IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Early Prediction Of Heart Disease

Sauleh Shabir
Department of Computer Science
Presidency University,
Bangalore, India
saulehshabirsk7@gmail.com

Sufiyan Ahmed Mujawar
Department of Computer Science,
Presidency University
Bangalore, India
sufiyanahmed4902@gmail.com

Sagarika Das
Department of Computer Science
Presidency University
Bangalore, India
dassagarika1932@gmail.com

Abstract — In the realm of healthcare and medical research, early disease prediction stands as a pivotal pillar for improving patient prognosis and curbing healthcare expenses. This study delves into the application of various machine learning algorithms like as Random Forest, Logistic Regression, Support Vector Machines (SVC), Decision Trees, and Gradient Boosting—for the accurate prediction of heart diseases. The primary goal is to develop robust predictive models that effectively analyze medical data, and contribute to timely identification and precise prognosis of cardiovascular conditions. This report synthesizes findings from an extensive literature review that scrutinizes multiple studies related to and around heart disease prediction using diverse machine learning methodologies. The analysis encompasses distinct approaches and datasets, showcasing the performance of algorithms in predicting heart diseases based on varying parameters and attributes. The comprehensive comparative analysis and evaluation conducted in this report aim to determine the superior-performing algorithms for heart disease prediction, contributing significantly to enhanced diagnostic precision and better patient care. This synthesis of various studies underscores the pivotal role of machine learning in revolutionizing healthcare, providing a roadmap for optimized models and potential real-world applications in heart disease diagnosis and prognosis.

Keywords—Early Heart Disease Prediction, Machine Learning, , KNN, SVC, Logistic Regression, Decision Trees, Gradient Boosting, Random Forest, Tkinter (GUI).

I. INTRODUCTION

Healthcare is the top focus for humanity. WHO guidelines states, good health is important for all individuals. It is said that appropriate health care services should be made available and one should always opt for regular checkup. A high number of almost 31% of deaths are due to heart caused diseases from all over the world. Early detection [1] and treatment of several heart diseases is very important and complex, especially in remote areas, because of the inadequate number of qualified doctors and diagnostic centers and various other resources that affect the accurate prognosis of heart disease.

Data collected from healthcare is generally huge in number and volumes and complex in structure. ML algorithms are able to handle the big data and mine them to find the meaningful information. The algorithms learn from previous data to do prediction on real data. This sort of machine learning model for such illness detection can encourage cardiologists to take quick actions, so the patients can get medicines within a small timeframe, thus saving a lot of lives

II. LITERATURE REVIEW

- Authors in [1] explore heart disease analysis and prediction using machine learning algorithms like k-nearest neighbor, decision tree, linear regression. It discusses dataset analysis, algorithm comparisons, and concludes that k-nearest neighbor performs best with 87% accuracy. The study emphasizes the importance of accurate predictions for early disease detection and hints at future improvements on approaches for machine learning for better heart disease prognosis.
- Authors in [5] focused on leveraging machine learning techniques for the heart disease prediction using a benchmark dataset containing 14 parameters related to heart health. They explored four prominent machine learning algorithms Decision Tree, Naive Bayes, and Random Forest - to develop a predictive model. Their aim was to analyze the performance and accuracy of these models in diagnosing heart diseases based on individual attributes and indicators. Ultimately, they found that the Random Forest algorithm exhibited the highest accuracy from the lot of 99%, outperforming the other algorithms tested in their dataset.
- Authors in [2] introduces a novel approach of using machine learning to predict heart disease by employing a hybrid model combining Decision Trees with Random Forest algorithms. They utilize the Cleveland heart disease dataset, showcasing accuracy levels of 88.7% through the hybrid model. The study demonstrates the significance of machine learning in healthcare, aiming to provide early detection and accurate predictions for cardiovascular diseases, addressing a critical health threat worldwide. The work explores various existing studies in heart disease prediction, highlighting the need for optimized models, and proposes a hybrid algorithm for enhanced accuracy, paving the way for potential real-world applications in heart disease diagnosis.
- Authors in [4] use an application using machine learning, particularly neural networks, to predict heart disease vulnerability based on easily accessible symptoms like age, sex, blood pressure, heart rate, etc. They've analyzed prevalent sensors

like Fitbit and Health Gear for data collection, employing multilayer perceptron (MLP) algorithms to train and test datasets. Results indicate high precision in predicting heart disease, aiming to provide early detection and reduce heart-related fatalities. Future work suggests expanding similar systems for other diseases using evolving technologies like big data and cloud computing.

III. METHOLOGY

A. Data collection

Data collection series is defined as the fact the system that gathers, gauge and examine the precise and accurate for probe. A probe can compare their model on the basis of gathered information. In most of the cases, information series is that the only and big step for probe, irrespective of the dimensions of probe. In our study, we use accurate dependent information set of Kaggle dataset provided in Table - I.

B. Data Preprocessing

Before beginning the model and algorithm of Machine Learning, we put gather the information that is to be implemented; this segment is completed in the steps given Features chosen are primarily given totally with the correlation matrix. we've got fourteen attributes in Table – I which are collected and arranged with respect to each other. The information of the chosen functions are defined in Table 1.

TABLE I. ATTRIBUTES

S. No.	Attribute	Desc.	Mean Value
1	age	in years	54.434
2	Sex	Male, Female	0.696
3	cp	Angina, abnang, notang, asympt	0.942
4	trstbps	Resting Blood Pressure in mm hg	131.612
5	Chol	Serum Cholesterol in mg/dl	246
6	fibs	fasting blood sugar- 1 if >120 mg/dl, 0 if <120 mg/dl	0.149
7	restecg	Electrocardiographic Results	0.53
8	thalach	Maximum Heart Rate observed	149.114
9	exang	exercise with angina has occurred	0.337
10	oldpeak	ST depression induced through exercise	1.072
11	slope	slope of the ST segment	1.385
12	thal	Number of major vessels ranging from 0 - 3 color by fluoroscopy	0.754
13	ca	Heart status	2.34
14	Target	Output Class	

C. Data Visualiztion

- Correlation Matrix and Heatmap:

A correlation matrix is a table which defines the correlation coefficients between variables. It's a useful way to find relationships, patterns in your dataset.
The correlation coefficient is a statistical measure that shows the extent to which two variables change together. It ranges from -1 to 1, where.
A heatmap is a graphical map of data points where values are shown in a matrix as colors. In a correlation matrix, the heatmap visually dipicts the strength and direction of the relation between attributes using a scale color.

- Histograms:

Histograms is a graphical representations of dataset distribution. They provide meaningfull insights into the underlying distribution of a continuous variable.

- Dataset Balance Check - Countplot:

This helps check the balance of the target variable, especially during classification tasks. A well balanced dataset has roughly same number of instances of a unique class, while an imbalanced dataset may have significantly more number of instances of a class than the others.
A countplot is a type of bar plot that displays the tally of observations from each category. It can be used to visualize the distribution of the target variables. Each bar represents the count of variables from each class.

D. Classification Algorithms used

- Logistic regression :

Logistic Regression is a modelling using analytical technique. It is used to analyse a dataset in which there is one or more independent attributes that decide a result. Logistic Regression is usually imported with a state value of 0. And then the training model is fitted and the prediction is done. The testing accuracy was 79%

- KNN Classifier :

K-nearest neighbors algorithm is powerful classification model generally utilized for grouping common variable using pattern recognition. It is widely used in prediction analysis. The algorithm identifies existing data points that are nearest to it. Using ‘sklearn.neighbors’, ‘KNeighbors Classifier’ is imported with kn_neighbors = 1. Then the training model is fitted and the prediction is done. The testing accuracy was 74%.

- Support vector machine :

Support Vector Machine or SVM is one of the famous Supervised Learning Techniques in ML. The benefits of this algorithm is that it creates the best suitable line or boundary based on decision take that can separate a n-dimensional space into vide classes so that we can easily verify the newly added data points in the correct category. From ‘sklearn’, ‘svm’ is imported and the kernel is kept as a linear and gamma as auto and C = 2. And the training model is fitted and the prediction is done. The testing accuracy was 80%.

- Random forest :

Random forest classifier is a very useful supervised classification tool. RF generates a forest of trees for a given dataset, rather than a single tree. Each of these trees make a classification from a given set of attributes. From ‘sklearn.ensemble’, ‘Random Forest Classifier’ is imported. The n_estimators is kept at 10 and random state at 0. Then the training model is fitted and the prediction is done. The testing accuracy was 87%.

- Decision Tree :

A Decision tree is a tree like diagram, the internal nodes represent values of an attribute, each branch denotes the outcome of the decision, each leaf node denotes a outcome. Decision Tree is imported where the random state was kept as 0 and then the training model was fitted and the prediction is done. The testing accuracy was 73%.

- Gradient Boosting :

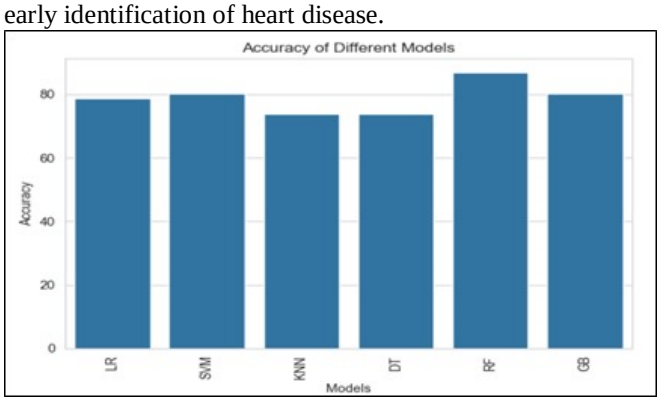
Gradient Boosting model encompasses several key metrics. The accuracy score is computed to gauge the overall correctness of the model's predictions. Additionally, a confusion matrix is employed to dissect the outcomes into true positives, true negatives, false positives, and false negatives, providing a more nuanced understanding of the model's performance.

RESULT

We found that the various machine learning models performed significantly differently in our experiment. Specifically, Random Forest and Support Vector Machines (SVM) have shown outstanding efficacy, surpassing both Gaussian naïve branches and decision trees.

With an impressive 87% accuracy percentage, the Random Forest model surpassed SVM by 6.8% and decision trees by almost 14 percent. These significant differences in performance indicate Random Forest's superior predictive capacity over other models.

In particular, Random Forest once again showed itself as the best model with a prediction accuracy of 87 percent when it succeeded in prediction of early heart disease. Not only did it perform better than SVM, which was the second best performing model, but it also demonstrated how much better Random Forest and SVM were in making predictions for the early identification of heart disease.



CONLCUSION

Together, studies [1], [5], [2] and [4] highlight the growing importance of machine learning in heart disease prognosis, with the aim of early detection and accurate prognosis. The various approaches used, from traditional algorithms such as k-nearest neighbor and decision trees to more advanced techniques such as hybrid models and neural networks, represent the evolving landscape of predictive modeling in healthcare. In particular, the high accuracy of these studies, up to 99%, reflects the potential of machine learning to greatly advance the diagnosis of heart disease.

Although each study emphasizes the importance of accurate predictions for early disease detection, the choice of algorithms varies: [5] emphasizes the dominance of Random Forest, [1] recommends the use of k-nearest neighbor, and [2] offers a hybrid. model In addition, [4] introduces the use of neural networks and sensor data for prediction and introduces the integration of modern technologies such as Fitbit and Health Gear.

Together, these findings suggest a promising future for machine learning applications in healthcare, particularly in the treatment of cardiovascular disease. The research not only promotes an ongoing dialogue about optimized models, but also paves the way for potential real-world applications and advances in heart disease diagnosis. As the field continues to advance, these multifaceted approaches will provide valuable insights into the ongoing search for accurate, effective, and easy-to-use methods to predict and prevent heart disease..

REFERENCES

[1] A. Singh and R. Kumar, "Heart Disease Prediction Using Machine Learning Algorithms," 2020 International Conference on Electrical and Electronics Engineering (ICE3), Gorakhpur, India, 2020, pp. 452-457, doi: 10.1109/ICE348803.2020.9122958.

[2] Modepalli, Kavitha & Gnaneswar, G. & Dinesh, R. & Sai, Y. & Suraj, R.. (2021). Heart Disease Prediction using Hybrid machine Learning Model. 1329-1333. 10.1109/ICICT50816.2021.9358597.

[3] Gavhane, Aditi et al. “Prediction of Heart Disease Using Machine Learning.” 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA) (2018): 1275-1278.

[4] Aditi Gavhane, Gouthami Kokkula, Isha Panday, Prof. Kailash Devadkar, “Prediction of Heart Disease using Machine Learning”, Proceedings of the 2nd International conference on Electronics,Communication and Aerospace Technology(ICECA), 2018.

[5] Sharma, Vijeta & Yadav, Shrinkhala & Gupta, Manjari. (2020). Heart Disease Prediction using Machine Learning Techniques. 177-181. 10.1109/ICACCCN51052.2020.9362842.