






Paper Type: Original Article

## Heart Disease Prediction using Machine Learning Techniques

Raghad Ahmed Gad <sup>1</sup> , Salma Saad Abdelshakour <sup>1</sup>  and Ahmed A. Eldouh <sup>2,\*</sup> 

<sup>1</sup> Computer Science and Information System, October 6 University, Giza, Egypt.

Emails: [Raghadahmed213@gmail.com](mailto:Raghadahmed213@gmail.com); [salmasaad.mohammad@gmail.com](mailto:salmasaad.mohammad@gmail.com).

<sup>2</sup> College of Engineering Technology, Ashur University, Baghdad, Iraq; [ahmed.abdelrahim@au.edu.iq](mailto:ahmed.abdelrahim@au.edu.iq).

Received: 01 Sep 2024

Revised: 30 Oct 2024

Accepted: 21 Nov 2024

Published: 23 Nov 2024

### Abstract

Throughout their growth, artificial intelligence and machine learning have shown beneficial in multiple areas, particularly with the huge amount of data that has been generated recently. Making quicker and more accurate decisions regarding illness forecasts may be more dependable. The model can be used in the visualization and analysis of diseases. The article compares between different machine learning algorithms. In addition to numerous machine learning techniques, the UCI dataset is employed. Testing was done on Logistic Regression, Naïve Bayes, Support Vector Machine, Random Forest and Decision Tree algorithms. After performing the mentioned algorithm, it indicates that the decision tree performs better than another algorithm, its an accuracy of 98.54%.

**Keywords:** Machine Learning; Artificial Intelligence; Heart Disease; Support Vector Machine; Linear Regression; Decision Tree; Random Forest; Naïve Bayes.

## 1 | Introduction

Machine learning is a type of artificial intelligence that enables software programs to enhance prediction accuracy on their own without explicit programming. It predicts new outcome values by using past data as income [1-19]. Machine learning has become an essential competitive advantage for many firms due to its growing breadth and application. It includes ensemble learning and supervised, and unsupervised classifiers that use past data, also referred to as training data to generate predictions or judgments. Machine learning approaches are being examined in the medical industry to simulate human actions or mental processes and identify disorders based on many incomes. The expression of heart disease points to a collection of disorders affecting the human heart. Remarkably, with 17.9 million deaths worldwide, cardiovascular illnesses are currently the core cause of death. The application of several ML algorithms for the diagnosis of cardiac problems has been the subject of numerous studies. Used the University of California Irvine database to predict cardiac illnesses using machine learning techniques and discovered that these algorithms performed better. This article aims to build a predictive model to detect heart disease based on patients' medical histories, motivated by these investigations. This model seeks to forecast the existence of heart disease using patient medical records and attribute data from the UCI repository. For diagnosis, fourteen patient characteristics are considered, including age, sex, blood pressure, serum cholesterol, and 'exang'. Five ML algorithms: Logistic Regression, Naïve Bayes, Random Forest, Support Vector Machine (SVM), and Decision Tree are used for classification and prediction purposes. The study presents promising results. Particularly noteworthy Decision Tree boasts an impressive accuracy rate of approximately 98.5%. The main points of this topic are:



Corresponding Author: [ahmed.abdelrahim@au.edu.iq](mailto:ahmed.abdelrahim@au.edu.iq)



<https://doi.org/10.61356/j.mawa.2024.5430>



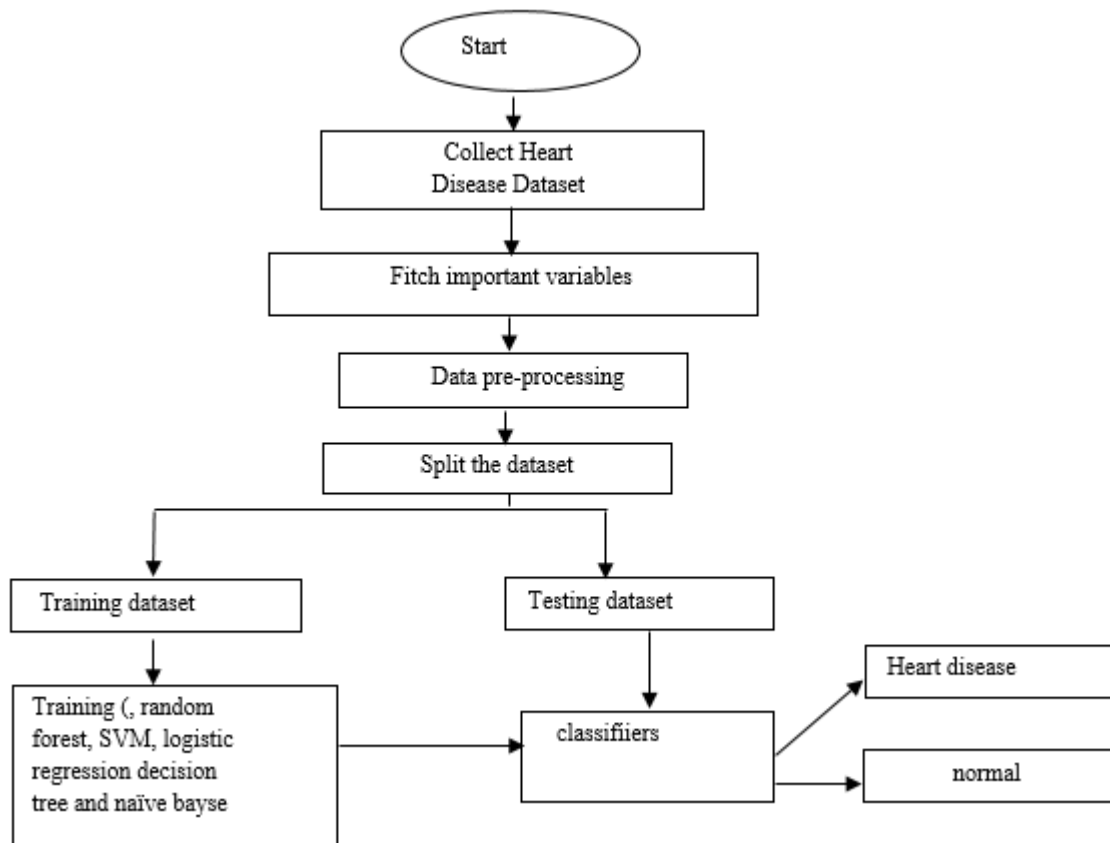
Licensee **Multicriteria Algorithms with Applications**. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0>).

- The Prediction of Heart Disease Using Machine Learning
- The use of a total of 1025 entries in one dataset with 14 different features.
- The testing of 5 machine learning algorithms. (Decision Tree, SVM, Naïve Bayes, Logistic Regression (LR), and Random Forest)

The paper is divided as follows: Sector two explains the methodology, Sector three explains various ML techniques, Sector four represents results and analysis, and Sector five contains the Conclusion.

## 2 | Methodology

This section outlines the methodology and analysis conducted in the current research as shown in Figure 1.



**Figure 1.** Process heart disease prediction.

The section explains the process of predicting heart diseases, starting from collecting the right dataset for this model and then extracting the essential features that affect the prediction process. After that, it covers the data preprocessing and splitting process, in which we handle the database and divide it into training and testing sets to employ various techniques on the trained dataset and see results by evaluating the accuracy using the testing data. This dataset was passed into five different ML techniques: Random Forest, Support Vector Machine, Naive Bayes (NB), Logistic Regression, and Decision Tree.

### 2.1 | Dataset and Attribute

Heart Disease Prediction Dataset from UCI

Dataset link: <https://www.kaggle.com/datasets/ketangangal/heart-disease-dataset-uci>

The database attributes describe the characteristics or properties of the dataset that play an essential role in analyzing and predicting outcomes of interest. The dataset consists of 14 features. The database was divided into training and testing partitions in an 80 to 20 percent ratio. The explanation of 14 features is discussed below:

Chest Pain (CP) refers to a sensation of discomfort or pain experienced in the chest area, often indicative of various medical conditions, most notably cardiovascular issues. Trestbps stands for "resting blood pressure" and is an important clinical parameter used to assess cardiovascular health, Thalach stands for the maximum heart rate achieved during an exercise stress test or physical activity. It is important, that Exang indicates exercise-induced angina, which is chest pain or discomfort experienced during physical exertion or exercise. Exang is a basic index used in the diagnosis and evaluation of heart disease risk. It measures as it provides insight into the cardiovascular fitness and capacity of an individual's heart, Oldpeak ST depression induced by exercise relative to rest. Not measured as an outcome, only gives information about cardiovascular fitness and the heart's reserve of a person's heart. The latter is a parameter used during stress testing to measure the severity of ischemia. It is a measure used during stress testing to assess the severity of coronary artery disease. Oldpeak values reflect the degree of myocardial ischemia or lack of oxygen supply to the heart muscle during physical activity. Ca refers to the number of major blood vessels (coronary arteries) that are visible using imaging techniques like fluoroscopy or angiography. That refers to thallium stress testing, which is a type of nuclear imaging used to check blood flow to the heart muscle. In this test, a radioactive substance called thallium is injected into the bloodstream, and images are taken to see how well blood is reaching the heart muscle both at rest and under stress.

## 2.2 | Pre-processing of Data

Data cleaning is a process of handling and removing noisy or missing values that negatively affect the dataset, ensuring accurate and trustworthy outcomes. This operation can be accomplished by using standard methods in Python to fill in the gaps left by noise or missing values. It is essential to alter the dataset, which can be done by using methods like aggregation, generalization, smoothing, and standardization. A crucial step in data preprocessing is integration, which entails resolving several problems to guarantee the smooth integration of different datasets. In certain instances, the dataset could be complex or challenging to understand, requiring a reduction to an appropriate format to achieve the best output. Managing unbalanced datasets is another factor to consider, which can be addressed through methods like undersampling and oversampling to handle the unbalanced data and have significant outcomes.

## 3 | Machine Learning Algorithms

### 3.1 | Logistic Regression

Logistic regression is a guided machine learning technique by labels that allows one to do binary classification tasks by predicting the probability of an outcome or an event. The model gives two possible outcomes: Yes→1 or no→0, in our case it tells us whether the patient has the disease or not.

### 3.2 | Support Vector Machine

SVM is an ML technique that is trained on labeled data, and it is utilized for both regression and classification. In n-dimensional space, the SVM method seeks to determine the best decision boundary, or line, for class separation. This boundary, also known as a hyperplane, makes it possible to quickly classify fresh data points. Support vectors are the vital vectors that help create the hyperplane and are identified by SVM.

### 3.3 | Naive Bayes

Naive Bayes is one of the techniques that classify new data based on the relation between inputs and outputs. It is used in classification tasks where the goal is to predict the category or class of given data based on the training data. In our case, the model starts to learn from the training data to see how often different features like age or cholesterol level will affect the outcome of whether a person has heart disease or not. Based on this information, when the model gets a new record from the testing set, the Naive Bayes algorithm checks the new patient's age, cholesterol level, and blood pressure, and sees how similar these features are to those of patients who had heart disease or did not.

### 3.4 | Decision Tree

It is a supervised learning algorithm. The main process is splitting the features into subsets looks like a tree structure; every node shows a feature, and every branch shows a decision rule. The decision tree learns how to make decisions or predictions based on this training data, and it then uses this information to predict the labels of new, unseen data.

### 3.5 | Random Forest

It is like the previous algorithm, but instead of combining branches, it combines multiple decision trees for the result. Each tree in the forest makes its prediction, and then the Random Forest algorithm combines these predictions to get the result.

## 4 | Experimental Results

### 4.1 | Result Comparing

It is like the previous algorithm, but instead of combining branches, it combines multiple decision trees for the result. Each tree in the forest makes its prediction, and then the Random Forest algorithm combines these predictions to get the result.

**Table 1.** Performance evaluation of reference paper.

Algorithm	Accuracy
Logistic regression	87.80%
SVM	67.80%
Naïve Bayes	82.93%
Decision Tree	98.54 %
Random Forest	92.54%

**Table 2.** Performance evaluation of our models and dataset.

Algorithm	First Dataset	Second Dataset	Third dataset
Logistic Regression	91.6%	90.8%	84.3%
Gradient Boosting	91.6%	90.8%	96.2%
Naïve Bayes	87.6%	84.7%	84.3%
Random forest	90.6%	90.3%	96.1%
Decision Tree	89.5%	86.5%	92.1%
K-NN	90.8%	89.8%	84.8%

So, we can see from previous tables that we improved Random Forest accuracy and Decision tree and then added a new technique, which is SVM, but it does not perform well.

### 4.2 | Analysis of Heart Disease Dataset

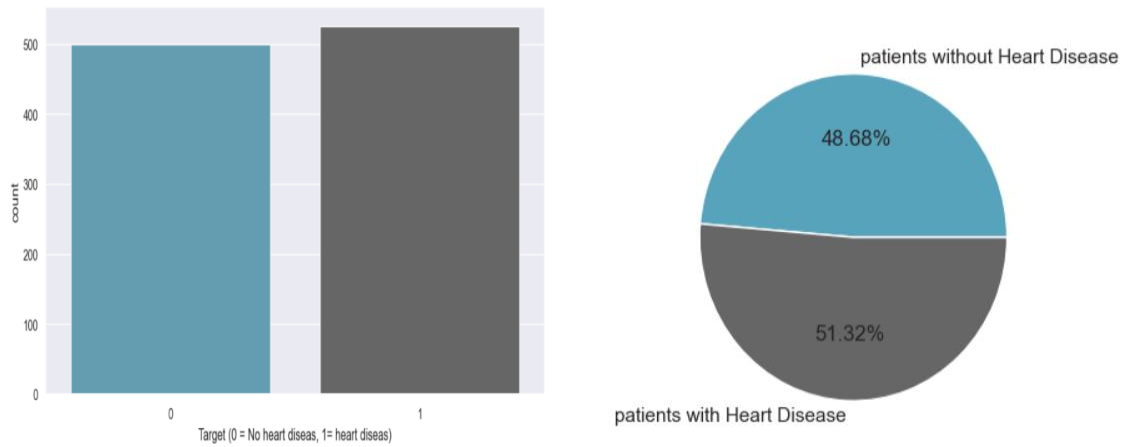


Figure 2. Target class.

The dataset has 1025 cases, among these 499 cases denoted 0 (No heart disease) a percentage is 48.68% while 526 cases denoted 1 (Have heart disease) a percentage is 51.32%.

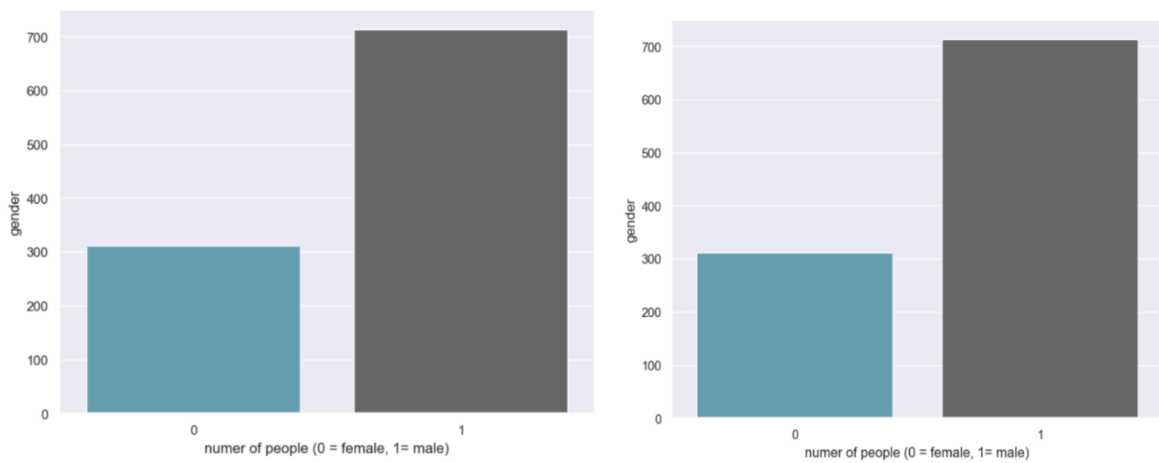


Figure 3. Count gender.

Figure 3 concludes that more males are more than females.

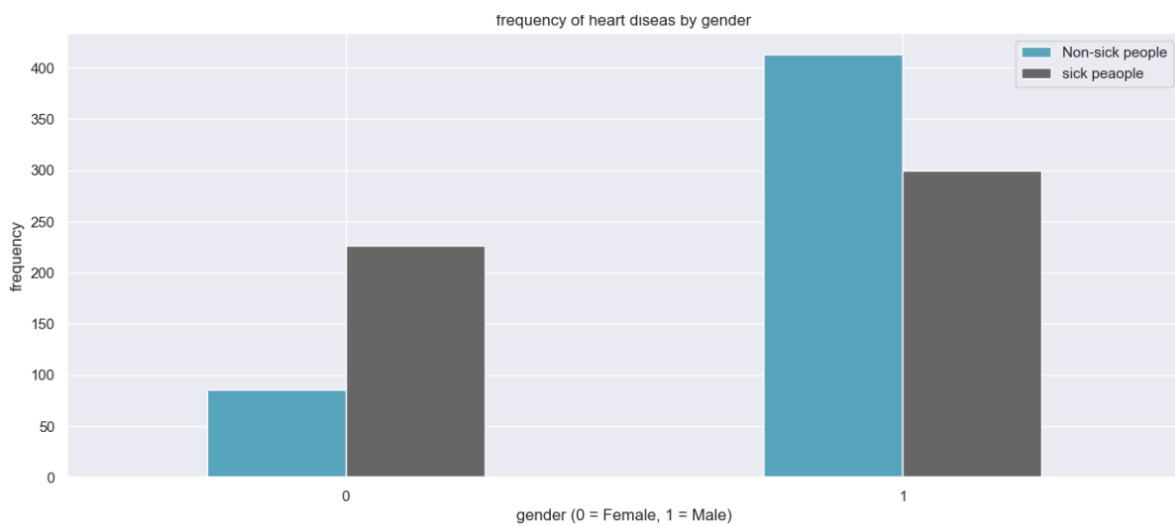


Figure 4. Frequency of heart disease by gender.

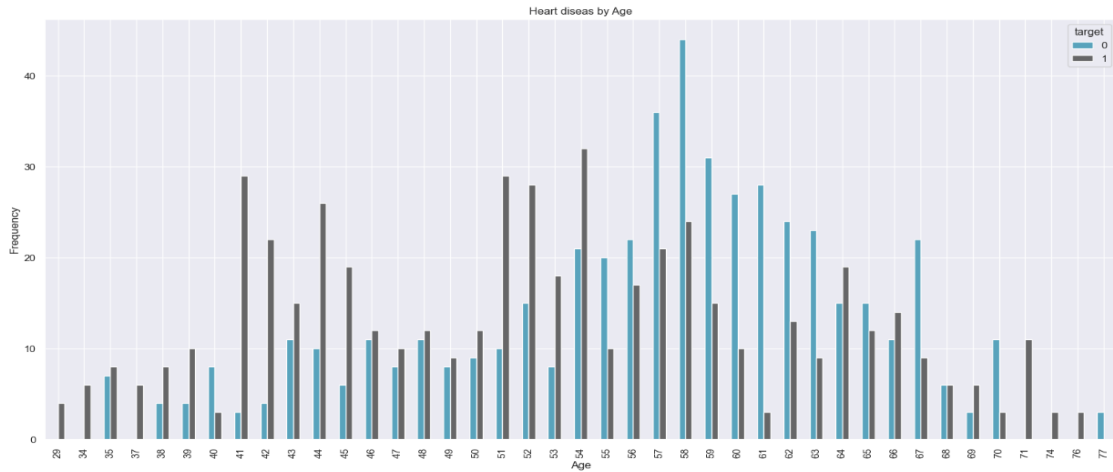


Figure 5. Comparison between Target and Age.

In Figure 4, the heart disease in females is less than in males

In Figure 5, the number of heart diseases at the age of 54 is the highest.

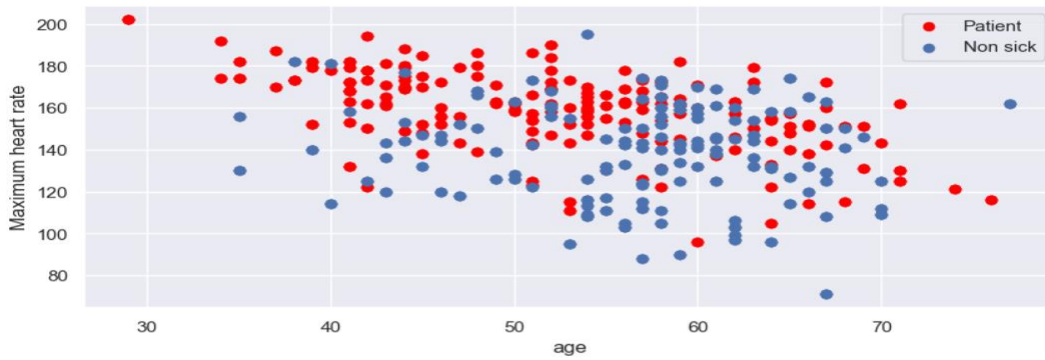


Figure 6. Disease Distribution between Maximum Heart Rate and Age.

In Figure 6, the number of heart diseases in ages between 50 and 60 with a heart rate between 140 and 180 is the highest.

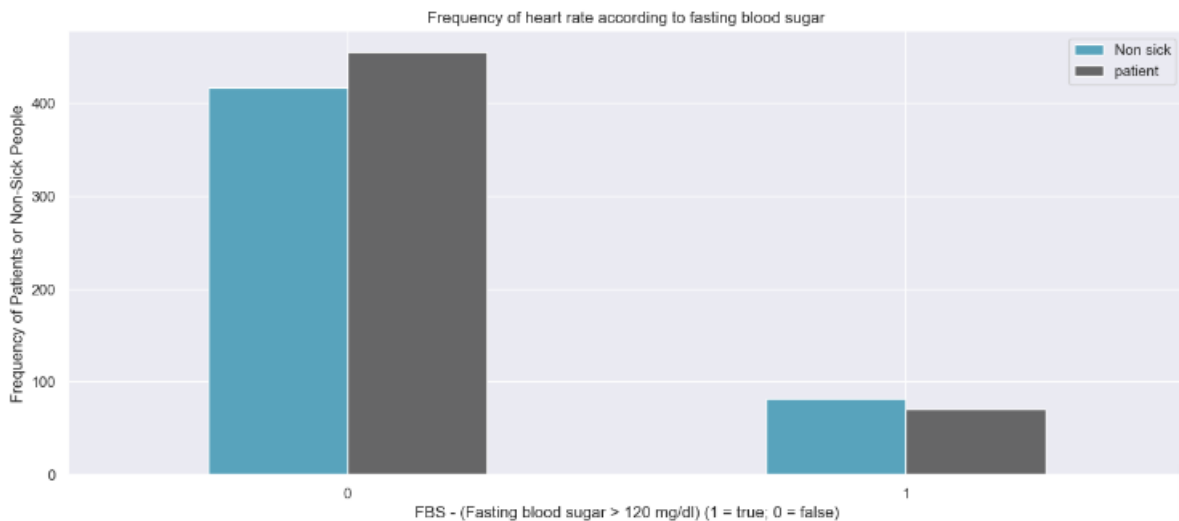
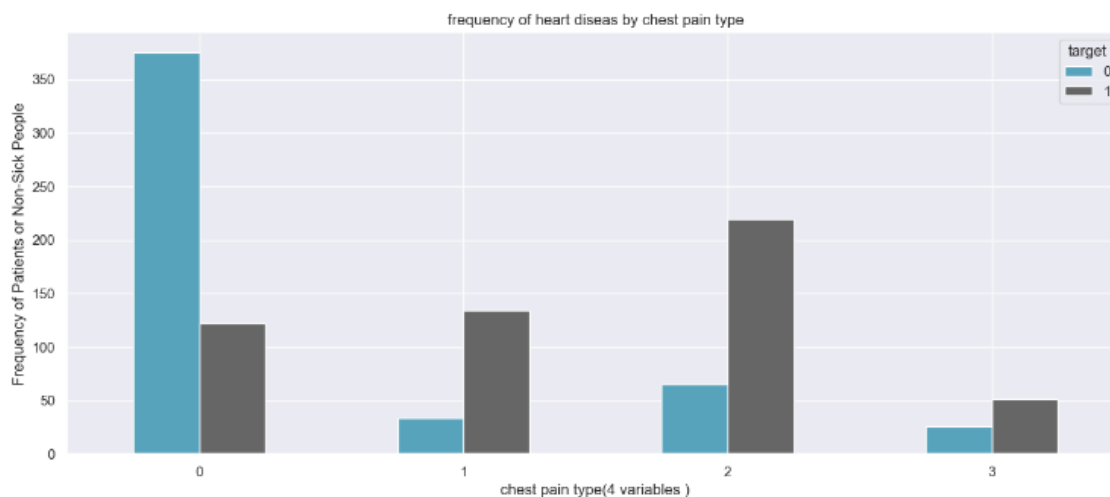


Figure 7. Frequency of Heart Disease According to Fasting Blood Sugar.



**Figure 8.** Frequency of Heart Disease by Chest Pain Type.

## 5 | Conclusion

It is essential to address heart diseases in the early stage; that will help us save as many lives as possible. Thanks to scientific progress, artificial intelligence, especially machine learning, helps us to define whether a person has heart disease or not. Decision Tree was the best one compared to the four other algorithms in predicting heart disease. With more research from medical professionals, the accuracy of prediction can grow even more. By using this model, we can deploy it on a website or application so it can be in a user-friendly interface. The doctor or even the user can check their case.

## Acknowledgments

The author is grateful to the editorial and reviewers, as well as the correspondent author, who offered assistance in the form of advice, assessment, and checking during the study period.

## Author Contributions

All authors contributed equally to this work.

## Funding

This research has no funding source.

## Data Availability

The datasets generated during and/or analyzed during the current study are not publicly available due to the privacy-preserving nature of the data but are available from the corresponding author upon reasonable request.

## Conflicts of Interest

The authors declare that there is no conflict of interest in the research.

## Ethical Approval

This article does not contain any studies with human participants or animals performed by any of the authors.

## References

- [1] M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, "Prediction of heart disease using random forest and feature subset selection," in *Innovations in bio-inspired computing and applications*. Springer, 2016, pp. 187–196.
- [2] P. Singh and I. S. Virk, "Heart Disease Prediction Using Machine Learning Techniques," 2023 Int. Conf. Artif. Intell. Smart Commun. AISC 2023, no. July, pp. 999–1005, 2023, doi: 10.1109/AISC56616.2023.10085584.
- [3] E. Fix and J. L. Hodges, "Discriminatory analysis. Nonparametric discrimination: Consistency properties," *International Statistical Review/Revue Internationale de Statistique*, vol. 57, no. 3, pp. 238–247, 1989.
- [4] S. A. Pattekari and A. Parveen, "Prediction system for heart disease using naïve bayes," *International Journal of Advanced Computer and Mathematical Sciences*, vol. 3, no. 3, pp. 290–294, 2012.
- [5] A. Singh and R. Kumar, "Heart disease prediction using machine learning algorithms," in 2020 International Conference on Electrical and Electronics Engineering (ICE3), 2020, pp. 452–457.
- [6] Y. Khourdifi and M. Bahaj, "Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization," *International Journal of Intelligent Engineering and Systems*, vol. 12, no. 1, pp. 242–252, 2019.
- [7] S. P. Shaji et al., "Prediction and diagnosis of heart disease patients using data mining technique," in 2019 international conference on communication and signal processing (ICCSP). IEEE, 2019, pp. 0848–0852.
- [8] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," in 2008 IEEE/ACS international conference on computer systems and applications. IEEE, 2008, pp. 108–115.
- [9] J.-K. Kim, J.-S. Lee, D.-K. Park, Y.-S. Lim, Y.-H. Lee, and E.-Y. Jung, "Adaptive mining prediction model for content recommendation to coronary heart disease patients," *Cluster computing*, vol. 17, no. 3, pp. 881–891, 2014.
- [10] H. Sharma and M. Rizvi, "Prediction of heart disease using machine learning algorithms: A survey," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 5, no. 8, pp. 99–104, 2017.
- [11] F. Przerwa, A. Kukowka, K. Kotrych, and I. Uzar, "Probiotics in prevention and treatment of cardiovascular diseases," *Herba Polonica*, vol. 67, no. 4, pp. 77–85, 2021.
- [12] A. N. Repaka, S. D. Ravikanti, and R. G. Franklin, "Design and implementing heart disease prediction using naïve Bayesian," in 2019 3rd International conference on trends in electronics and informatics (ICOEI). IEEE, 2019, pp. 292–297.
- [13] V. Ramalingam, A. Dandapath, and M. K. Raja, "Heart disease prediction using machine learning techniques: a survey," *International Journal of Engineering Technology*, vol. 7, no. 2.8, pp. 684–687, 2018.
- [14] M. Shouman, T. Turner, and R. Stocker, "Using decision tree for diagnosing heart disease patients," in *Proceedings of the Ninth Australasian Data Mining Conference-Volume 121*, 2011, pp. 23–30. Machine learning," *International Journal of Research and Technology*, vol. 9, no. 04, pp. 659–662, 2020.
- [15] H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, "Heart disease prediction using machine learning algorithms," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1022, no. 1, 2021, Doi: 10.1088/1757-899X/1022/1/012072.
- [16] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms," *Mobile Information Systems*, vol. 2018, 2018.
- [17] R. Katarya and S. K. Meena, "Machine learning techniques for heart disease prediction: a comparative study and analysis," *Health and Technology*, vol. 11, no. 1, pp. 87–97, 2021.
- [18] A. Jagtap, P. Malewadkar, O. Baswat, and H. Rambade, "Heart disease prediction using machine learning," *International Journal of Research in Engineering, Science, and Management*, vol. 2, no. 2, pp. 352–355, 2019.
- [19] Wikipedia contributors. (2022, June 22). Machine learning. In Wikipedia, the Free Encyclopedia. Retrieved 06:31, June 26, 2022, from [https://en.wikipedia.org/w/index.php?title=Machine\\_learning&oldid=1094363111](https://en.wikipedia.org/w/index.php?title=Machine_learning&oldid=1094363111).

**Disclaimer/Publisher's Note:** The perspectives, opinions, and data shared in all publications are the sole responsibility of the individual authors and contributors, and do not necessarily reflect the views of Sciences Force or the editorial team. Sciences Force and the editorial team disclaim any liability for potential harm to individuals or property resulting from the ideas, methods, instructions, or products referenced in the content.