SSRG International Journal of Computer Science and Engineering
ISSN: 2348–8387 / https://doi.org/10.14445/23488387/IJCSE-V10I4P101

Volume 10 Issue 4, 1-7, April 2023
© 2023 Seventh Sense Research Group®

Original Article

Estimating Heart Disease Used by Data Mining and Artificial Intelligence Techniques

R. Surendiran

School of Information Science, Annai College of Arts and Science, Kumbakonam, India

Corresponding Author : surendiranmca@gmail.com

Received: 24 February 2023 Revised: 02 April 2023 Accepted: 16 April 2023 Published: 28 April 2023

Abstract *- The heart is the most dangerous and significant part of the human physique. Life is entirely reliant upon the healthy operation of our hearts. It is a significant cause of death in the modern biosphere. One of the most critical health problems facing people today is heart illness. It is reportedly the leading cause of demise around the globe. Medical specialists frequently find it challenging to predict a cardiac illness early on. Many valuable hidden facts and information in the health sector today might be used to make predictions, particularly in treatment. Data mining remains a process for examining massive datasets before producing substantial and practical outcomes using exceptional AI-based apparatuses. This article aims to anticipate cardiovascular or heart illness using 3 AI devices Decision tree, naive Bayes, and a neural system. These techniques' determination is assessed based on many particular & factors with improvements for greater accuracy. The correctness based on various characteristics of each approach will then be compared. Afterwards, the most reliable method determines whether a man or woman resolves to develop coronary heart ailment. Medical professionals can utilize this method to anticipate diseases early so that prompt treatment can the given if he is persevering.*

Keywords *- Data mining, Artificial intelligence, Heart, Illness, Estimate.*

1. Introduction

Nowadays, one of the diseases with the most significant global prevalence is cardiovascular disease. According to estimates, it was responsible for 15% of all-natural fatalities in 2017—or around 17.9 million demises [15]. A circulatory disease is a long-lasting form of heart illness that may be recognized in its initial stages by taking measurements of several health indicators, together with blood pressure, fat, heart rate, and glucose levels [15]. The circulatory disease impacts not only individuals' health but also the nation's economies and costs [17]. Many machine learning and data mining methods are now being created and investigated to forecast various illnesses.

Similarly, several data mining, machine learning, and mixture procedures remain researched, created, and then examined to identify and forecast the initial stages of cardiac illness [27–32]. Identifying heart illness from a patient's data is known as a heart disease diagnosis. When a patient has many diseases, medics may not be talented at identifying them promptly [11] accurately. The authors of [21] examined various studies on the prognosis of cardiac illnesses from various years and concluded that data mining approaches are more accurate. Classification algorithms' ability to analyze massive data sets makes them commonly employed in the healthcare industry. Naive Bayesian, Sustenance Vector Mechanism, Adjacent National, Decision Tree, Uncertain Lucidity, Fuzzy Constructed Neural System, Non-natural Neural System, and Hereditary Procedures are approximately the most regularly used approaches in healthcare [1].

2. Related Work

Many strategies for heart disease prediction, including classification and feature selection techniques, have been scrutinized, established, and investigated by different academics and journalists. The authors combined the Linear Method (LM) and Random Forest traits to create the hybrid HRFLM technique (RF). They were 88.4% accurate in their predictions [1]. In a 2019 study, the authors aimed to improve prediction accuracy by utilizing various feature selection strategies. A UCI heart illness data set is subjected to individual applications of the following data mining methods: Decision Sapling, Logistic Reversion, Logistic Reversion SVM, Naive Bayes, and then Accidental Forest. The results are compared to those of previous studies. It is ultimately determined that Logistic Reversion, which achieved a correctness of 84.84%, is the most excellent eye assortment method for forecasting heart illness. [2]. Their test findings also revealed that the VOTE approach with NB and LR, which performed the best in data mining, had an accuracy rate of 87.4% when forecasting heart disease [3].

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

The models' presentation remained validated using the 10-crease cross-authentication method [3]. In instruction to enhance the detection of heart illness, the writers 2019 created an automatic analytic organization based on the two numerical models and DNN (2-DNN MODEL). Their suggested approach focused on the two primary issues—underfitting also overfitting—then presented an analytic organization that neither under fits nor overfits the working out information. Their suggested perfect produced a testing exactness of 93.33% [4].

The writers 2019 developed an automatic diagnostic method grounded on the two numerical models and DNN to improve the identification of heart disease (2-DNN MODEL). They provided an analytic organization that neither under fits nor overfits the working out information in their approach, concentrating on the two main problems of underfitting and overfitting. Testing accuracy with their proposed model was 93.33% [4]. d ANN performed better on the identical data set, with an accuracy of 78.14% [5]. In instruction to increase the correctness of heart disease organization, researchers created a hybrid strategy in 2019 that combines several strategies that take advantage of the Debauched Correlation-Based Feature Collection (FCBF) performance to reduce terminated signals. This approach demonstrated more significant than 90% accuracy [6]. Few writers [8] have made use of a group of classifiers. The collaborative algorithms bagging, improving, assembling, and majority elective were used during the research. The accuracy of the ensemble algorithms was enhanced by using the proper feature set selection methods. Using the feature set FS2[8], mainstream voting produced the highest accuracy.

The authors used the Star Log and Cleveland heart disease datasets in 2020 to predict coronary heart disease. In their balanced education, the profound neural system achieved better than attained correctness of 98.15% with the Star log dataset, while SVM obtained an accuracy of 97.36% with the Cleveland dataset [9]. According to the authors, this model's construction and control are less complex than other proposed replicas. The BP procedure was rummage-sale to train the network, along with a bipolar sigmoid purpose activation function and preprocessing techniques PCA then LDA to lessen the dataset's dimensionality. The characteristics closest to the hyperplane are designated using the SVM-LDA approach. The k crease authentication technique is used to authenticate the system. After 50 rounds, the network converges. By utilizing PCA, the suggested model attains 94.53% organization correctness for the judgment of heart illness [10]

The decision tree algorithm outperformed the other supervised machine learning (ML) algorithms tested by the authors in [13] concerning internal strength and accuracy. The consequences were attained by consuming two arithmetical software stages, R-Studio besides Rapid Miner, and remained associated with demonstrating this. The comparative analysis of the heart disease detection system was carried out by the authors of [18] utilizing the top 10 data removal classification procedures [33]. The accuracy of machine learning algorithms depends on the datasets used for training and testing, according to research done by the authors in [19]. They implemented a prediction algorithm and came to this result.

Support Vector Machine had the highest prediction accuracy, according to experiments conducted in 2011 by scientists in [23] using organization algorithms. A sophisticated and intellectual coronary heart disease prediction system (IHDPS) was suggested by Sellappan & Palaniappan in [24], employing three data removal approaches (naive Bayes, choice trees, and neural systems). Dramatists K. Srinivas, B. Kavita Rani, and then A. Govardhan discussed forecasting a heart attack using various data mining methods. They included Decision Tree, Naive Bayes, and ANN techniques [26]. Algorithms for statistical learning were created using data mining tools like TANAGRA.

3. Faces of Heart Disease

- For both men and women, heart sickness is the principal cause of humanity. In 2009, men were described for over half of all heart disease-related fatalities.
- Almost 610,000 Americans—or 1 in every four fatalities—die from heart disease yearly.
- Coronary heart disease, which claims the survival of more than 370,000 individuals annually, is the most prevalent kind of heart disease.
- Every 43 seconds, a heart attack occurs in the US. In America, a being dies from a heart illness-related incident every miniature.
- In the United States, heart disease—which affects people of most racial/ethnic groups, including Black Americans, Hispanics, and whites—is the foremost cause of mortality. Heart disease is the second leading basis of disease for American Indians, Pacific Islanders, and Asian Americans.
- The United States loses $108.9 billion annually to coronary heart disease alone. The price of prescription drugs, medical services, and missed wages are all included in this total.

4. Risk

Key heart disease hazard influences include smouldering, high blood pressure, and high LDL cholesterol. These three risk aspects are present in at least 50% of People. A person's risk for heart disease may also increase due to several additional medical issues and lifestyle decisions, such as:

1. Diabetes,
2. Obesity and being overweight
3. A meagre diet
4. A lack of exercise
5. Abuse of alcohol in excess

In this study, classification algorithms are used to examine heart disease predictions. With medical data, these hidden shapes might be demoralized for strength analysis. Data mining skill offers a practical method for identifying recent and enduring trends in the data. Healthcare administrators can utilize the detected information to obtain better services.

5. Proposed Organization

Based on the results of our literature analysis, they concluded that the three methodologies listed below are more accurate and effective in classifying and predicting. Thus, they tested out these three techniques:

1. Neural System
2. Decision Sapling
3. Adolescent Bayes

Figure 1 shows the attribute v/s accuracy 50 testing data graph, and th e Figure 2 shows the attribute v/s accuracy 50 testing data graph.

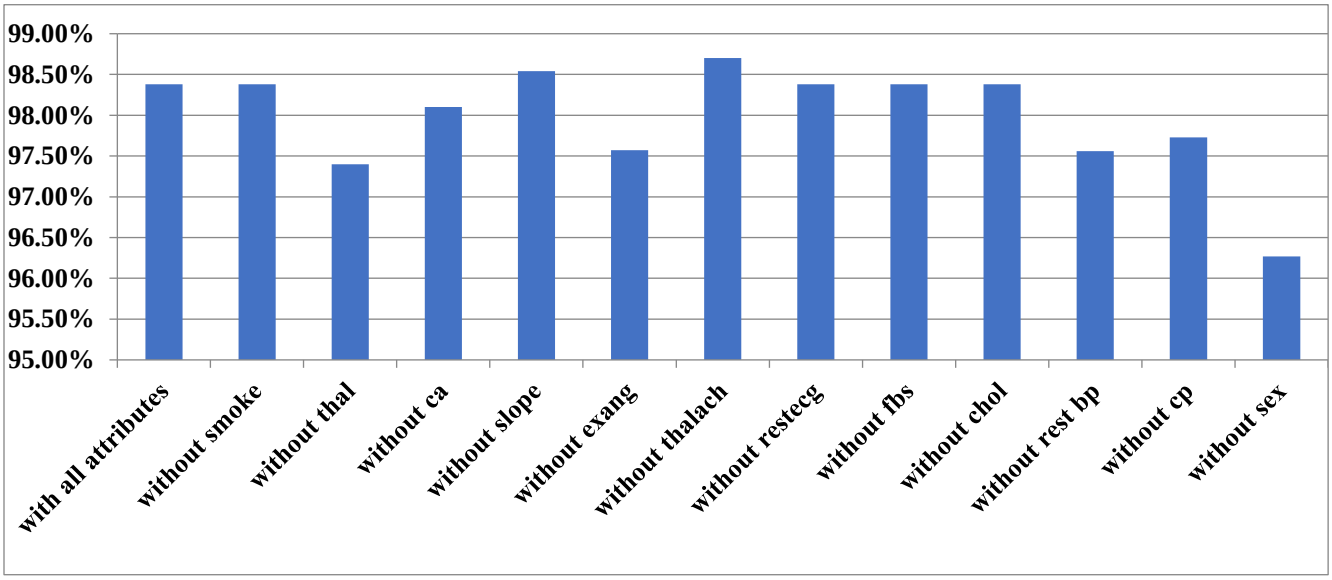

Fig. 1 Attribute v/s accuracy 50 testing data graph

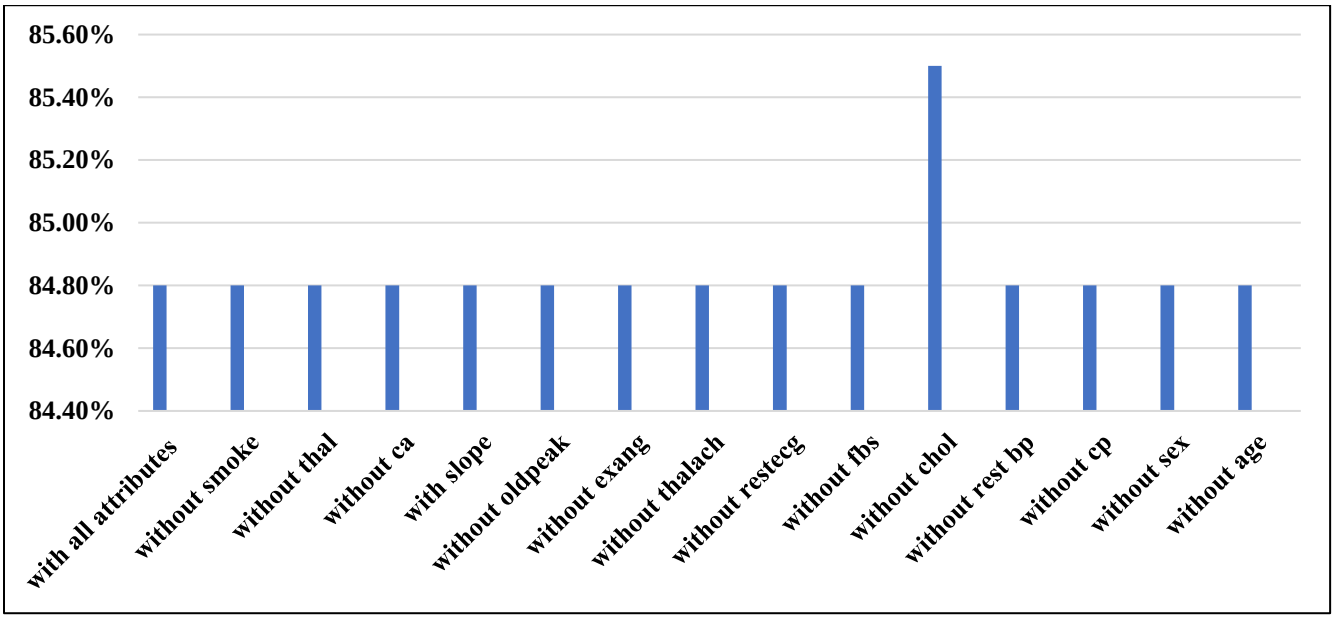

Fig. 2 Attribute v/s accuracy 200 testing data graph

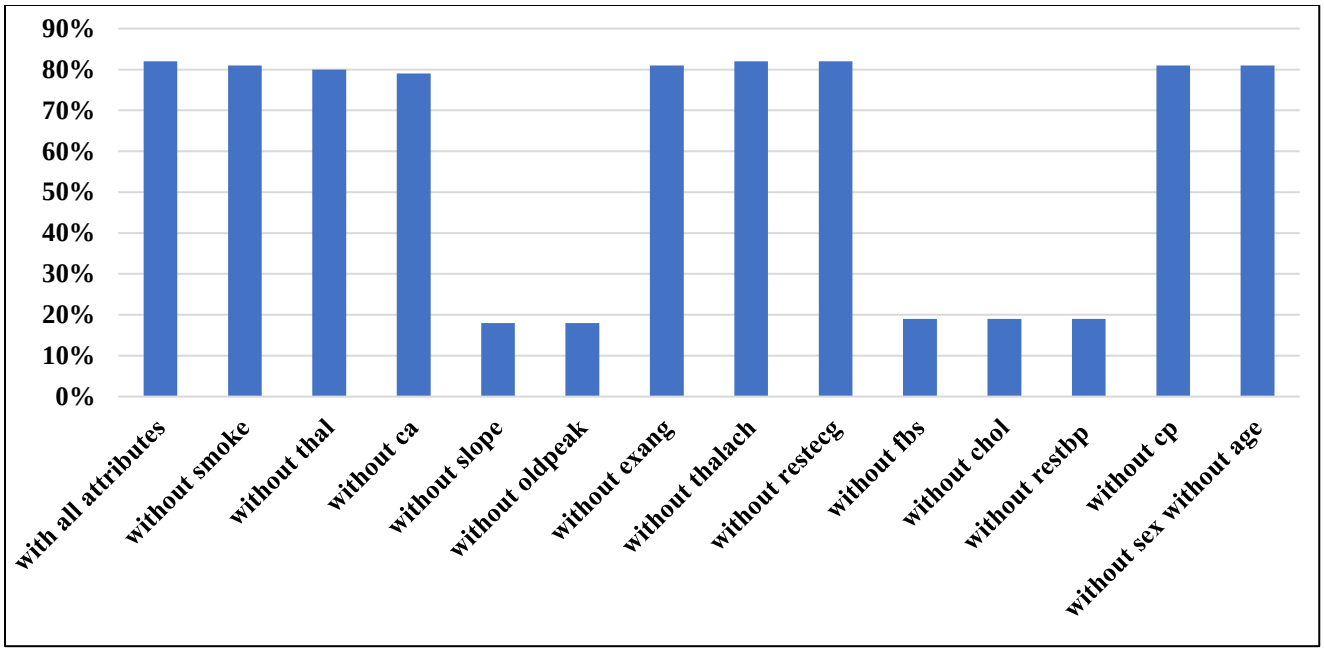

Fig. 3 Attribute v/s accuracy with hidden layer 5, epoch 500, folds 3, learning rate graph.

6. Implementation Grades and Discussion

Two administered learning algorithms, Nave Bayes and Conclusion Tree, have remained utilized in this study to envisage the presence of heart disease based on individual patient medical data. As only two classes are involved in the choice, it is a binary classification issue. This study utilized 14 variables in the Cleveland heart ailment data set. The qualities' specifics are as surveys are shown in Figure 3.

6.1. Characteristics

Contribution characteristics for the classification will be as surveys:

- Stage (in years) - a constant
- sexuality (categorical variable, 0: female, 1: male)
- The definite capricious called cp (Chest Pain Nature) - Its values range from 1 to 4, with each number designating a particular form of chest pain: value 1: the average angina value two: an unusual angina value Value of non-angina pain at position 3 4: asymptomatic
- The continuous variable rest by (Undeveloped Blood Gravity) in mmHg
- the continuous variable chol (Serum Cholesterol)
- abstaining blood sugar of more than 120 mg/dl (1= true, 0 = false) group element
- Consequences of the inactive electrocardiogram - Value 0: Average - Value 1: having an aberrant ST-T wave (T wave overturns, ST elevation or ST unhappiness of >0.05mV). Value 2 indicates the presence of leftward ventricular hypertrophy that meets Estes' criterion.
- thalach (Determined Heart Rate)
- exang (Exertional Encouraged Angina) 0: False 1: True category-specific adjustable
- gradient (Gradient of Peak Trial), with 1: Up, 2: Flat, and 3: Down distinct adjustable
- ca (Quantity of Foremost Pitchers): categorical variable, values 0, 1, 2, and 3.
- thal (standard, fixed, and reversible defects)
- Continual variable for the old peak (ST depression)
- The category variable num (Diagnosis of Heart Disease) is 0 or 1

7. Conclusion Tree

Figure 1 illustrates the Choice Tree evaluated used for various challenging data, demonstrating how the correctness may be increased by eliminating nearly all of the features and retesting. The findings for the three information removal algorithms used to estimate heart illness are displayed in numerous tables and diagrams after modifying the dataset, i.e., growing, then lowering the training and challenging information.

8. Decision Trees

Decision tree models are frequently employed in data mining to analyze the data and generate the tree and its governing rule. When examples are to be classified or placed into classes, the prophecy can be to forecast categorical standards (cataloguing trees). A decision node in a conclusion tree specifies that some evaluation be conducted on only a quality value, per one subdivision and sub-tree, for each potentially valid test result. Each node in a conclusion tree is also a foliage protuberance that illustrates the value of the mark characteristic or lesson of the samples.

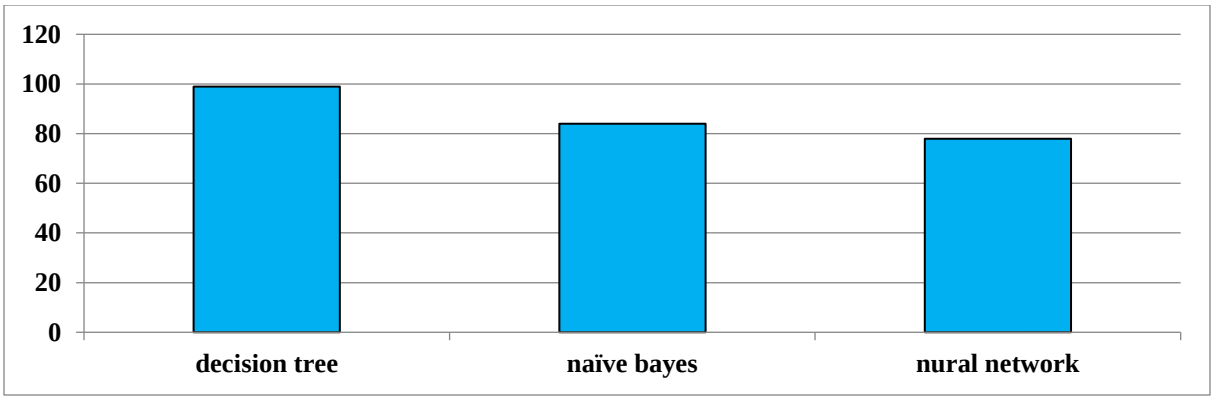

Fig. 4 Accuracy level for 3 data mining techniques

9. Naïve Bayes

The Bayes classifier is founded on the Bayes statement and makes independent predictions for each variable. A Naive Bayesian model benefits big datasets since it is simple to construct and does not require tedious, repetitive parameter estimates. The Naive Bayesian classifier works well despite its simplicity and is popular because it frequently outperforms more advanced classification techniques.

By subtracting P (c), P (x), and P (x | c), one may apply the Bayes proposition to arrive at the following possibility, P (c | x). The Naive Bayes classifier functions on the assumption that the effect of an analyst's (x) influence on a specific lesson's (c) impact is self-governing of the impacts of the extra predictors' (y) influences. Class-restricted independence is the term for this presupposition.

$$P(c\ /\ x) = \frac{P(x\ /\ c)P(c)}{P(x)}$$

$$P(c\ /\ x) = P(x_1|c) \times P(x_2|c)x \cdot \ldots xP(x_n|c) \times P(c)$$

The graph below demonstrates how the accuracy may be increased by eliminating some features and rerunning the Naive Bayes test for various quantities of testing data.

10. Neural Network

Relevant output data sets are connected to groups of contribution data sets using a multilayer perceptron architecture of an artificial neural network. This predictable rectilinear perceptron subtype employs three or more layers of nerve cells (nodes) for nonlinear stimulation. It performs better than a perceptron in this area because it can distinguish between data that is neither linearly discernible nor distinguishable by a hyper-plane.

11. Modal Assessment

The curve for the three data mining approaches' accuracy is shown in Figure 4.

12. Conclusion and Future Possibility

They may infer from the graphs mentioned above produced by our solution that as hidden layer numbers grow, the output gets less accurate and takes more time, making it less efficient. Also, accuracy fell after the learning rate. They found that using hidden layers, a faster learning rate and a more extensive training dataset resulted in the neural network's accuracy reaching 81.08%. The outcome changed as soon as they altered the characteristics. The accuracy of a decision tree was reduced when the qualities for cholesterol and chest discomfort were removed because both are crucial indicators of heart disease. However, when the sex feature was removed, the accuracy remained the same, leading us to believe that it had no bearing on illness prediction. The Nave Bayes method is independent of other qualities. Therefore they also tried to test its accuracy by deleting other attributes, but the findings did not change significantly.

To improve accuracy in a neural network, they tested with various hidden layers, knowledge rates, and Attribute changes. When they raised the numeral of secreted layers, the accuracy was better, but the calculation time increased, which was terrible for making predictions. However, once they decreased the number of secreted layers, the accuracy was better, and the calculation time decreased, making the results more dependable. After examining the graphs mentioned earlier, they concluded that the decision tree provided 98.53% more correct outcomes than the other tactics, which provided 85.02% (Naive Bayes) to 81.82%, respectively (neural system). In place of seen in the graph below, utilizing additional training datasets and analyzing the datasets may increase the system's effectiveness and reliability. By employing additional training datasets and then analyzing them, they can grow the system's efficiency and dependability. To be more specific, they can attempt to include more elements like junk food, exercise, and smoking. Likewise, there is room to improve this organization by fusing these methodologies to create a hybrid perfect that is more effective than using them separately.

References

[1] Senthilkumar Mohan et al., "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," *IEEE Access*, vol. 7, pp. 81542-81554, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[2] Saba Bashir et al., "Improving Heart Disease Prediction Using Feature Selection Approaches," *2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, pp. 619-623, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[3] Mohammad Shafenoor Amin, Yin Kia Chiam, and Kasturi Dewi Varathan, "Identification of Significant Features and Data Mining Techniques in Predicting Heart Disease," *Telematics and Informatics*, vol. 36, pp. 82-93, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[4] Liaqat Ali et al., "An Automated Diagnostic System for Heart Disease Prediction Based on x^2 Statistical Model and Optimally Configured Deep Neural Network," *IEEE Access*, vol. 7, pp. 34938-34945, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[5] Srabanti Maji, and Srishti Arora, "Decision Tree Algorithms for Prediction of Heart Disease," *Information and Communication Technology for Competitive Strategies*, pp. 447-454, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[6] Youness Khourdifi, and Mohamed 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, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[7] Neha Walia, and Arvind Kalia, "Programming Languages for Data Mining: A Review," *International Journal of Computer Trends and Technology*, vol. 68, no. 1, pp. 38-41, 2020. [CrossRef] [Publisher Link]

[8] C. Beulah Christalin Latha, and S. Carolin Jeeva, "Improving the Accuracy of Prediction of Heart Disease Risk Based on Ensemble Classification Techniques," *Informatics in Medicine Unlocked*, vol. 16, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[9] Safial Islam Ayon, Md. Milon Islam, and Md. Rahat Hossain, "Coronary Artery Heart Disease Prediction: A Comparative Study of Computational Intelligence Techniques," *IETE Journal of Research*, vol. 38, no. 4, pp. 2488-2507, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[10] Kavita Burse et al., "Various Preprocessing Methods for Neural Network-Based Heart Disease Prediction," *Smart Innovations in Communication and Computational Sciences,* vol. 851, pp. 55-65, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[11] Monther Tarawneh, and Ossama Embarak, "Hybrid Approach for Heart Disease Prediction Using Data Mining Techniques," *Advances in Internet, Data and Web Technologies*, pp. 447-454, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[12] S.Sreekanth, and Dr. P.C. Rao, "Anomaly Detection using Data Mining Methods", *International Journal of Computer Trends and Technology*, vol. 67, no. 12, pp. 20-23, 2019. [CrossRef] [Publisher Link]

[13] Juan-Jose Beunza et al., "Comparison of Machine Learning Algorithms for Clinical Event Prediction (Risk of Coronary Heart Disease)," J*ournal of Biomedical Informatics*, vol. 97, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[14] Amanda H. Gonsalves et al., "Prediction of Coronary Heart Disease Using Machine Learning: An Experimental Analysis," *Proceedings of the 2019 3rd International Conference on Deep Learning Technologies*, pp. 51-56, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[15] Sravani Nalluri et al., "Chronic Heart Disease Prediction using Data Mining Techniques," *Data Engineering and Communication Technology*, pp. 903-912, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[16] Dr.I.Lakshmi, and Y.Yogalakshmi, "A Study on Stress Management Strategies Factors Among Employees in It Sector Using Data Mining," *International Journal of Computer and Organization Trends*, vol. 9, no. 3, pp. 22-25, 2019. [CrossRef] [Publisher Link]

[17] Chandra Babu Gokulnath, and S. P. Shantharajah, "An Optimized Feature Selection Based on Genetic Approach and Support Vector Machine for Heart Disease," *Cluster Computing*, vol. 22, no. 6, pp. 14777-14787, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[18] I. Ketut Agung Enriko, "Comparative Study of Heart Disease Diagnosis using Top Ten Data Mining Classification Algorithms," *Proceedings of the 5th International Conference on Frontiers of Educational Technologies*, pp. 159-164, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[19] Archana Singh, and Rakesh Kumar, "Heart Disease Prediction using Machine Learning Algorithms," *2020 International Conference on Electrical and Electronics Engineering (ICE3),* pp. 452-57, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[20] Shekharesh Barik et al., "Heart Disease Prediction using Machine Learning Techniques," *Advances in Electrical Control and Signal Systems*, pp. 879-88, 2020. [CrossRef] [Publisher Link]

[21] Amogh Powar et al., "Data Mining & Artificial Intelligence Techniques for Prediction of Heart Disorders: A Survey," *2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (Vitecon),* pp. 1-7, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[22] Kumari, M., and Godara, S, "Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction," 2011.

[23] Wesley Odeh Odumu, and Ezekiel Endurance Chukwuemeke Igbonoba, "Big Data Mining Model to Predict Electronic Payment System using Machine Learning", *International Journal of Recent Engineering Science*, vol. 9, no. 2, Mar-Apr. 2022, pp. 7-17. [CrossRef] [Publisher Link]

[24] Chaitrali S. Dangare, and Sulabha S. Apte, "Improved Study of Heart Disease Prediction System Using Data Mining Classification Techniques," *International Journal of Computer Applications*, vol. 47, no. 10, pp. 44-48, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[25] Ashish Chhabbi et al., "Heart Disease Prediction Using Data Mining Techniques," *International Journal of Research in Advent Technology*. [Publisher Link]

[26] Dr. Hassabelrasul Yousuf AL Tom Shihabeldeen, "An Application in Islamic Financial Sector Using Data Mining," *SSRG International Journal of Economics and Management Studies*, vol. 6, no. 11, pp. 107-111, 2019.[CrossRef] [Publisher Link]

[27] Resul Das, Ibrahim Turkoglu, and Abdulkadir Sengur, "Effective Diagnosis of Heart Disease Through Neural Networks Ensembles," *Expert Systems with Applications: An International Journal*, vol. 36, no. 4, pp. 7675-80, 2009. [CrossRef] [Google Scholar] [Publisher Link]

[28] Resul Das, Ibrahim Turkoglu, and Abdulkadir Sengur, "Diagnosis of Valvular Heart Disease through Neural Networks Ensembles," *Computer Methods Programs Biomedicine*, vol. 93, pp. 185–191, 2009. [CrossRef] [Google Scholar] [Publisher Link]

[29] C. Gokulnath et al., "Preservation of Privacy in Data Mining by using PCA Based Perturbation Technique," *International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM)*, pp. 202-206, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[30] Revathi.S et al., "Increasing Crop Production Based on Data Mining Concepts," *SSRG International Journal of Computer Science and Engineering*, vol. 5, no. 10, pp. 3-6, 2018. [CrossRef] [Publisher Link]

[31] Ismail Babaoglu et al., "Assessment of Exercise Stress Testing with Arti- Ficial Neural Network in Determining Coronary Artery Disease and Predicting Lesion Localization," *Expert Systems with Applications,* vol. 36, no. 2, pp. 2562–2566, 2009. [CrossRef] [Google Scholar] [Publisher Link]

[32] K. Rajeswari, V. Vaithiyanathan, and T.R. Neelakantan, "Feature Selection in Ischemic Heart Disease Identification using Feed Forward Neural Networks," *Procedia Engineering*, vol. 41, pp. 1818–1823, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[33] Xindong Wu et al., "Top 10 Algorithms in Data Mining," *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1-37, 2008. [CrossRef] [Google Scholar] [Publisher Link]

[34] Ashitosh Tilve et al., "Pneumonia Detection using Deep Learning Approaches," *2020 International Conference on Emerging Trends in Information Technology and Engineering (IC-ETITE*), pp. 1-8, 2020. [CrossRef] [Google Scholar] [Publisher Link]