Dhaka Univ. J. Sci. 73(2): 171-175, 2025 (July)

ISSN 1022 - 2502, E-ISSN 2408-8528
DOI: https://doi.org/10.3329.dujs.v73i2.82776

A Comparative Analysis of Various Machine Learning Techniques in Diagnosis of Heart Disease

Adeeb Shahriar Zaman and Md. Shapan Miah
Department of Mathematics, University of Dhaka, Dhaka-1000, Bangladesh.

(Received: 10 November 2024; Accepted: 25 May 2025)

Abstract

Nowadays, heart illness is somewhat common truth. Both the death rate and frequency are rising daily. In this study, three models-a simple logistic model using linear regression, K-Nearest Neighbors (K-NN), and Support Vector Machine (SVM)-are used for training and testing data in order to accurately predict heart disease. Following preprocessing, a ratio was used to divide the data into train and test sets. The results showed that the Support Vector Machine (SVM) approach was the most accurate model in terms of prediction accuracy, while the K-Nearest Neighbors (K-NN) and Logistic model with linear regression were relatively lesser accurate with respect to the same value of all parameters. To put it succinctly, these three prediction models were able to accurately forecast heart disease and exceeded accuracy. These findings suggest that these models are very practical and effective, and they can give physicians valuable information to help them identify and treat patients with heart disease more accurately.

KeyWords: Heart disease, Logistic Regression, K-NN algorithm, SVM algorithm, Machine Learning.

I. Introduction

Heart disease is an extremely prevalent condition, and each year more people die from it. Recent studies on heart disease have focused on early diagnosis, therapy, and prediction[1,2]. Among these, algorithms for machine learning (ML) have emerged as a promising technique for heart disease prediction. Large-scale data processing and analysis, sampling, model building, and future prediction are all possible using machine learning algorithms[3]. By examining clinical, physiological, and imaging data, machine learning algorithms can create prediction models and determine a patient's risk or hazard[4]. To forecast improved outcomes, algorithms like SVM, K-NN, and regression models with sigmoid functions have been employed.

The appeal of ML algorithms in heart disease prediction is in great importance. Firstly, it can help physician to assess risk of patients more correctly and providing treatments.

Second, it can increase diagnosis efficiency and accuracy. Lastly, it can lower the heart disease death rate[9].

In recent times, there have been some studies done on predictive modeling in cardiovascular diseases prediction[12,13]. We used our own codes to review some of these works to acquire better results. We used three robust machine learning algorithms which are Logistic Regression, K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) to predict the presence of heart disease in a patient and compared their accuracies, precision, recall and F1 score to find the best method among them.

Even though machine learning algorithms are used to detect cardiac problems, they still require clarifications and enhancements. We used a set of data in this research that may contain noise. The accuracy of models will increase as data quality continues to improve[9].

II. Data Sources

The dataset utilized in this work comes from the publicly accessible Kaggle platform[7]. It consists of 303 patients' data along with 14 features. These features are age, gender, chest pain, resting blood pressure, cholesterol, fasting blood pressure, resting electrocardiographic results, maximum heart rate, exercise induced angina (1 = yes; 0 = no), old peak (ST depression induced by exercise relative to rest), slope (the slope of the peak exercise ST segment), number of major vessels, thalassemia and target. Target 1 denotes a person with heart disease, while target 0 denotes a person with a normal heart. There were no missing values in the dataset. 13 of the 14 features' values are in integer format and the remaining one is in decimal format. We have shown the data of 5 patients along with 10 features in Table 1.

Table 1. Partial data

Age	Sex	Resting BP	Cholesterol	Fasting BS	Max HR	Exercise Angina	Old Peak	ST Slope	Heart Disease
63	M	145	233	Yes	150	No	2.3	0	1
44	M	130	233	No	179	Yes	0.4	2	1
54	F	135	304	Yes	170	No	0	2	1
55	F	128	205	No	130	Yes	2	1	0
38	M	138	175	No	173	No	0	2	1

Correspondence Author email: shapan@du.ac.bd

III. Statistical Analysis of Data

Each feature was statistically analyzed and maximum, minimum, average, standard deviation and median of each feature were calculated firstly.

Table 2. Feature List

Feature	Max	Min	Average	Standard Deviation	Median
Age	77	29	54.36634	9.08210	55
Resting BP	200	94	131.6238	17.5381	130
Cholesterol	564	126	246.264	51.8307	240
Fasting BS	1	0	0.148515	0.35619	0
Old Peak	6.2	0	1.039604	1.16107	0.8
Max HR	202	71	149.6469	22.9051	153
Heart Disease	1	0	0.544554	0.49883	1

IV. Model Introduction

Logistic model with simple Linear regression

The goal of logistic model is to estimate the probability of occurrence. The value range for the prediction should therefore be between 0 and 1. This method is used for classification problem only.

In linear regression model we have regression equation as $\hat{y} = b_1x_1 + b_2x_2 + ... + b_kx_k + b_0$. Here $\hat{y}$ is the dependent variable and $(x_1, x_2, ..., x_k)$ are the independent variables and b_i's are the regression coefficients. The sigmoid function is given by $f(z) = \dfrac{1}{1+e^{-z}}$. We can observe that the functional value approaches 1 as z tends to ∞ and 0 as z tends to $-\infty$. So, for any real value of z, the functional value will be in between 0 and 1. Now we will infuse the linear regression equation into the sigmoid function as follows.

$f(\vec{b}) = \dfrac{1}{1+e^{-(b_1x_1+b_2x_2+.....+b_kx_k+b_0)}}$. Determining the optimal $\vec{b}$ coefficients will help us to get a better result. Usually, 0.5 is considered as the threshold value. If $\hat{y}$ is greater than 0.5, the target variable is predicted to be 1 and if $\hat{y}$ is less than 0.5, the target variable is predicted to be 0. Thus, we can use a continuous model to find the target value which is either 0 or 1 in a classification problem.[10].

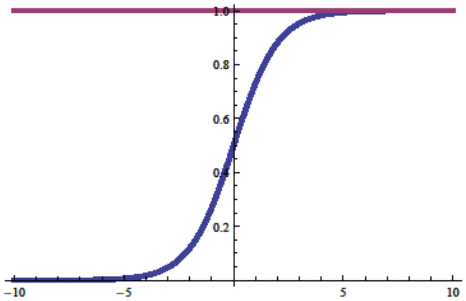

Fig. 1. Sigmoid function.

K-Nearest Neighbors (K-NN) Algorithm

A straightforward yet incredibly effective classification algorithm is K-Nearest Neighbors (KNN). It uses a distance metric to classify. Additionally, the algorithm is non-parametric. Finding the number of nearest neighbors is the first step in the K-NN approach. The distance between each training example and the query instance is then determined. Once the distance has been sorted, the minimum distance is used to identify the closest neighbors. The prediction is then made using the query instance's prediction value to be the simple majority of the nearest neighbor category[10].

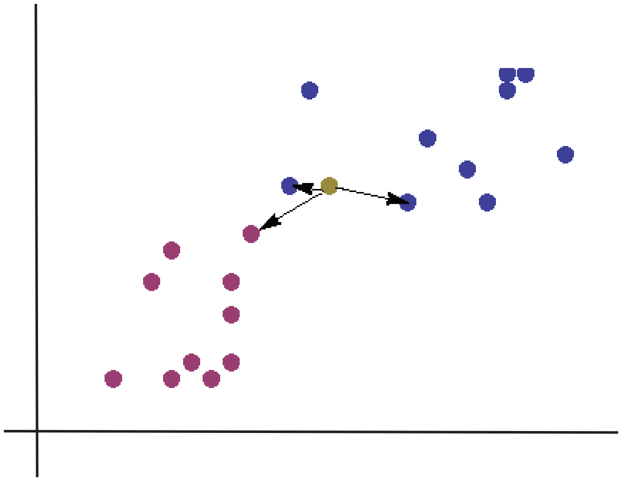

Fig. 2. K-NN algorithm.

Support Vector Machine Algorithm

One of the most often used supervised learning algorithms for both regression and classification problems is the Support Vector Machine (SVM). But it's mostly applied to machine learning classification problems.

In order to make it simple to classify fresh data points in the future, the SVM method aims to draw the best line or decision boundary that can divide n-dimensional space into classes. A hyperplane is the optimal choice boundary.

In order to create the hyperplane, SVM selects the extreme points and vectors. These are the points lying closest to the decision boundary or the hyperplane. The algorithm is known as a Support Vector Machine (SVM) because these extreme situations are referred to as support vectors[10].

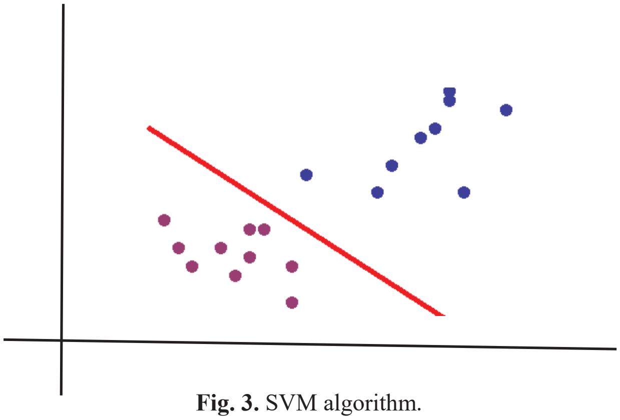

Fig. 3. SVM algorithm.

V. Results and Analysis

We separated our data set into two sections using a ratio 1: 1 after first preparing the data. In other words, 50% of the data are used to test the model, and the remaining 50% are used to train the model. Parameters including precision, recall, F1 score, confusion matrix, and accuracy are used to assess the models.

The confusion matrix consists of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). Here, 'Positive' denotes the prediction of the presence of heart disease and 'Negative' denotes the absence of it and 'True' implies the prediction is correct whereas 'False' implies it is incorrect. The confusion matrix for the logistic model with the simple linear regression model, the K-NN model, and the SVM model, are shown in Figures 4, 5 and 6, respectively.

		Predicted Class	
		0	1
Actual Class	0	TN 48	FP 21
	1	FN 8	TP 75

Fig. 4. Logistic regression model.

		Predicted Class	
		0	1
Actual Class	0	TN 55	FP 14
	1	FN 13	TP 70

Fig. 5. K-NN model.

		Predicted Class	
		0	1
Actual Class	0	TN 52	FP 17
	1	FN 8	TP 75

Fig. 6. SVM model.

Table 3. Modelling Evaluation

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	80.921%	78	90	84
K-NN model	82.236%	83	84	84
SVM model	83.552%	82	90	86

Precision, Recall, F1 score and Accuracy of the models are defined as follows:

- Precision = $\frac{TP}{TP+FP}$
- Recall = $\frac{TP}{TP+FN}$
- F1 Score = $\frac{2*Precision*Recall}{Precision+Recall}$
- Accuracy = $\frac{TP+TN}{TP+FP+TN+FN}$

From the above confusion matrix and model evaluation indexes the best performance in terms of accuracy of heart disease prediction was obtained by Support Vector Machine (SVM) method with accuracy 83.55% whereas the Logistic model with linear regression and K-nearest Neighbors methods have obtained the accuracy 82.24% and 80.92% respectively. All the above ML algorithms have accuracy more than 80% and can predict the heart disease of the patients fairly accurately. Fig.7 shows the accuracies of the models.

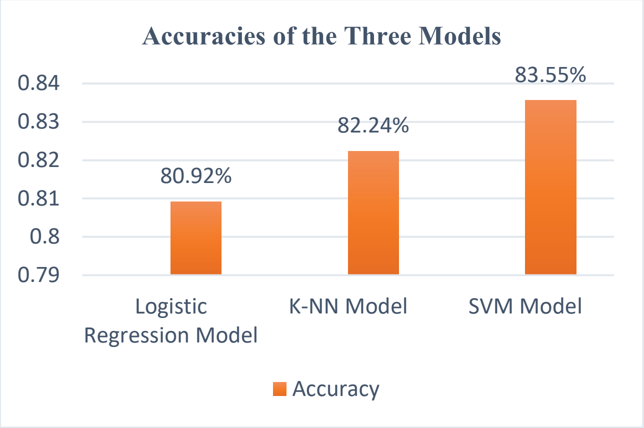

Fig. 7. Bar chart of the accuracy of the model.

We also performed 10-fold cross validation technique in order to prevent the issue of overfitting and bias. The mean accuracies of Logistic Regression, K-NN and SVM were respectively 84.4%, 79.6% and 85.7%. This shows that the SVM model is better than the other models in diagnosing heart diseases.

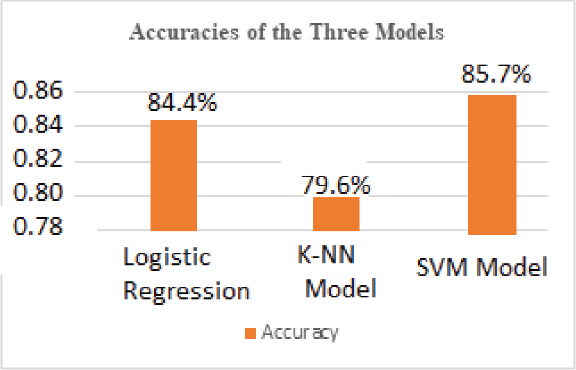

Fig. 8: Bar chart of the accuracy of the model after 10-fold cross validation.

VI. Conclusion

All over the world heart disease has become very common fatal disease [11]. Many people die from heart disease nowadays for not proper identification of disease. In this paper we have tried to employ three ML algorithms to identify heart disease. Among them the best performer was Support Vector Machine (SVM) model with accuracy 83.55% and recall 90%. This model has better result compared with the others if we consider both accuracy and recall. For medical diagnosis cases, recall parameter is essential since it shows the measure of all people who really have heart disease, how many the model actually detected. Considering these two parameters, we can be sure that the SVM is model is superior than the others.

The entire data set was divided into two categories in this paper: train and test sets, respectively, based on the ratio 1: 1. We trained our models using the train data set and evaluated them using the test data set.

The confusion matrix helps us to observe the classification of the test set. We observed that in SVM method, both true positive (TP) and true negative (TN) are high while false positive (FP) and false negative (FN) are low. This means that this model is able to correctly predict whether a patient has heart disease or not.

In summary, we can say that all the three ML models are able to predict the heart disease. Among them the Support Vector Machine (SVM) method with accuracy 83.55% is more accurate than Logistic with Linear Regression and KNN methods with accuracy 80.92% and 80.24% respectively.

Acknowledgement

We are grateful to Tanveer Ahmed Khan Fahim for his assistance with the cross-referencing in our paper.

References

1. Zang, Hengyi, et al. 2024 Evaluating the Social Impact of AI in Manufacturing: A Methodological Framework for Ethical Production. Academic Journal of Sociology and Management, **2(1)**, pp. 21-25.

2. Dong, Xinqi, et al. 2024 The Prediction Trend of Enterprise Financial Risk based on Machine Learning ARIMA Model." Journal of Theory and Practice of Engineering Science, **4(1)**, pp. 65-71.

3. Liu, Shun, et al. 2023: Financial Time-Series Forecasting: Towards Synergizing Performance and Interpretability Within a Hybrid Machine Learning Approach. 2023. arXiv preprint arXiv: 2401.00534

4. Cai, G., J. Qian, Song, T., Q. Zhang, and Liu, B., 2023. A deep learning-based algorithm for crop Disease identification positioning using computer vision. International Journal of Computer Science and Information Technology, **1(1),** pp.85-92.

5. Liu, B., Cai, G., Qian, J., Song, T. and Zhang, Q., 2023. Machine learning model training and practice: a study on constructing a novel drug detection system. International Journal of Computer Science and Information Technology, **1(1)**, pp.139-146.

6. Liu, B. 2023. Based on Intelligent Advertising Recommendation and Abnormal Advertising Monuitoring System in the Field of Machine Learning. International Journal of Computer Science and Information Technology, **1(1)**, pp. 17-23.

7. Data of heart disease patients, https://www.kaggle.com/datasets/yasserh/heart-disease-dataset

8. Aurelien Geron, 2019 Hands on Machine Learning with Scikit-Learn, Keras and Tensoflow, O'Reilly Media, Inc, 2nd edition.

9. Rammal F H, Z A. Emam Heart Failure Prediction Models using Big Data Techniques[J]. International Journal of Advanced Computer Science and Applications (IJACSA), 2018,9.

10. Bishop, C.M. and N.M. Nasrabadi,2006. Pattern recognition and machine learning **4(4)** 738). New York: springer

11. Deaths from cardiovascular disease surged globally, https://world-heart-federation.org/news/deaths-from-cardiovascular-disease-surged-60-globally-over-the-last-30-years-report/ (Accessed: 24 May 2025).

12. Hayudini, M.A., D.A.K. Kiram, Kiram, M.M. Abduljalil, A.H., Latorre, N.J. and Sahibad, F.B., 2025. Predictive Modeling in Cardiovascular Disease: An Investigation of Random Forests. Natural Sciences Engineering and Technology Journal, **5(1),** pp.393-404.

13. Krupadanam, C. and Narendran, S., 2025, March. Analysis of convolutional neural network algorithm for cardiac diagnosis compared with accuracy of artificial neural network algorithm. In AIP Conference Proceedings **3252(1)** p. 020071). AIP Publishing LLC.