engineering proceedings

MDPI

Proceeding Paper

Heart Disease Prediction Using ML †

Abdul Rehman Ilyas [1,*], Sabeen Javaid [1] and Ivana Lucia Kharisma [2]

[1] Department of Software Engineering, University of Sialkot, Sialkot 51040, Pakistan; sabeen.javaid@uskt.edu.pk
[2] Department of Informatics Engineering, Nusa Putra University, Sukabumi 43152, West Java, Indonesia; ivana.lucia@nusaputra.ac.id
* Correspondence: abdulrehmanilyas8088@gmail.com
† Presented at the 7th International Global Conference Series on ICT Integration in Technical Education & Smart Society, Aizuwakamatsu City, Japan, 20–26 January 2025.

Abstract

The term heart disease refers to a wide range of conditions that impact the heart and blood vessels. It continues to be a major global cause of morbidity and mortality. The narrowing or blockage of blood vessels, which can result in major medical events like heart attacks, angina (chest pain) or strokes, is a common issue linked to heart disease. In order to lower the risk of serious complications and facilitate prompt medical intervention, early diagnosis and prediction are essential. This study developed predictive models that can precisely identify people at risk by applying a variety of machine learning algorithms to a structured dataset on heart disease. Blood pressure, cholesterol, age, gender, and other health-related indicators are among the 13 essential characteristics that make up the dataset. Numerous machine learning models such as Naïve Bayes, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree, Random Forest, and others were trained using these features. Using the RapidMiner platform, which offered a visual environment for data preprocessing, model training, and performance analysis, all models were created and assessed. The best-performing model was the Naïve Bayes classifier which achieved an impressive accuracy rate of 90% after extensive testing and comparison of performance metrics like accuracy precision and recall. This outcome shows how well the model can predict heart disease in actual clinical settings. By supporting individualized health recommendations, enabling early diagnosis, and facilitating timely treatment, the effective application of such models can significantly benefit patients and healthcare professionals. Furthermore, heart disease incidence can be considerably decreased by identifying and addressing modifiable risk factors such as high blood pressure, elevated cholesterol, smoking, diabetes, and physical inactivity. In summary, machine learning has the potential to improve the identification and treatment of heart-related disorders. This study highlights the value of data-driven methods in healthcare and indicates that incorporating predictive models into standard medical procedures may enhance patient outcomes, lower healthcare expenses, and improve public health administration.

Keywords: heart disease; machine learning; early diagnosis; predictive model; risk factors

Academic Editors: Debopriyo Roy, George F. Fragulis and Peter Ilic

Published: 10 October 2025

Citation: Ilyas, A.R.; Javaid, S.; Kharisma, I.L. Heart Disease Prediction Using ML. *Eng. Proc.* **2025**, *107*, 124. https://doi.org/10.3390/engproc2025107124

1. Introduction

One of the most important organs in the human body, the heart, pumps blood through a huge network of blood vessels to make sure that all of the cells and tissues receive oxygen and vital nutrients. An essential component of maintaining life and ensuring the healthy

operation of all body systems, the heart beats 100,000 times per day on average. The absence of a healthy and effective heart deprives the body of the oxygenated blood it requires, which can result in a number of health issues and, in extreme situations, death. Heart disease, sometimes referred to as cardiovascular disease (CVD), is now one of the world's leading causes of death. This chronic illness, which affects millions of people of all ages genders and socioeconomic backgrounds, is still on the rise at an alarming rate. By 2030, heart disease may kill about 23.6 million people a year if appropriate preventive measures are not taken, according to the World Health Organization (WHO). The most prevalent and dangerous type of heart disease is atherosclerotic heart disease, which is brought on by the accumulation of fatty deposits in the arteries. Unhealthy lifestyle choices, like smoking, poor eating habits, a lack of physical activity, chronic stress, and irregular medical checkups, are the main causes of heart disease. Heart attacks, strokes, and heart failure are made more likely by these risk factors, which progressively weaken the heart and blood vessels. Promoting heart health and delaying the onset of cardiovascular diseases requires avoiding bad habits, managing stress, eating a balanced diet, and exercising frequently. There is a pressing need for creative, effective, and scalable approaches to risk assessment and early detection, due to the rising incidence and severity of heart disease. Conventional diagnostic techniques are resource-intensive and frequently involve invasive procedures, despite their effectiveness. On the other hand, machine learning (ML) has become a potent instrument in the fields of healthcare and medical research, providing data-driven solutions that can accurately predict health outcomes and spot patterns. The use of machine learning techniques for heart disease prediction is investigated in this study using a publicly accessible dataset that includes 13 features such as age, gender, blood pressure, cholesterol, and other health-related indicators. Predictive models were constructed using a variety of supervised learning algorithms such as Random Forest, K-Nearest Neighbors (KNN), Naïve Bayes, Support Vector Machine (SVM), and Decision Tree. RapidMiner, a user-friendly platform popular for data science and machine learning applications, was used to implement and assess these models. The predictive performance of each algorithm was evaluated through testing and training. Since it showed the highest accuracy among the models tested, the Random Forest algorithm was chosen to serve as the foundation for the suggested predictive model. Measures like accuracy sensitivity and efficiency were used to assess the model's performance further in order to guarantee its dependability and robustness in practical applications. In order to help healthcare professionals diagnose heart conditions quickly and accurately, this study intends to contribute to the development of intelligent decision-support systems by utilizing machine learning for early heart disease prediction. The integration of these models into routine clinical practice and personal health monitoring systems could ultimately improve patient outcomes, lower mortality rates, and encourage the proactive management of heart health.

2. Literature Review

The authors of [1] describe how they used a dataset that included 13 features, such as medical history, to develop a machine learning model for predicting heart disease. They used KNN and SVM and the Decision Tree produced the highest accuracy of 98–83%. Real-time prediction was made possible by the model's deployment using Flask [2]. A dataset of 270 observations and 22 features is used in this paper [3] to predict heart disease. Neural networks, CN2 rule inducers, SVM, SGD, kNN, and Decision Trees were among the models used. SVM, SGD, and Decision Tree models produced the best accuracy of 87.69%. The authors of [4] use a dataset of 270 observations and 22 features to present machine learning models for the prediction of heart disease. They made use of neural networks, CN2 rule inducer, SVM,, SGD kNN, and Decision Trees. The algorithms of Decision

Tree, SVM, and SGD with 20-fold, 10-fold, and 5-fold cross-validation had the highest accuracy (87.69%). Using a combined dataset with 11 features from sources in Cleveland, Hungary, Switzerland, Statlog, and Long Beach, VA, the authors of [5] present machine learning models to predict heart disease. To balance and divide the data into training and testing, they employed the SVM-SMOTE model. When Random Forest, Logistic Regression, and Support Vector Machine were employed, Random Forest had the highest accuracy at 92.9%. SVM had the highest accuracy at 89.7%, while Logistic Regression reached 86.1%. A study [6] on machine learning methods for predicting heart disease was presented by the authors. The paper makes use of various medical datasets. ANN, Random Forest, SVM, KNN, Naïve Bayes, and Decision Trees were among the algorithms reviewed in the paper. A 88.7% accuracy was attained by a HRFLM model. In [7], the authors presented the importance of feature selection and model complexity for improving accuracy. The SVM algorithm was used in [8]. Despite the study's encouraging findings, the authors employed machine learning models such as Multilayer Perceptron (MLP), Random Forest, Support Vector Machine, J. 48 and Naive Bayes (NB) for the prediction of heart disease. With an accuracy of 94.9%, the Random Forest model was the most accurate. In order to predict heart disease, the authors of [9] used a Kaggle dataset with 59,000 rows and 11 attributes. Training accounted for 70% and testing for 30%. They made use of Random Forest, Multilayer Perceptrons, and Decision Trees. The accuracy of the models ranged from 86. 37% to 87.28%. Number of machine learning techniques were employed, such as neural networks, which produced an accuracy of 94.2%. Using data from four different medical facilities, the authors created an ensemble machine learning model to identify coronary heart disease and combined classifiers to increase accuracy using an adaptive boosting algorithm [10]. Following the selection of 29 features, the average accuracy across datasets was 85.27% and the precision was even higher at 85.84% [11]. Using machine learning methods, the authors of the paper [12] created a model for predicting heart disease. To increase accuracy, they employed a hybrid model known as HRFLM (Hybrid Random Forest with Linear Model) which combines Random Forest (RF) and Linear Model (LM) techniques. The accuracy of this hybrid approach was 88.7%. To forecast heart disease, the authors employed SVM, K-Nearest Neighbors, Logistic Regression, Gradient Boosting, and Extreme Gradient Boosting. The highest accuracy of 99.98% was attained by the Extreme Gradient Boosting model [13]. The highest accuracy of the authors' methods, which included Naïve Bayes and Decision Tree with SVM and FCMIM, was 92.37%in [14]. SVM (RBF) received 82%, ANN received 81%, and Logistic Regression received 88%. Patients in the dataset used by the author ranged in age from 29 to 79 years and it contained physiological characteristics, risk factors, and gender information, all of which are crucial for the diagnosis of heart disease [15]. SVM, KNN, Random Forest, Decision Tree, XGBoost, Naïve Bayes, and Logistic Regression were the algorithms used. The Random Forest model proved to be the most successful in predicting heart disease in the study with the highest accuracy of 86.89% [16]. The authors employed Logistic and other machine learning models, such as Random Forest, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Artificial Neural Networks, and Regression. A 96% accuracy was attained by the SVM algorithm [17]. With an overall average accuracy of 87.5%, the KNN algorithm obtained the highest accuracy of 88.52%. Outlier removal and dataset balancing were part of the preprocessing that led to the author of this paper [18] selecting 13 pertinent features from 79 attributes. XGBoost, Support Vector Machine (SVM), and Naïve Bayes were among the machine learning algorithms that were employed. At 87% accuracy, the XGBoost model is the most accurate. A prediction model was developed using the dataset. They employed algorithms such as Naive Bayes, KNN, SVM, Decision Trees, and Logistic Regression. The feature selection model produced accurate results for the prediction of heart disease [19].

3. Methodology

We have used Rapid Miner Studio for the methodology, which is a user-friendly software used to apply different machine learning algorithms to find the best accuracy. In Figure 1, the full process of how we apply different steps and achieve the results is explained.

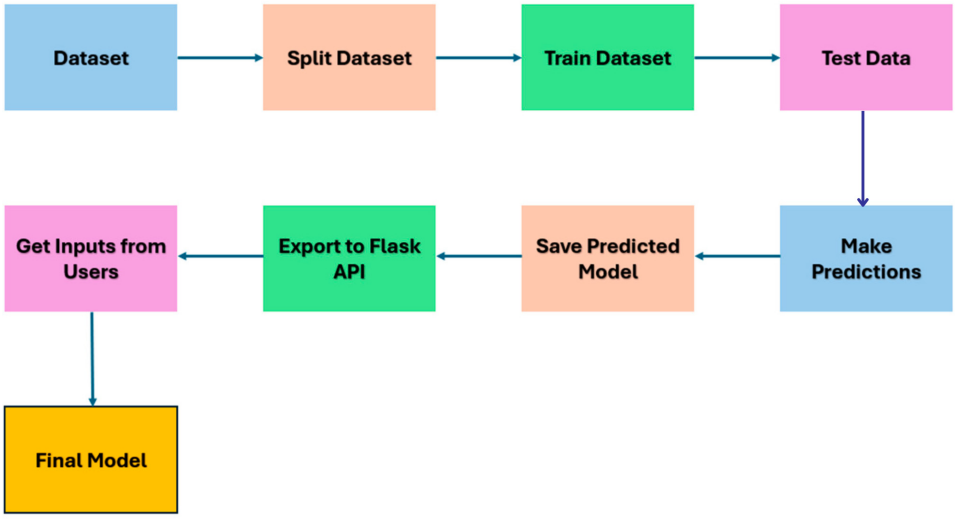

Figure 1. Flow diagram.

In this paper, the dataset used is a heart disease dataset which has been taken from kaggle.com. The dataset contains 14 attributes and 1025 values. The target label helps in predicting the presence of disease in patients with values of 0 = no disease and 1 = disease, as shown in Figure 2.

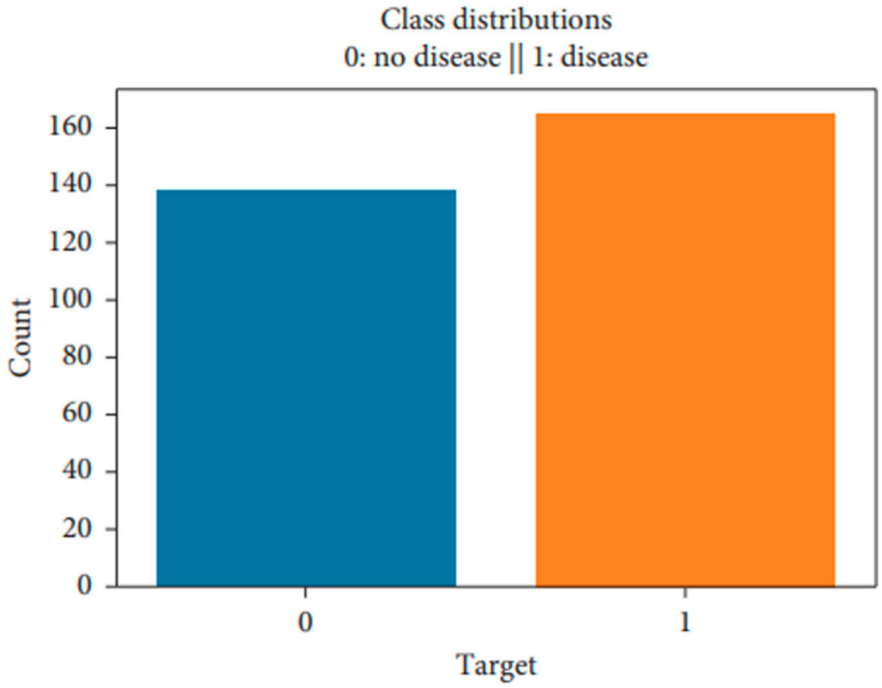

Figure 2. Predicting the presence of disease.

In past papers, many authors have achieved very good results. As detailed in the Methodology section, Rapid Miner was used for training and testing machine learning algorithms in Heart Failure Prediction using Naive Bayes, KNN, Random Forest, Linear

Regression, Logistic Regression, and Decision Tree, as these are very efficient in predicting the performance and accuracy of any dataset. In this process, we go beyond simply achieving high accuracy and delve deeper into evaluating the model's performance through various metrics. This research was conducted using a heart disease dataset by using an accuracy formula, Classification Error, Precision, Recall, and F1 Measure. The formulas for calculating accuracy and classification errors are as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\text{Classification Error} = 1 - \text{Accuracy}$$

3.1. Classification Algorithms

A supervised learning classification algorithm technique is used to forecast the results of the current data. As seen in Figure 3, the dataset has been split into training (0.7) and testing (0.3). The working of the individual classifiers is explained below.

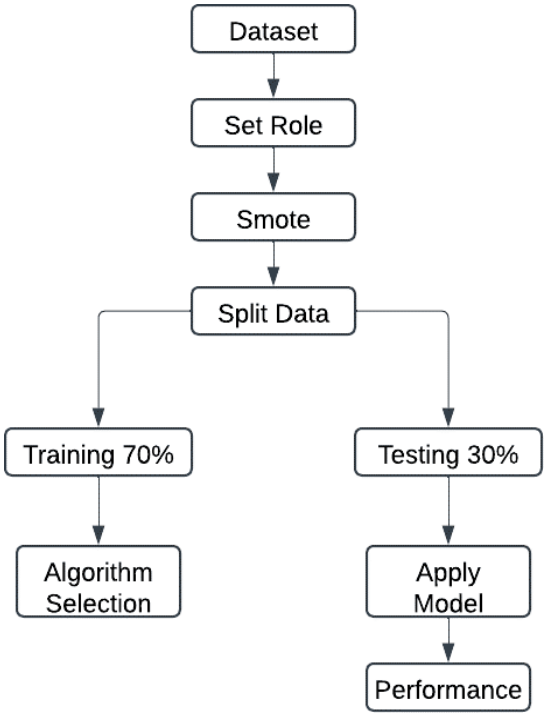

Figure 3. Framework.

3.2. Random Forest

The supervised machine learning algorithm is basically an ensemble method. Random Forest is a collection of different Decision Trees which use subsamples in order to achieve the maximum chances of accurate prediction in the model for predicting Heart Failure built into Rapid Miner. For our dataset, we carried out some hyper parameter tuning to help our model learn patterns without memorizing. The Random Forest model workflow is illustrated in Figure 4. The distribution of prediction values across ten bins is illustrated in Figure 5. As shown in Figure 6, the predictions are divided into two ranges based on the threshold of 0.5.

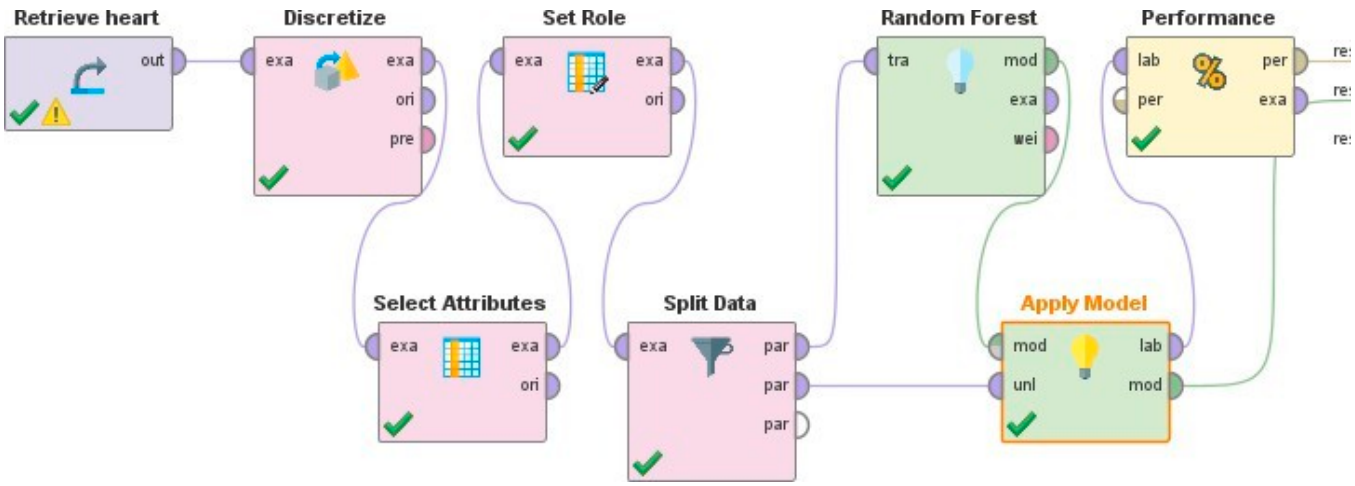

Figure 4. Random Forest model.

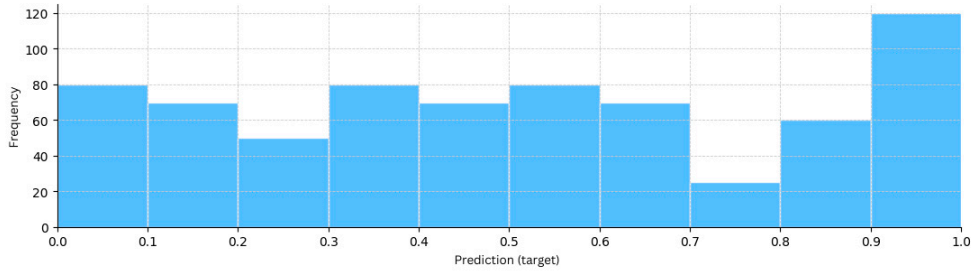

Figure 5. Histogram of prediction (target) values across ten bins.

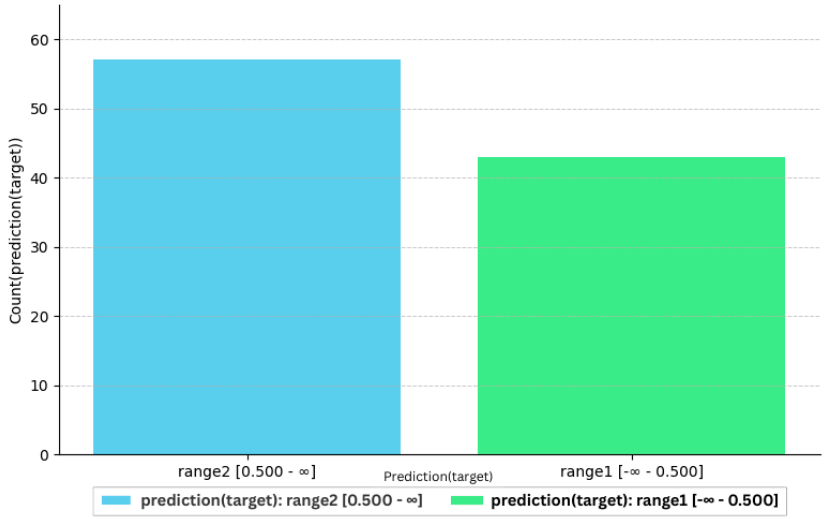

Figure 6. Distribution of predictions split into two ranges.

3.3. KNN

KNN predicts the unknown sample by considering its K closest Decision Tree neighbors in the training dataset based on similar characteristics.

3.4. Naïve Bayes

Naive Bayes is a classifier algorithm based on the probabilities of the given data points and past knowledge. A key limitation of Naive Bayes is that it assumes no connection between pieces of information given to it for prediction.

3.5. Decision Tree

In Decision Trees, prediction starts from the root towards the node and is very beneficial for handling multi-class datasets. Upon comparing the data point with a specific rule at a node, the Decision Tree pursues the branch relevant to that value and proceeds to the next node.

4. Results

This machine learning model was trained using Rapid Miner to check whether a person has heart disease or not by applying different algorithms. Now, we will discuss what results we obtain after applying different algorithms in Rapid Miner Studio. Every model gives a different accuracy. Table 1 shows the evaluation results of all algorithms.

Table 1. Results of algorithms.

Algorithm	Accuracy	Precision	Recall	F1 Score
Decision Tree	79.32%	80.57%	79.04%	79.04%
Random Forest	85.39%	85.43%	85.34%	85.39%
KNN	77.30%	77.27%	77.27%	77.27%
SVM	82.73%	84.56%	82.43%	83.48%
Logistic Regression	80.28%	80.82%	80.13%	80.47%
Linear Regression	82.60%	83.80%	83.75%	83.77%
Naïve Bayes	90.00%	93.51%	92.90%	93.20%
Ensemble	87.95%	88.20%	87.80%	88.00%

5. Conclusions and Future Research Directions

Disease diagnosis monitoring and treatment have all changed as a result of the increasing use of cutting-edge technologies in healthcare. The importance of artificial intelligence (AI) and machine learning (ML) in delivering intelligent healthcare solutions is growing as the world continues to transform into a digital ecosystem. This study helps bring about that change by using a variety of supervised learning algorithms to predict heart disease based on health-related attributes, including Random Forest, Decision Tree, Naïve Bayes, Logistic Regression, and K-Nearest Neighbors. With an impressive accuracy of 90%, the Naïve Bayes classifier outperformed the other algorithms, according to a comparative analysis. This shows how machine learning models can effectively identify people who are at risk of heart disease, which will ultimately help medical professionals make prompt and well-informed decisions. Early diagnosis and intervention can be made more accessible and dependable by integrating such models into healthcare systems, which will lower the prevalence of serious complications and deaths linked to cardiovascular diseases. However, even with these encouraging outcomes, this study also identifies areas that could use improvement in the future. A particular dataset with only 13 features served as the basis for the current investigation. Future research might concentrate on growing the dataset to include more thorough and varied patient records, including stress levels, genetic history, lifestyle factors, ECG readings, and real-time monitoring information from wearable technology. The model's learning capacity can be improved by increasing the amount and diversity of data, which will result in predictions that are even more accurate. Additionally, investigating ensemble approaches and deep learning strategies may improve model performance even more. Results from the use of hybrid models, which integrate the advantages of several algorithms, could be more reliable and broadly applicable. Furthermore, the development of real-time predictive systems that are integrated into healthcare applications or Internet of Things devices that continuously monitor patient vitals and provide real-time risk assessment could be the goal of future research. The

personalization of prediction models is another area that could be developed. More precise and contextually aware predictions can result from customizing algorithms according to each patient's unique profile, which includes demographics, medical history, and regional health trends. Furthermore, putting explainable AI (XAI) techniques into practice would help medical professionals trust and understand these systems better. In summary, there is great potential for improving early diagnosis, treatment planning, and lifesaving outcomes through the use of machine learning in heart disease prediction. Future studies should try to expand on these findings by utilizing richer datasets, more complex models, and real-world implementations as technology continues to transform healthcare. This brings us closer to a time when AI-powered solutions will be a common feature of preventative and individualized healthcare.

Author Contributions: A.R.I. conceptualized the study, designed the research framework, and supervised the project. S.J. carried out data collection, preprocessing, model implementation, evaluation of results, and prepared figures and tables. I.L.K. provided supervision and guidance, critically reviewed the methodology and results, validated findings, and contributed to manuscript editing and final approval. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data will be made available upon reasonable request to the first author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ahmad, G.N.; Fatima, H.; Ullah, S.; Saidi, A.S.; Imdadullah. Efficient Medical Diagnosis of Human Heart Diseases Using Machine Learning Techniques with and Without GridSearchCV. *IEEE Access* **2022**, *10*, 80151–80173. [CrossRef]
2. Airehrour, D.; Gutierrez, J.; Kumar Ray, S. GradeTrust: A Secure Trust Based Routing Protocol for MANETs. In Proceedings of the 25th International Telecommunication Networks and Applications Conference (ITNAC), Sydney, Australia, 18–20 November 2015; pp. 65–70. [CrossRef]
3. Anbuselvan, P. Heart Disease Prediction Using Machine Learning. *Int. J. Eng. Res. Technol.* **2020**, *9*, 515–518. [CrossRef]
4. Bharti, R.; Khamparia, A.; Shabaz, M.; Dhiman, G.; Pande, S.; Singh, P. Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning. *Math. Probl. Eng.* **2021**, *2021*, 8387680. [CrossRef] [PubMed]
5. Bhatt, C.M.; Patel, P.; Ghetia, T.; Mazzeo, P.L. Effective Heart Disease Prediction Using Machine Learning Techniques. *Algorithms* **2023**, *16*, 88. [CrossRef]
6. Chang, V.; Bhavani, V.R.; Xu, A.Q.; Hossain, M.A. An Artificial Intelligence Model for Heart Disease Detection Using Machine Learning Algorithms. *Healthc. Anal.* **2022**, *2*, 100016. [CrossRef]
7. Ghumbre, S.U.; Ghatol, A.A. Heart Disease Diagnosis Using Machine Learning Algorithm. In *Advances in Intelligent Systems and Computing, Proceedings of the International Conference on Information Systems Design and Intelligent Applications 2012 (INDIA 2012), Visakhapatnam, India, 5–7 January 20212*; Springer: Berlin/Heidelberg, Germany, 2012; Volume 132, pp. 217–225. [CrossRef]
8. Haq, A.U.; Li, J.P.; Memon, M.H.; Nazir, S.; Sun, R.; García-Magariño, I. A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms. *Math. Probl. Eng.* **2018**, *2018*, 3860146. [CrossRef]
9. Jindal, H.; Agrawal, S.; Khera, R.; Jain, R.; Nagrath, P. Heart Disease Prediction Using Machine Learning Algorithms. *IOP Conf. Ser. Mater. Sci. Eng.* **2021**, *1022*, 012072. [CrossRef]
10. Li, J.P.; Haq, A.U.; Din, S.U.; Khan, J.; Khan, A.; Saboor, A. Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare. *IEEE Access* **2020**, *8*, 107562–107582. [CrossRef]
11. Limbitote, M.; Damkondwar, K.; Mahajan, D.; Patil, P. A Survey on Prediction Techniques of Heart Disease Using Machine Learning. *Int. J. Eng. Res. Technol.* **2020**, *9*, 450–453. [CrossRef]
12. Mohan, S.; Thirumalai, C.; Srivastava, G. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. *IEEE Access* **2019**, *7*, 81542–81554. [CrossRef]

13. Nagavelli, U.; Samanta, D.; Chakraborty, P. Machine Learning Technology-Based Heart Disease Detection Models. *Math. Probl. Eng.* **2022**, *2022*, 7351061. [CrossRef] [PubMed]
14. Ahmad, A.A.; Polat, H. Prediction of Heart Disease Based on Machine Learning Using Jellyfish Optimization Algorithm. *Diagnostics* **2023**, *13*, 2392. [CrossRef] [PubMed]
15. Pires, I.M.; Marques, G.; Garcia, N.M.; Ponciano, V. Machine Learning for the Evaluation of the Presence of Heart Disease. *Procedia Comput. Sci.* **2020**, *177*, 432–437. [CrossRef]
16. Shukur, B.S.; Mijwil, M.M. Involving Machine Learning Techniques in Heart Disease Diagnosis: A Performance Analysis. *Int. J. Electr. Comp. Eng.* **2023**, *13*, 2177–2185. [CrossRef]
17. Taylor, O.E.; Ezekiel, P.S.; Okuchaba, F.B.D. A Model to Detect Heart Disease Using Machine Learning Algorithm. *Int. J. Comp. Sci. Eng.* **2019**, *7*, 1–5. [CrossRef]
18. Yilmaz, R.; Yagin, F.H. Early Detection of Coronary Heart Disease Based on Machine Learning Methods. *Int. Med. J.* **2022**, *4*, 1–6. [CrossRef]
19. Diwaker, C.; Tomar, P.; Solanki, A.; Nayyar, A.; Jhanjhi, N.Z.; Abdullah, A.; Supramaniam, M. A New Model for Predicting Component-Based Software Reliability Using Soft Computing. *IEEE Access* **2019**, *7*, 147191–147203. [CrossRef]