International Journal of applied exercise physiology

ISSN: 2322-3537
Vol-14 Issue-01 June 2025

An Analysis of Machine Learning Algorithms for Heart Disease Diagnosis

[1] S. Venkateswara Rao, [2] K. Akhila,

[1]Assistant Professor, Megha Institute of Engineering & Technology for Women, Ghatkesar.
[2] MCA Student, Megha Institute of Engineering & Technology for Women, Ghatkesar.

Abstract—

Heart disease is a major global health concern today, affecting a large percentage of the global population. It mostly affects the United States, but also India, and every year it claims a large number of lives there. Clinical studies and medical professionals agree that heart disease does not strike out of the blue but rather develops over time as a result of a person's chronically unhealthy lifestyle choices and the abnormalities in their bodies' biochemical processes. People seek medical attention at hospitals after experiencing these symptoms, and they undergo a battery of costly tests and treatments. This study's findings may help raise awareness about the patient's health before symptoms of the illness ever manifest. In this study, data was gathered from various sources and then divided into two sets: one set was used as a training dataset, while the other was used as a test dataset.

I summarized the results of my experiments with several classifier algorithms in an effort to improve their accuracy. These methods include k-nearest neighbor, logistic regression, decision tree classifier, support vector machine, random forest, and naive bayes. Similar to or even better than previous algorithms, SVM, Logistic Regression, and KNN produced accurate results. This research presents a new way to determine, with basic prefixes like sex, glucose, blood pressure, heart rate, etc., which factors are susceptible to heart disease. Applying various devices and conducting clinical trials for the actual experiment is the next step for this article. Machine learning, decision tree classifiers, random forest classifiers, support vector machines, logistic regression, k-nearest neighbors, gaussian NB, bernoulli NB, and big data are all terms that may be used.

INTRODUCTION

Following the heart, the brain is the most important organ in the human body. Its primary function is to pump blood throughout the body, making it an essential organ; yet, there are a number of situations in which it may not function properly. Although there are many forms of cardiovascular illness, the two most prevalent are heart failure and coronary artery disease (CAD). A narrowing or blocking of the coronary arteries is the primary cause of coronary artery disease [1]. With approximately 26 million individuals already affected and an estimated 2% yearly increase, coronary artery disease (CAD) is the top cause of mortality from cardiovascular disease. In 2005, there were 17.5 million fatalities worldwide [2]. Spending on CAD by the US government in 2008 amounts to $35 billion [3].
In terms of risk factors, medical research has identified two broad types: those that are immutable and those that are modifiable. Considerations such as sex, age, and family history are immutable. In contrast, modifiable factors include things like smoking, diet, exercise, BP, cholesterol, etc. Several methods exist for predicting the occurrence of cardiovascular disease, which is a major global health concern that requires accurate diagnosis. Although it might be a little pricey, angiograms are a preferred method. Consequently, a machine learning model may be used to forecast the patient's state in lieu of this conventional method. We have made great strides in the last year thanks to machine learning in our ability to forecast the state of heart disease for a medical dataset. Heart disease may be predicted using several common characteristics, such as:

- Age
- Gender
- Current Smoker
- Cigs per Day
- Diabetes
- Prevalent hypertensive
- Blood Pressure Medication

International Journal of applied exercise physiology

ISSN: 2322-3537
Vol-14 Issue-01 June 2025

• Systolic Blood Pressure
• Diastolic Blood Pressure
• Body Mass Index
• Total Cholesterol Level
• Heart Rate

This research proposes a classification machine learning system that can identify high-risk factors based on the aforementioned attributes and provide varying degrees of accuracy.

RELATED WORKS

By using several classification methods, this article is able to forecast the onset of cardiac illness in a patient. Many studies have used various methods to predict the occurrence of heart disease. Applying a logistic strategy yielded a 77.0% accuracy rate in early heart disease investigations [6]. The performance score was 78.9% in a comparable year, and another investigation used the CLASSIT idea cluster system [7]. University of California Irvine (UCI) heart disease dataset was the primary data source for the study and subsequent results [8]. K. Pramanik et al. suggest a hybrid algorithm that combines KNN and ID3 for the purpose of diagnosing cardiac illness. As a preprocessing algorithm, KNN is also useful for preprocessing data. The classifier uses the ID3 algorithm for heart disease prediction, and the treatment KNN method is used for perfect grade classification [9]. The authors S. Nashif, Mohammad Hasan Imam, Md. Rakib Raihan, and Md. Rasedul Islam [10] suggest using cloud computing and machine learning for the purpose of predicting the occurrence of cardiac diseases. With the help of SVM, the Arduino-based system was able to provide more accurate readings (e.g., blood pressure, heart rate, temperature) every ten seconds.

Architecture Diagram

III. PROPOSED METHODOLOGY

Data Collection Source

Data on cardiovascular illness is culled from Kaggle [4], which has 4238 rows and 16 columns. There is a direct or indirect relationship between heart illness and the distinctive features shared by all columns. Therefore, I made use of all qualities first, and then, after considering risk factors and predictions, I decreased their use to improve performance and eliminate significant risk factors. Aside from the other features, this dataset only contains one goal value. Each of those characteristics has a gender and an age associated with it, with a value of 0 for male and a value of 1 for female. Information such as their smoking history, blood pressure, glucose levels, heart rate, body mass index, and stock records from the past, among other things. There are records where the value is either 0 or 1, with 0 indicating no record and 1 indicating a record that is present. B. Modeling and Algorithms I gathered the dataset after first searching it from several sources. Once I had my data, I ran it through a number of feature engineering and statistical calculation operations. At that point, I had to deal with the junk and NULL values in that dataset. To improve the dataset's predictive power, I first build a new model using a variety of supervised classification machine learning methods, and then I add features and target values. As a last point, this model also has the potential to act independently of a system to execute the specified input values and produce the desired output; nevertheless, this remains an area of future investigation.

International Journal of applied exercise physiology

ISSN: 2322-3537
Vol-14 Issue-01 June 2025

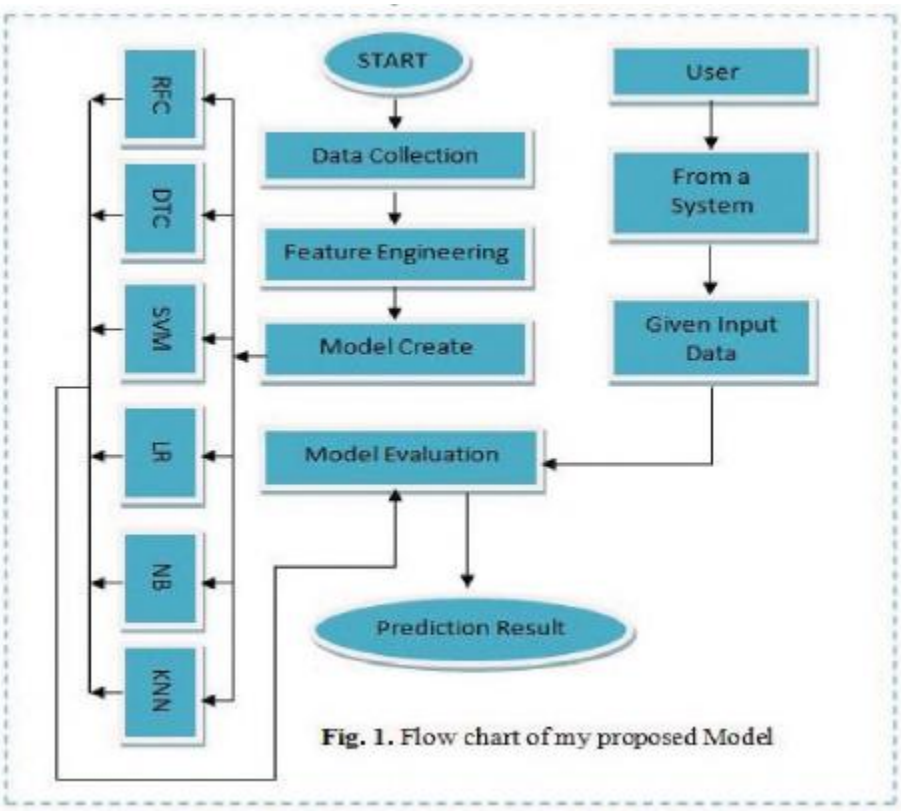

Fig. 1. Flow chart of my proposed Model

IV. RESULT & DISCUSSION

One method for evaluating the efficacy of machine learning classification algorithms is the confusion matrix. The confusion matrix allows us to compute F1, recall, accuracy, and precision.

CONFUSION MATRIX

		Actual values	
		Positive	Negative
Predicted Values	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Precision: Preciseness equals (True Positive + True Negative) divided by Total I discovered that KNN, SVM, and LR achieved the greatest and equal accuracy (87%) out of seven machine learning classification methods. In contrast, RFC and BNB both forecast 86%, which is comparable to or slightly higher than DTC's 76%. All GNB is precisely 81% at that moment. In conclusion, the performance of the first three algorithms is superior than that of other algorithms, however DTC's performance is the worst.

International Journal of applied exercise physiology

ISSN: 2322-3537
Vol-14 Issue-01 June 2025

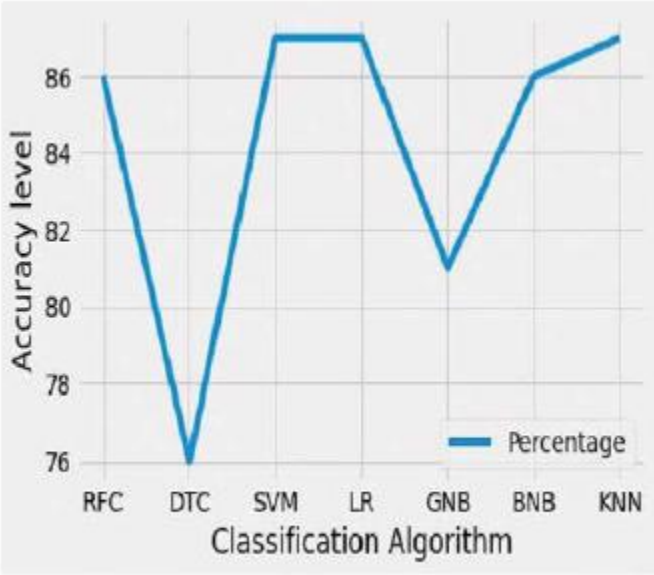

Figure 2. Accuracy of various techniques.

TABLE I. PRECISION, RECALL, AND F1 OF DIFFERENT TECHNIQUES.

Techniques	Precision 0	Precision 1	Recall 0	Recall 1	F1 Score 0	F1 score 1
RFC	.88	.39	.99	.06	.93	.11
DTC	.89	.21	.83	.30	.86	.26
SVM	.87	00	1.0	00	.93	00
LR	.87	.50	1.0	.03	.93	.05
GNB	.89	.28	.90	.25	.90	.26
BNB	.87	.20	.99	.01	.93	.02
KNN	.87	00	1.0	00	.93	00

B. Precision:

The ratio of the number of positive instances that were accurately anticipated to the total number of positive cases is known as precision. The accuracy of each algorithm is shown below this bar graph [5].

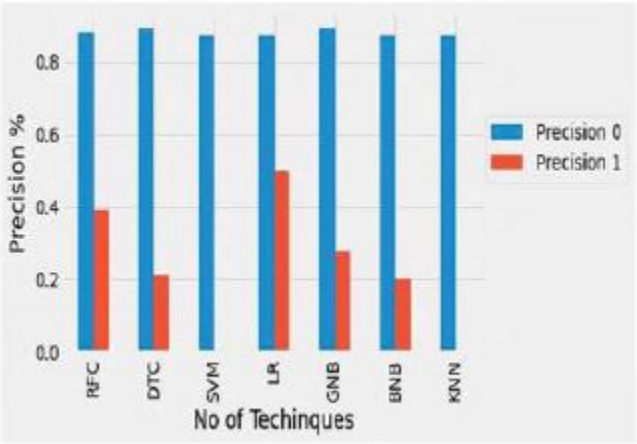

Figure 3. Precision of various techniques.

C. Recall:

International Journal of applied exercise physiology

ISSN: 2322-3537
Vol-14 Issue-01 June 2025

Recall is the fraction of true positives out of the total number of examples in the training set that were accurately predicted. Calculate the recall by dividing the sum of the true positives and false negatives by the total number of TRs. The formula for precision is HP divided by the sum of TP and FP.

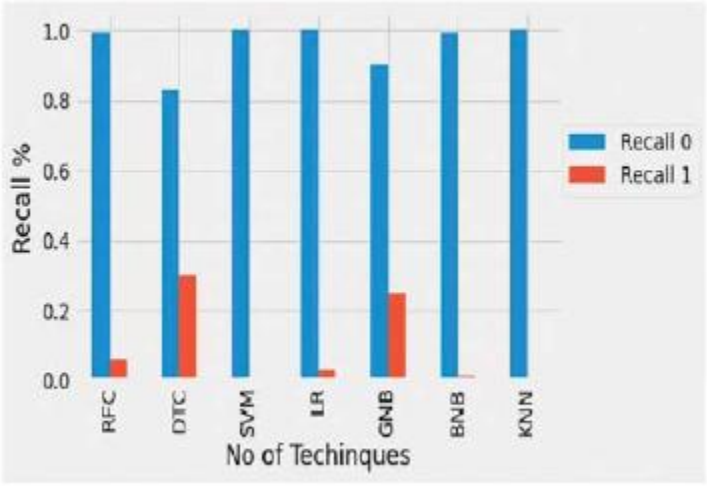

Figure 4. Recall of various techniques.

D. F1 Measure:

The F1 measure is the mean of the recall and accuracy. Consequently, this metric was used for both positive and negative results. In terms of practicality, F1 outshines accuracy. F1 is appropriate in very varied situations, although accuracy is superior when the false positive and false negative costs are close. Your F1 score is equal to the product of two times the precision and recall, divided by the sum of the two.

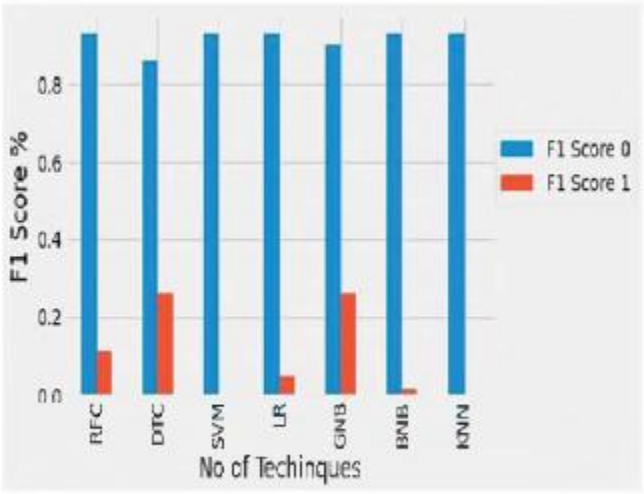

Figure 5. F1-Score of various techniques.

V. ANALYZING RISK FACTOR

International Journal of applied exercise physiology

ISSN: 2322-3537
Vol-14 Issue-01 June 2025

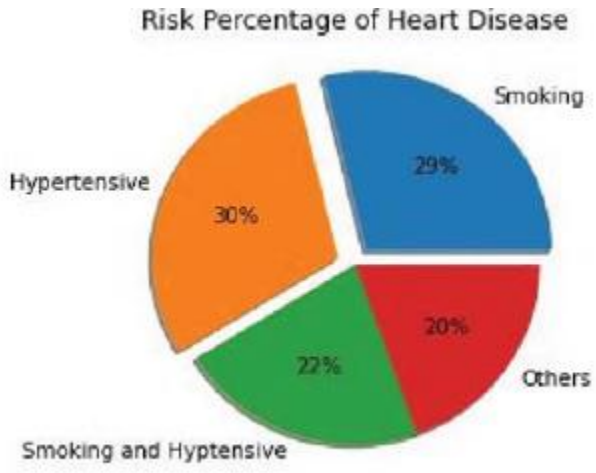

Figure 6. Analyzing of Risk Factor.

I examined the data and discovered that only 644 persons are impacted by heart disease for a different cause, out of about 4,239 data points and 16 features. From all of these factors, I derived that smoking and hypertension are the most important ones; the former accounts for about 29% and the latter for 30%. In contrast, the sum of the two is close to 22%, whereas the sum of the other reasons is only close to 20%. In conclusion, the results of this analysis show that these two main factors account for over 80% of the causes of heart disease.
also has the ability to forecast diabetes, cancer, and other diseases based on heart disease protocols; moreover, it can use a novel algorithm to achieve enhanced precision and efficiency.

VII. CONCLUSIONS

Applying a machine learning classification method to this study allows us to find that KNN, Logistic Regression, and SVM provide the best results when compared to other techniques. These three were crucial to my study; of the other four, Decision Tree Classifier was the worst and two were very close to having superior accuracy.
In addition to improved accuracy, users may use these models from inside a system and learn more about their present heart disease state. To further improve the accuracy of the current result, we may additionally compare other data mining approaches and hybrid procedures.

REFERENCES

[1] L. Ali et al., "An Optimized Stacked Support Vector Machines Based Expert System for the Effective Prediction of Heart Failure," IEEE Access, vol. 7, pp. 54007-54014, 2019, doi: 10.1109/ACCESS. 2019.2909969.
[2] A. Javeed, S. Zhou, L. Yongjian, I. Qasim, A. Noor, and R. Nour, "An Intelligent Learning System Based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection," IEEE Access, vol. 7, pp. 180235-180243, 2019, doi: 10.1109/ACCESS.2019.2952107.
[3] M. Gjoreski, A. Gradisek, B. Budna, M. Gams, and G. Poglajen, "Machine Learning and End-to-End Deep Learning for the Detection of Chronic Heart Failure from Heart Sounds," IEEE Access, vol. 8, pp. 20313-20324, 2020, doi: 10.1109/ACCESS.2020. 2968900.
[4] Aman ajmera, https://www.kaggle.com/ amanajmera1/framinghamheart- study-dataset
[5] Koo Ping Shung, "https://towardsdatascience.com /accuracy-precisionrecall- or-f1-331fb37c5cb9"
[6] R. Detrano et al., "International application of a new probability algorithm for the diagnosis of coronary artery disease," Am. J. Cardiol., vol. 64, no 5, pp. 304-310, 1989.
[7] J. H. Gennari, P Langley, and D. Fisher, "Models of incremental concept formation," Artif. Intell., vol. 40, no. 1-3, pp. 11-61, 1989
[8] K. Bache and M. Lichman, "UCI Machine Learning Repository," University of California Irvine School of Information, vol. 2008, no. 14/8. p. 0, 2013.
[9] Beant Kaurh, Williamjeet Singh, "Review on Heart Disease Prediction System using Data Mining

International Journal of applied exercise physiology

ISSN: 2322-3537
Vol-14 Issue-01 June 2025

Techniques", c IJRITCC, Vol.2, Issue: 10, p.p.3003-08, 2014.
[10] Amin, M.S., Telematics and Informatics, https://doi.org/10.1016Zj.tele.2018.11.007