Heart Disease Prediction

Kruthi
Computer Science and Engineering
PES University
Bangalore, India
kruthi163@gmail.com

S Niharika
Computer Science and Engineering
PES University
Bangalore, India
niharikasurapuram@gmail.com

Avantika Padmaraj Hombannavar
Computer Science and Engineering
PES University
Bangalore, India
avantikahombannavar@gmail.com

***Abstract*— Heart disease is the leading cause of death in the world. Hence heart disease prediction requires perfection and exactness for diagnosis and analyses. In this paper we carried out research on heart disease from data analytics point of view. We used data analytics to detect and predict the disease. Starting with a pre-processing phase, where we analyzed the most relevant features by the correlation matrix, followed by various Machine Learning algorithms such as SVM, KNN, XGB, ensemble and stacking were used for the development of the model. Results show that stacking, which is a combination various models (level wise) gave the highest accuracy of 98.93%**

Keywords—Machine Learning · Heart Disease · prediction

I. Introduction

Cardiovascular diseases (CVDs) is one of the deadliest diseases and the principal cause of death globally. An estimated 17.9 million people died from CVDs in 2019, representing 32% of all global deaths. Of these deaths, 85% were due to heart attack and stroke. Heart attacks and strokes are usually acute events and are mainly caused by a blockage that prevents blood from flowing to the heart or brain.

Due to prolonged years of exposure to unhealthy lifestyles, heart disease clinically presents itself in the early stages of life, as well as in old age. Most cardiovascular diseases can be prevented by addressing behavioral risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity, and harmful use of alcohol. Thus, it is very important and concerning to detect cardiovascular disease as early as possible so that management with counseling and medicines can begin. But this is not easy, especially in developing countries, due to the lack of diagnostic centers, resources, and qualified doctors. In addition, it is also more expensive and computation intensive, and it takes a lot of time for evaluations to be carried out

Machine learning is playing an essential role in the medical side. Recently, to solve difficult issues, a range of data mining techniques and machine learning techniques are built and applied.

Our study of this problem is part of data science applications, where we detect cardiac patients based on necessary attributes and predict whether the patient tested has heart disease or not. We have utilized supervised learning techniques for predicting heart disease at an initial phase. This paper compares various machine learning algorithms such as logistic regression, k-nearest neighbor (KNN), support vector machine (SVM), adaboost, gradient boosting, ensemble, and stacking.

The rest of this paper is structured as follows: Section 2 discusses the literature review, existing methods and techniques available. Section 3 describes the proposed methodology. In Section 4, experimental results and the comparison between classification techniques are presented. Finally, Section 5 ends with a conclusion of current work and some notes on future enhancement.

II. Literature Review

In previous studies, researchers have put in their efforts in finding the best model for predicting heart diseases. Most of the past research looked into identifying features that contribute to better heart prediction accuracy. The table below presents some of the works done by researchers

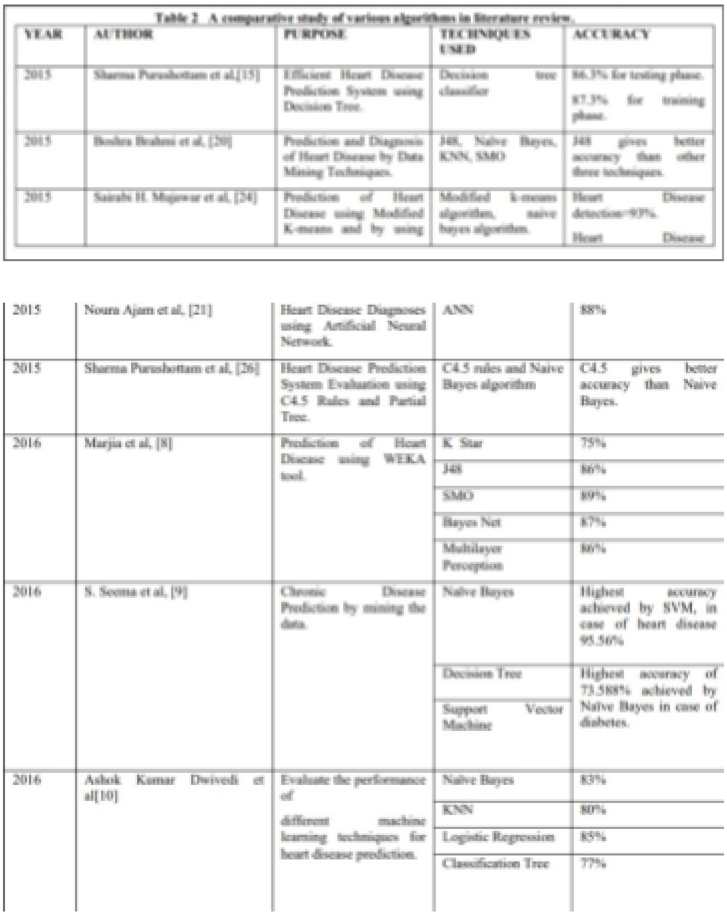


Table 2 A comparative study of various algorithms in literature review.

YEAR	AUTHOR	PURPOSE	TECHNIQUES USED	ACCURACY
2015	Sharma Purushottam et al.[15]	Efficient Heart Disease Prediction System using Decision Tree.	Decision tree classifier	86.3% for testing phase. 87.3% for training phase.
2015	Boshra Brahmi et al, [20]	Prediction and Diagnosis of Heart Disease by Data Mining Techniques.	J48, Naïve Bayes, KNN, SMO	J48 gives better accuracy than other three techniques.
2015	Sairabi H. Mujawar et al, [24]	Prediction of Heart Disease using Modified K-means and by using	Modified k-means algorithm, naive bayes algorithm.	Heart Disease detection=93%. Heart Disease
2015	Noura Ajam et al, [21]	Heart Disease Diagnoses using Artificial Neural Network.	ANN	88%
2015	Sharma Purushottam et al, [26]	Heart Disease Prediction System Evaluation using C4.5 Rules and Partial Tree.	C4.5 rules and Naive Bayes algorithm	C4.5 gives better accuracy than Naive Bayes.
2016	Marjia et al, [8]	Prediction of Heart Disease using WEKA tool.	K Star	75%
			J48	86%
			SMO	89%
			Bayes Net	87%
			Multilayer Perception	86%
2016	S. Seema et al, [9]	Chronic Disease Prediction by mining the data.	Naïve Bayes	Highest accuracy achieved by SVM, in case of heart disease 95.56%
			Decision Tree	Highest accuracy of 73.588% achieved by Naïve Bayes in case of diabetes.
			Support Vector Machine	
2016	Ashok Kumar Dwivedi et al[10]	Evaluate the performance of different machine learning techniques for heart disease prediction.	Naïve Bayes	83%
			KNN	80%
			Logistic Regression	85%
			Classification Tree	77%

III. Proposed Methodology

The proposed methodology is shown in the diagram Fig1:

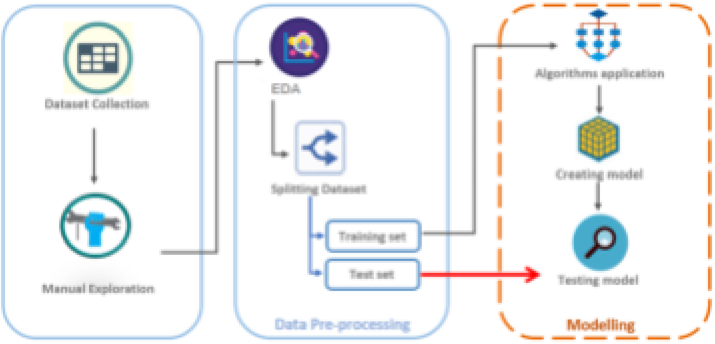
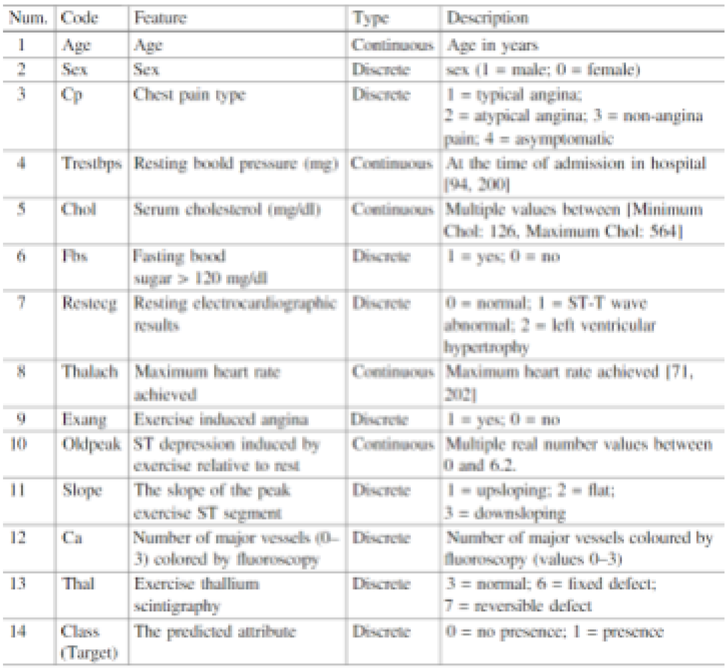


Fig 1: proposed methodology

A. *Dataset*

We have used the well-known Cleveland dataset, collected from UCI machine learning repository, which is available on Kaggle [5]. The Cleveland database was selected for this research because it is a commonly used database for machine learning researchers with comprehensive and complete records. In this field, the dataset is a collection of medical analytical reports with a total of 303 records with 14 medical features with the last column giving the output variable. The medical attributes and their description are shown in Fig 2.



Num.	Code	Feature	Type	Description
1	Age	Age	Continuous	Age in years
2	Sex	Sex	Discrete	sex (1 = male; 0 = female)
3	Cp	Chest pain type	Discrete	1 = typical angina; 2 = atypical angina; 3 = non-angina pain; 4 = asymptomatic
4	Trestbps	Resting boold pressure (mg)	Continuous	At the time of admission in hospital [94, 200]
5	Chol	Serum cholesterol (mg/dl)	Continuous	Multiple values between [Minimum Chol: 126, Maximum Chol: 564]
6	Fbs	Fasting bood sugar > 120 mg/dl	Discrete	1 = yes; 0 = no
7	Restecg	Resting electrocardiographic results	Discrete	0 = normal; 1 = ST-T wave abnormal; 2 = left ventricular hypertrophy
8	Thalach	Maximum heart rate achieved	Continuous	Maximum heart rate achieved [71, 202]
9	Exang	Exercise induced angina	Discrete	1 = yes; 0 = no
10	Oldpeak	ST depression induced by exercise relative to rest	Continuous	Multiple real number values between 0 and 6.2.
11	Slope	The slope of the peak exercise ST segment	Discrete	1 = upsloping; 2 = flat; 3 = downsloping
12	Ca	Number of major vessels (0–3) colored by fluoroscopy	Discrete	Number of major vessels coloured by fluoroscopy (values 0–3)
13	Thal	Exercise thallium scintigraphy	Discrete	3 = normal; 6 = fixed defect; 7 = reversible defect
14	Class (Target)	The predicted attribute	Discrete	0 = no presence; 1 = presence

Fig 2 : Detailed information of attributes

B. *Data Pre-processing*

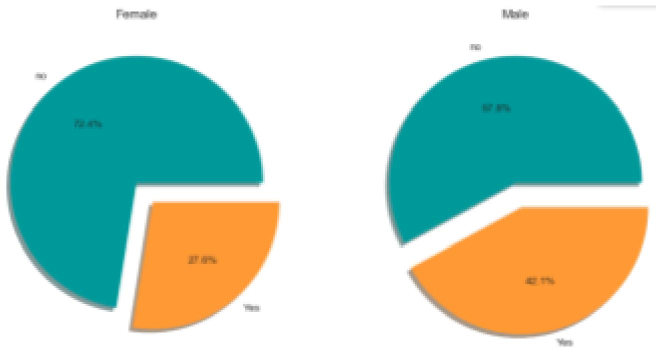
1) Data Cleaning: The dataset had no missing values.

2) Exploratory Data Analysis:

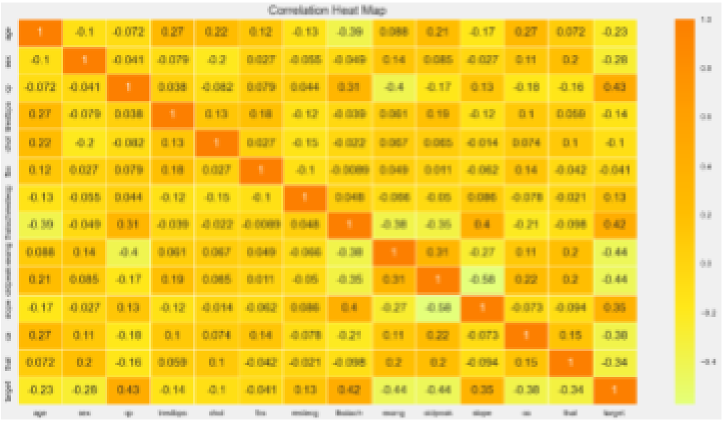
- Univariate analysis - For categorical features bar plots are used to calculate the number of each category in a particular variable. For numerical features, probability density plots are used to look at the distribution of the variable.

Fig 3 shows that the probability of a male having heart disease is greater than that of a female.

Fig 3



- Bivariate analysis – Between various features and the target
- Multivariate analysis – We used correlation matrix to identify the important features. It showed cp,thal, and slope.



3) Data splitting:

We have used 80% training dataset and 20% dataset used as testing dataset the system.

4) Standardization of Data:

The dataset was found to be imbalanced. Hence we applied a resampling technique where over-sampling of the minority category data is done.

Then we have applied feature scaling to standardize the continuous features present in the data in a fixed range.

C. Modeling

We have used 7 models and they are as follows

a)Logistic Regression:

LR is the supervised ML learning method. It is established on the association between dependent and independent variable as seen in Fig.5 variable “a” and “b” are dependent variable and independent variable and relation between them is shown by equation of line which is linear in nature that why this approach is called linear regression. It gives a relation equation to predict a dependent variable value “b” based on a independent variable value “a” as we can see in the Fig.5 so it is concluded that linear regression technique give the linear relationship between a(input) and b(output).

b)SVM:

The objective of the SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. The dimension of the hyperplane depends upon the number of features.

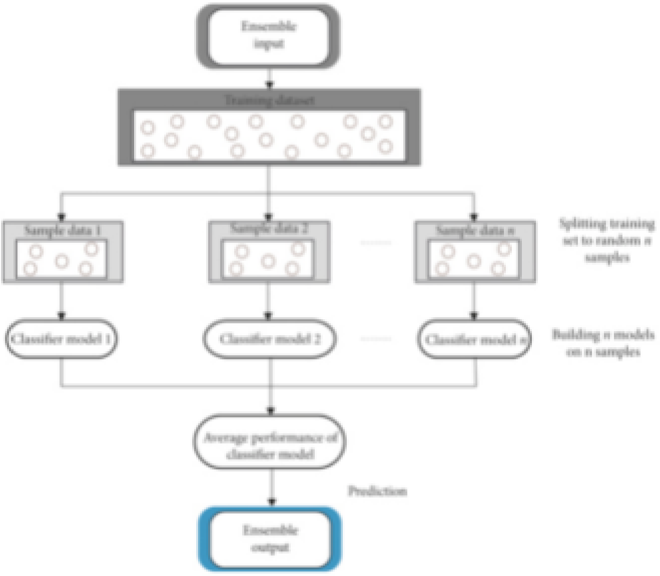
c)KNN: is a nonparametric technique of lazy learning to enable the prediction of the new sample classification. It is utilized in several groups. It can be utilized in both the forecast problems of regression and classification. However, it is often utilized in classification when it applies to industrial problems as it fairs across all criteria examined when assessing a technique's functionality, but it is utilized mostly because of its ease of understanding and lower computation time

d)Adaboost:

An AdaBoost classifier is an ensemble method that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases.

e) Ensemble model:

Ensemble techniques are methods that can be utilized to enhance the performance of a classifier. It is an effective classification method that combines a weak classifier with a strong classifier to improve the weak learner's efficiency . The ensemble technique is used in the proposed technique to enhance the accuracy of various algorithms for diagnosing heart disease. Compared to an individual classification, the purpose of combining multiple algorithms is to obtain better performance.We have included 4 weak learners in our model- Decision tree, KNN, Adaboost and Naïve Bayes algorithms.



f) XGB:

XGBoost is a popular and efficient open-source implementation of the gradient-boosted trees algorithm. The objective here is to minimize this loss function by adding weak learners using gradient descent. Each predictor is trained using the residual errors of the predecessor as labels.

e) Stacking:

is an ensemble learning technique that combines multiple base classification model predictions into a new data set. This new data are treated as the input data for another classifier. We have included 4 learners in our model- Gradient boost, Decision trees, Random forest and Logistic regression.

Level-0 Models (Base-Models): Models fit on the training data and whose predictions are compiled.We have used XGB, Decision tree and random forest as the base models.

Level-1 Model (Meta-Model): Model that learns how to best combine the predictions of the base models. We have used logistic regression for level 1

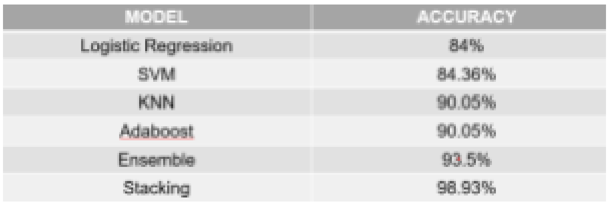
IV. Results

For evaluation metrics confusion matrix and accuracy.

Accuracy is one of the most important performance metrics for classification. It is defined as the proportion between the correct classification and the total sample, as shown in the following equation:

$$\mathrm{Acc} = \frac{(\mathrm{TP} + \mathrm{TN})}{(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})}.$$

The table below shows the accuracy we have got for various different machine learning algorithms.



MODEL	ACCURACY
Logistic Regression	84%
SVM	84.36%
KNN	90.05%
Adaboost	90.05%
Ensemble	93.5%
Stacking	98.93%

The proposed algorithm stacking shows the highest accuracy of 98.93% .

V. Conclusions

Heart disease prediction is challenging and very important in the medical field. However, the mortality rate can be drastically controlled if the disease is detected at the early stages and preventative measures are adopted as soon as possible. Further extension of this study is highly desirable to direct the investigations to real-world dataset.

References

[1] Salhi, Dhai Eddine, Abdelkamel Tari, and M. Kechadi. "Using machine learning for heart disease prediction." *International Conference on Computing Systems and Applications*. Springer, Cham, 2020.

[2] Boukhatem, Chaimaa, Heba Yahia Youssef, and Ali Bou Nassif. "Heart Disease Prediction Using Machine Learning." 2022 Advances in Science and Engineering Technology International Conferences (ASET). IEEE, 2022.

[3] Gao, Xiao-Yan, et al. "Improving the accuracy for analyzing heart diseases prediction based on the ensemble method." *Complexity* 2021 (2021).

[4] Mohan, Senthilkumar, Chandrasegar Thirumalai, and Gautam Srivastava. "Effective heart disease prediction using hybrid machine learning techniques." *IEEE access* 7 (2019): 81542-81554.

[5] https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset