foods like whole grain cereals, pulses, fruits and vegetables.
- Do exercise regularly.
- Do take probiotics regularly as they promote healthy bacteria in your gut.
- Both green and black varieties of tea may help reduce "bad" cholesterol (LDL).
- Having regular cups can also improve artery function.
- Skip bottled versions and brew it yourself for the biggest benefits.

Don'ts:

- Avoid refined foods and their products like white rice, maida, white bread...
- Avoid caffeine and alcohol as they make you dehydrated.
- Avoid frozen and processed foods. Avoid red meat, oily and fat foods.
- Researchers estimate that cleaning up smoggy air could prevent nearly 8,000 heart failure hospitalizations each year.
- Breathing it in contributes to atherosclerosis, a hardening of the arteries.
- Just moving farther from big roadways can reduce your risk.

[go back to home page](#)

Figure 17: Diet Plan Page

The above figure is directed from the results page to diet plan page. Based on the results of the user and how risky the disease is the system provides a diet plan to the user to maintain health conditions. By following the diet plan user can bring his health conditions from severe to normal.

4.2 Research Implications

The proposed methodology aims to predict whether the patient is suffering from the heart disease or not. This automation helps doctors to analyze the critical condition of the patients. Hence it also helps in improvement of treatments. A user interface is created to take the input from the user and the model predicts the presence of heart disease and recommends diet plans. This is useful for improving the user's health effectively.

5. CONCLUSION

Heart Disease Prediction is a very common problem now. This proposed user interface platform helps everyone to register and login and give in the data and get to know their health status. Based on the given data it predicts whether the heart disease is present or not. This proposed system helps to identify disease in a very early stage to prevent death rate. A database is also created so that all the patient's data can be stored. It is a hybrid system approach successfully used for heart disease prediction with higher accuracy rate.

6. FUTURE ENHANCEMENTS

In future, model for direct service of the patients from the old age homes or other home care centers to the Intensive Care Unit (ICU) through ambulance services can be planned. An artificially intelligent system will take the data of clinical parameters from old age homes or other care centers. The model gets the single output that will reveal distinct stages of patients in terms of healthy, first/second stage of sickness and critical stage. The system will show green color if the status of the person is healthy, and the respective person will be informed via SMS that you are 'Healthy'. Otherwise, if the person is at the first/second stage of sickness, then a SMS 'Do frequent monitoring' will be sent to his/her mobile number.

6. REFERENCES

- [1] Pronab Ghosh, Sami Azam, Mirjam Jonkman, Asif Karim , F. M. Javed Mehedi Shamrat, Eva Ignatious, Shahana Shultana, Abhijith Reddy Beeravolu, Friso De Boer, "Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms with relief and LASSO Feature Selection Techniques", 10.1109/ACCESS.2021.3053759, VOLUME 9, 2021
- [2] Tsatsral Amarbayasgalan, Van-Huy Pham, Nipon Theera-Umpon (Senior Member, Ieee), Yongjun Piao And Keun Ho Ryu (Life Member, Ieee), "An Efficient Prediction Method for Coronary Heart Disease Risk Based on Two Deep Neural Networks trained on well-ordered training dataset", 10.1109/Access.2021.3116974, Volume 9, 2021
- [3] Norma Latif Fitriyani, Muhammad Syafrudin, Ganjar Alfian, (Member, Ieee), And Jongtae Rhee, "HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System", 10.1109/ACCESS.2020.3010511, VOLUME 8, 2020.
- [4] Sarria E. A. Ashri, M. M. El-Gayar , And Eman M. El-Daydamony, "HDPF: Heart Disease Prediction FrameworkBased on Hybrid Classifiers and Genetic Algorithm", 10.1109/ACCESS.2021.3122789, Volume 9, 2021.
- [5] Aqsa Rahim, Ghulam Ishaq Khan, Yawar Rasheed, Farooque Azam, Muhammad Waseem Anwar, "An Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases", 10.1109/ACCESS.2021.3098688, IEEEAccess 2021.