WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Heart Disease Prediction Using Machine Learning Algorithms

Divya Pratap Singh[1], Vaibhav Pratap Singh[2], Sidhant Kumar Singh[3], Shivam Vatsa[4] Sanny Kumar[5]

[1,2,3,4] Student, Dept. Of Computer Science Engineering(Data Science), Noida Institute of Engineering and Technology, Greater Noida.
[5] Faculty, Dept. Of Computer Science Engineering(Data Science), Noida Institute of Engineering and Technology, Greater Noida.
Email: 011divyam@gmail.com

ABSTRACT –

The prevalence of heart disease is rising quickly every day, making early detection of heart disease frightening and crucial. Heart disease diagnosis is a difficult task that needs to be completed quickly and expertly. The purpose of this research is to identify patients who, in light of several medical characteristics, are more likely to suffer heart disease. Using the patient's medical information, we constructed a heart disease prediction model to determine the likelihood of receiving a heart disease diagnosis or not. We employed a wide range of machine learning methods, including Decision Tree Classifier, Gradient.

Boosting Classifier, Random Forest, K Neighbors Classifier, Support Vector Machine (SVM), Neural Network, Random Forest, and Logistic Regression, to forecast and determine which patient has heart illness. The model's ability to forecast cardiac illnesses more accurately in any individual was controlled using a very useful strategy. The suggested model's strength was quite pleasing; it combined Random Forest Classifier and Deep Learning to detect signs of heart disease in a specific individual, showing good accuracy when compared to other classifiers. The suggested heart disease prediction model will lower costs while improving medical care. Significant information from this study can be used to predict who will have heart disease. Python is used in the model's implementation, and the dataset was gathered from the Kaggle repository.

Keywords - Support Vector Machine(SVM), Random Forest, Deep Learning, Heart disease, Neural Network, Logistic Regression

Introduction

Machine learning (ML), a branch of artificial intelligence (AI), enables computers to autonomously learn from data and enhance their capabilities without being explicitly programmed. The core of machine learning involves developing models that independently analyze and learn from data, beginning their learning from observations, experiences, or directives to identify patterns and improve decision-making over time. The primary aim of these systems is to simulate human learning, consistently honing their abilities by interacting with actual data from the real world.

Heart diseases are a significant concern globally. The World Health Organization indicates that cardiovascular conditions are the foremost cause of death worldwide, taking approximately 17.9 million lives annually. This initiative utilizes medical history to pinpoint individuals at increased risk for heart disease, particularly those exhibiting indicators such as high blood pressure or chest pain. Early identification allows for more timely and effective treatment options that are less invasive.

Utilizing three data mining techniques—Random Forest Classifier, KNN (K-Nearest Neighbors), and Logistic Regression—this study achieves an 87.5% accuracy rate, surpassing previous models that employed only one method. By analyzing a dataset from the UCI repository containing various patient medical attributes, the study predicts heart conditions with KNN showing the highest accuracy at 88.52%. This method not only supports early risk detection but also helps to lower healthcare costs, showing substantial promise for improving patient outcomes.

Related Work

This work is motivated by extensive research into machine learning algorithms for identifying cardiovascular diseases, accompanied by a targeted literature review. A variety of methods such as the random forest classifier, KNN, and logistic regression have been utilized to predict cardiovascular conditions effectively. Results indicate that each algorithm possesses unique capabilities to meet specific objectives. In practice, both traditional and recent machine learning and deep learning models, including the IHDPS, were explored to determine decision boundaries and simplify key factors like family history of heart disease.

However, the predictive accuracy of the IHDPS model falls short compared to newer models that employ artificial neural networks and other advanced technologies to forecast coronary heart disease. McPherson [*et al*.] successfully utilized neural network techniques to pinpoint risk factors for atherosclerosis and coronary artery disease, providing reliable disease status predictions.

Furthermore, R. Subramanian and colleagues have advanced the use of neural networks in diagnosing and predicting conditions like blood pressure and heart disease. Developing a deep neural network that assimilates specific disease-related features is crucial for ensuring accurate predictions when applied to test datasets. This network, designed with approximately 120 hidden layers processed by an output perceptron, is recommended for diagnosing heart conditions under a supervised learning framework. Trained on historical data, this model's accuracy was validated with new data by a physician, demonstrating its effectiveness in clinical applications.

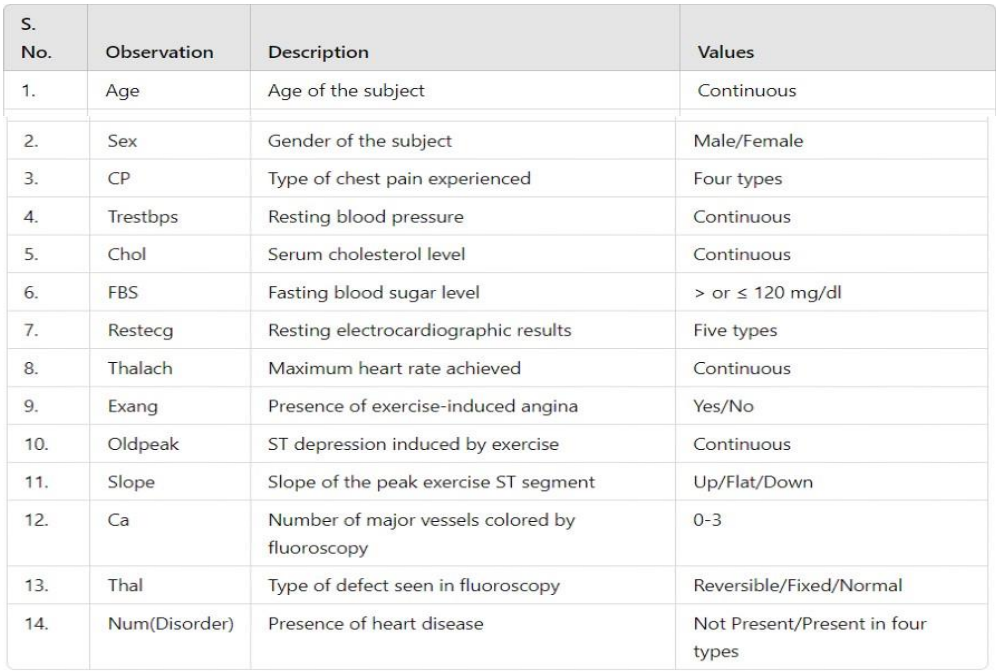
Data Source

A carefully organized dataset comprising 304 patients from various age groups was selected based on their medical history, including incidents of heart problems. Heart disease encompasses a variety of cardiac-related conditions which, according to the World Health Organization (WHO), are the predominant cause of death among middle-aged individuals. The dataset utilized in this study is rich in medical records and essential variables like age, resting blood pressure, fasting blood sugar level, and other relevant medical attributes. These data points are critical for determining whether a patient has been diagnosed with heart disease.

The dataset, sourced from the UCI repository, contains 13 critical medical attributes essential for assessing heart disease risk and differentiating between patients who are at risk and those who are not. This information allows for the classification of patients into groups according to their likelihood of developing heart disease, thereby enabling more focused and effective interventions. This heart disease dataset is crucial for identifying patterns that indicate the likelihood of heart conditions.

Structured into training and testing sections, this dataset contains 303 rows and 14 columns, with each row representing an individual patient's record. "Table 1" in the dataset provides a comprehensive list of all the attributes, serving as a reference for data analysis and pattern recognition in the study of heart disease risk.

Table 1. Various Attributes used are listed



S. No.	Observation	Description	Values
1.	Age	Age of the subject	Continuous
2.	Sex	Gender of the subject	Male/Female
3.	CP	Type of chest pain experienced	Four types
4.	Trestbps	Resting blood pressure	Continuous
5.	Chol	Serum cholesterol level	Continuous
6.	FBS	Fasting blood sugar level	> or ≤ 120 mg/dl
7.	Restecg	Resting electrocardiographic results	Five types
8.	Thalach	Maximum heart rate achieved	Continuous
9.	Exang	Presence of exercise-induced angina	Yes/No
10.	Oldpeak	ST depression induced by exercise	Continuous
11.	Slope	Slope of the peak exercise ST segment	Up/Flat/Down
12.	Ca	Number of major vessels colored by fluoroscopy	0-3
13.	Thal	Type of defect seen in fluoroscopy	Reversible/Fixed/Normal
14.	Num(Disorder)	Presence of heart disease	Not Present/Present in four types

Methodology

This research presents an analysis of several machine learning techniques, including K closest neighbors (KNN), logistic regression, and random forest classifiers. These algorithms can help medical analysts or practitioners identify heart disease accurately. Examining journals, papers that have been published, and data on cardiovascular illness from recent times are all part of this documentation. The suggested model's methodology provides a framework [13]. The process known as methodology consists of stages that convert provided data into identifiable data patterns for users' understanding. The suggested methodology (Figure 1) is broken down into steps: data collection is the first stage, significant values are extracted in the second stage, and preparation is the third stage where the data are examined. Data cleansing, normalization, and missing value handling are all covered by data preprocessing, depending on the algorithms employed [15]. The suggested model employs KNN, Random Forest Classifier, and Logistic Regression as the classifiers for the pre-processed data once it has been pre-processed. At last, we implemented the suggested model and assessed it using a range of performance criteria to determine its correctness and overall effectiveness. A heart disease prediction system (HDPS) that works well is shown in this was created with various classifiers. For prediction, this model makes use of 13 medical characteristics, including age, sex, blood pressure, cholesterol, chest pain, and fasting sugar [17].

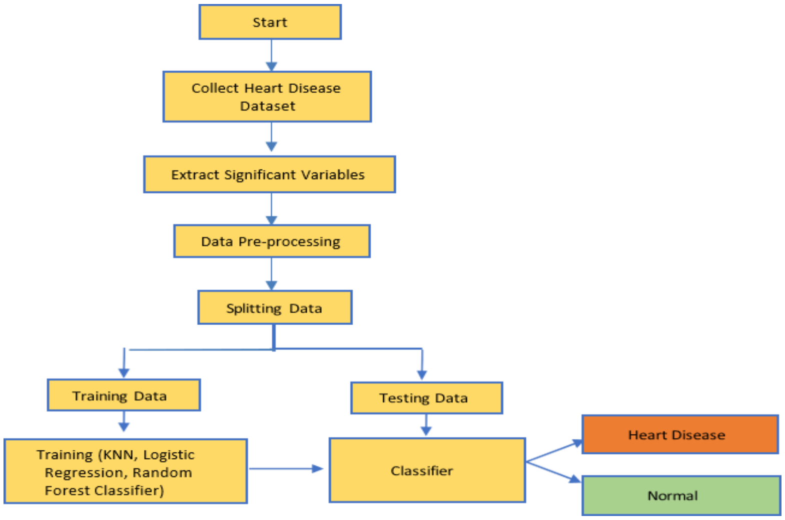

Start
Collect Heart Disease Dataset
Extract Significant Variables
Data Pre-processing
Splitting Data
Training Data
Testing Data
Training (KNN, Logistic Regression, Random Forest Classifier)
Classifier
Heart Disease
Normal

FIG 1. Proposed Model

Results and Discussions

From these findings, it is evident that while the majority of researchers employ various algorithms, including SVC and Decision Tree, to identify patients with heart disease, KNN, Random Forest Classifier, and Logistic Regression provide superior outcomes that surpass their methods [23]. Our employed algorithms outperform the algorithms of the previous researchers in terms of accuracy, cost-effectiveness, and speed. Furthermore, KNN and logistic regression yielded a maximum accuracy of 88.5%, which is higher than or nearly similar to the accuracies of earlier studies. Additionally, our research indicates that when it comes to predicting which patient would be diagnosed with heart disease, KNN and logistic regression perform better than random forest classifier. This demonstrates the superiority of KNN and logistic regression in the diagnosis of heart disease. Figures 2, 3, 4, and 5 depict a plot of the number of patients who have been classified and predicted by the classifier based on age group, resting blood pressure, sex, and chest pain.

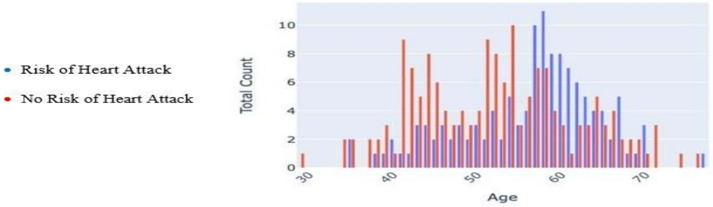


Fig 2. Shows the Risk of Heart Attack on the basis of their age.

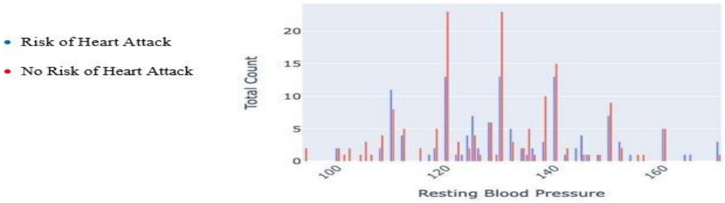


Fig 3. Shows the Risk of Heart Attack on the basis of their Resting Blood Pressure.

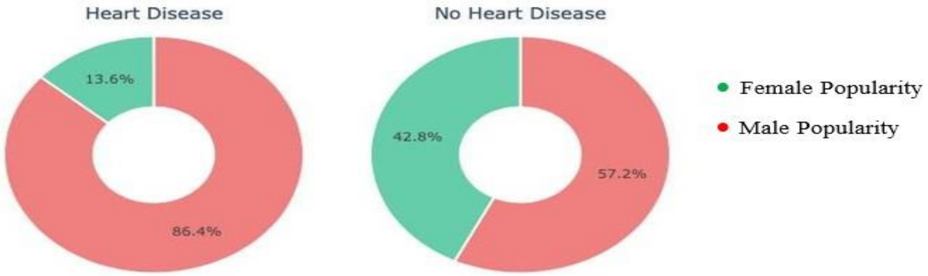


Fig 4. Shows the patients having or not having heart disease on thebasis of Sex.

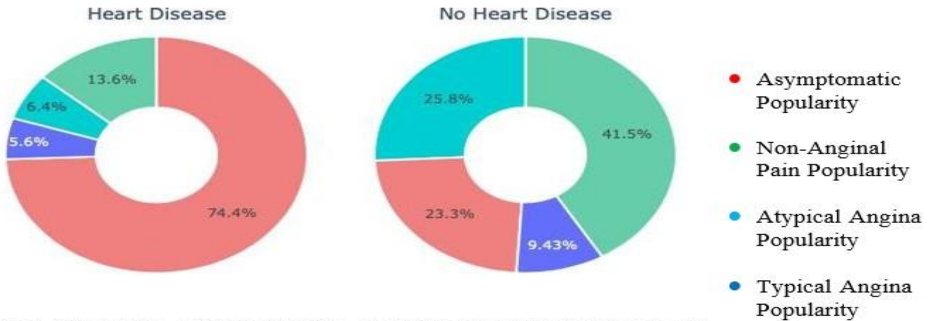


Fig 5. Shows the patients having or not having Heart Disease on thebasis of type of Chest Pain.

Conclusion

Three machine learning classification modeling techniques have been used to create a model for the detection of cardiovascular disease. By extracting the patient medical history that causes a deadly heart illness from a dataset containing the patient's medical history, including chest pain, blood pressure, sugar levels, and other conditions, this method predicts who will have cardiovascular disease. Based on the patient's clinical data, this heart disease detection system offers assistance if the patient has already received a heart disease diagnosis. The proposed model was constructed using the following algorithms: KNN, Random Forest Classifier, and Logistic Regression [22]. Our model has an accuracy of 87.5%. Increased training data ensures that the model has a better chance of correctly predicting whether or not a given individual has heart disease [9].

These computer-aided tools allow us to anticipate patients more accurately and quickly while also significantly lowering costs. We can work with a variety of medical databases since machine learning approaches are superior to human prediction, benefiting both patients and physicians. In light of this, we can conclude that this research helps us predict the patients who are diagnosed with heart problems by cleaning the dataset, using logistic regression and KNN, and producing an average accuracy of 87.5% on our model—better than the 85% accuracy of the prior models. It is also determined that, at 88.52%, KNN has the highest accuracy of the three algorithms we have used. "Figure 6" indicates that heart disease affects 44% of the individuals included in the study.

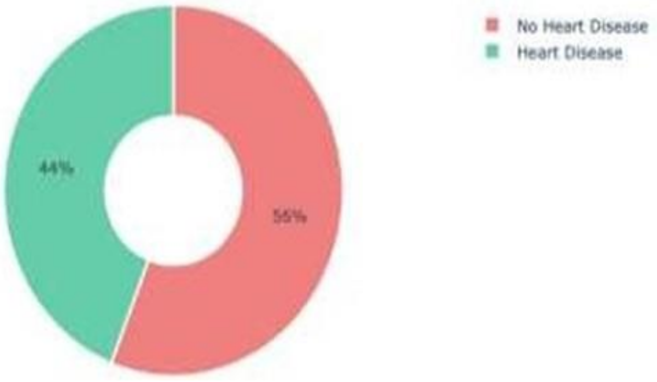


Figure 6. Shows the total number of patients having or not having Heart Disease.

REFERENCES

1. Soni J, Ansari U, Sharma D & Soni S (2011). *Predictive data mining for medical diagnosis: an overview of heart disease prediction.* International Journal of Computer Applications, *17*(8), 43-8

2. Dangare C S & Apte S S (2012). *Improved study of heart disease prediction system using data mining classification techniques*. *International Journal of Computer Applications*, *47*(10), 44-8.

3. Ordonez C (2006). *Association rule discovery with the train and test approach for heart disease prediction*. *IEEE Transactions on Information Technology in Biomedicine*, *10*(2), 334-43.

4. Shinde R, Arjun S, Patil P & Waghmare J (2015). *An intelligent heart disease prediction system using k-means clustering and Naïve Bayes algorithm*. *International Journal of Computer Science and Information Technologies*, *6*(1), 637-9.

5. Bashir S, Qamar U & Javed M Y (2014, November). *An ensemble-based decision support framework for intelligent heart disease diagnosis*. In *International Conference on Information Society (i-Society 2014)* (pp. 259-64). IEEE.
6. Jee S H, Jang Y, Oh D J, Oh B H, Lee S H, Park S W & Yun Y D (2014). *A coronary heart disease prediction model: the Korean Heart Study*. *BMJ open*, *4*(5), e005025.

7. Ganna A, Magnusson P K, Pedersen N L, de Faire U, Reilly M, Ärnlöv J & Ingelsson E (2013).*Multilocus genetic risk scores for coronary heart disease prediction*. *Arteriosclerosis, thrombosis, and vascular biology*, *33*(9), 2267-72.

8. Jabbar M A, Deekshatulu B L & Chandra P (2013, March). *Heart disease prediction using lazy associative classification*. In *2013 International Multi-Conference onAutomation, Computing,Communication, Control and Compressed Sensing (iMac4s)* (pp. 40- 6). IEEE.

9. Dangare Chaitrali S and Sulabha S Apte. *"Improved study of heart disease prediction system using data mining classification techniques." International Journal of Computer Applications* 47.10 (2012): 44-8.

10. Soni Jyoti. "Predictive data mining for medical diagnosis: *An overview of heart disease prediction*." *International Journal of Computer Applications* 17.8 (2011): 43-8.

11. Chen A H, Huang S Y, Hong P S, Cheng C H & Lin E J (2011, September). *HDPS: Heart disease prediction system*. In *2011 Computing in Cardiology* (pp. 557-60). IEEE.

12. Parthiban, Latha and R Subramanian. *"Intelligent heart disease prediction system using CANFIS and genetic algorithm." International Journal of Biological, Biomedical and Medical Sciences* 3.3 (2008).

13. Wolgast G, Ehrenborg C, Israelsson A, Helander J, Johansson E & Manefjord H(2016). *Wireless body area network for heart attack detection [Education Corner]. IEEE antennas and propagation magazine*, *58*(5), 84-92.

14. Patel S & Chauhan Y (2014). *Heart attack detection and medical attention using motion sensing device -kinect. International Journal of Scientific and Research Publications*, *4*(1), 1-4.

15. Zhang Y, Fogoros R, Thompson J, Kenknight B H, Pederson M J, Patangay A & Mazar S T (2011). *U.S. Patent No. 8,014,863*. Washington, DC: U.S. Patent and Trademark Office.

16. Raihan M, Mondal S, More A, Sagor M O F, Sikder G, Majumder M A & Ghosh K (2016, December). *Smartphone based ischemic heart disease (heart attack) risk prediction using clinical data and data mining approaches, a prototype design*. In *2016 19th International Conference on Computer and Information Technology (ICCIT)* (pp. 299-303). IEEE.

17. Buechler K F & McPherson P H (1999). *U.S. Patent No. 5,947,124*. Washington, DC: U.S. Patent and Trademark Office.

18. Takci H (2018). *Improvement of heart attack prediction by the feature selection methods. Turkish Journal of Electrical Engineering & Computer Sciences*, *26*(1), 1-10.

19. Worthen W J, Evans S M, Winter S C & Balding D (2002). *U.S. Patent No. 6,432,124*. Washington, DC: U.S. Patent and Trademark Office.

20. Acharya U R, Fujita H, Oh S L, Hagiwara Y, Tan J H & Adam M (2017). *Application of deep convolutional neural network for automated detection of myocardial infarction using ECG signals*. *Information Sciences*, 415, 190-8.

21. Brown N, Young T, Gray D, Skene A M & Hampton J R (1997). *Inpatient deaths from acute myocardial infarction*, 1982-92: analysis of data in the Nottingham heart attack register. *BMJ, 315*(7101), 159-64.

22. Piller L B, Davis B R, Cutler J A, Cushman W C, Wright J T, Williamson J D & Haywood L J (2002). *Validation of heart failure events in the Antihypertensive and Lipid Lowering Treatment to Prevent Heart Attack Trial (ALLHAT) participants assigned to doxazosin and chlorthalidone*. *Current controlled trials in cardiovascular medicine*, *3*(1), 10.

23. Folsom A R, Prineas R J, Kaye S A & Soler J T (1989). *Body fat distribution and self-reported prevalence of hypertension, heart attack, and other heart disease in older women*. *International journal of epidemiology*, *18*(2), 361-7.

24. Kiyasu J Y (1982). *U.S. Patent No. 4,338,396*. Washington, DC: U.S. Patent and Trademark Office.