**International Journal of Computer Sciences and Engineering**
Vol.**11**, Special Issue.**1**, pp.**76-80**, November **2023**
ISSN: 2347-2693 (Online)
Available online at: www.ijcseonline.org

ISSN: 2347-2693 (E)

Research Paper

# Methodical Prediction of Cardiovascular Disease Using Consolidated Machine Learning Classification Algorithms and Analysis

**Aparna Datta[1*], Sreeja Ghosh[2]**

[1]Master of Computer Applications, Meghnad Saha Institute of Technology, Kolkata, India
[2]Computer Science and Engineering, Asansol Engineering College, Asansol, India

**Corresponding Author: sarkar23aparna@gmail.com*

***Abstract:*** Heart disease has been a serious threat to mankind. According to research 7 out of 10 people die due to heart failure. In this paper, we have proposed a framework using which we can determine if a person has heart ailments or not. We have used various ML classification algorithms such as Logistic Regression, SVM, Random Forest, Decision tree, KNN, MLP, and Neural Network to determine the existence of heart disease. The best result has been obtained by Random Forest. Timely detection of a disease can save many people's lives, thereby controlling the mortality rate to some extent.

***Keywords:*** Cardiovascular disease, KNN, SVM, Neural Network, Random Forest, Decision Tree, MLP, Logistic regression

## 1. Introduction

'Heart diseases' or 'cardiovascular diseases' (CVD) are a growing threat to mankind. The heart is always considered the most intricate organ of the body. Thus, the study of the heart and the threats revolving around it has always been a challenge to scientists. But from time-to-time biologists were able to discover leading risk factors for heart diseases, including high BP, high low-density lipoprotein (LDL) cholesterol, diabetes, smoking and second-hand smoke exposure, obesity, unbalanced diet, physical inactivity, etc., and provided mankind with the best aid possible.

But with the rapid growth of the population and increasing heart diseases arrives the crisis of such experts. According to the 'Centres for Disease Control and Prevention (CDC)', "approximately every 40 seconds an American will have a heart attack. Every year, 805,000 Americans have a heart attack, 605,000 of them for the first time. About 12 percent of people who have a heart attack will die from it." [11]. Early detection and accurate diagnosis of CVDs are crucial for timely intervention and effective management of such patients. In such conditions, the growth of technology is now a boon in the applied medical science field. With advancements in technology came ML algorithms, which emerged as a promising solution in this period of crisis.

Machine learning algorithms use large amounts of datasets to identify patterns and make predictions based on data that can indicate and help in the early detection of CVD patients. In this paper, we will explore the role of classification for CVD detection and identification, its potential benefits, challenges while using them, limitations, and their future application in the field of medical science.

We have taken the heart disease dataset from Kaggle [9].Out of the 14 features, we have worked with 10 attributes after feature importance and scaling. The data is given as input to different models and the existence of any cardiovascular disease is examined.

Related studies in this field are stated in the next section. Then our proposed framework and methodology is explained followed by results, comparison, and conclusion.

## 2. Related Work

Our proposed idea has been an area of interest of many researchers. We have covered a few relevant works, as follows.

Table 1

| Reference | Year | Algorithm used | Accuracy |
|---|---|---|---|
| Mohan et al. [1] | 2019 | Hybrid random forest with a linear model (HRFLM) | 88.7% |
| DINESH K G, et al. [2] | 2018 | Logistic regression-86.52%. Random Forest-80.99%. Naïve Bayes- 84.27% Gradient boosting- 84.27% SVM- 79.77% | 86.52% |
| Shah et al. [3] | 2020 | Decision tree, Naïve Bayes, KNN, Random Forest. | 91.6% |
| Bharti et al. [4] | 2021 | Logistic regression, KNN, Random Forest, Decision Tree, SVM | 92.7% |
| Ul Haq et al. [5] | | Logistic regression, K-NN, ANN, SVM, NB, DT, and random forest | |
| Mamun Ali et al. [6] | 2021 | KNN.MLP, Random Forest | 89.6% |
| Ping Li et al. [7] | 2020 | MLP, SVM, Fuzzy Logic | 92.32% |
| Dwivedi et al. [8] | 2016 | ANN, SVN, Decision Trees, Naïve Bayes | 81.83% |
| Singh et al. [10] | 2020 | Linear Regression, Decision Tree, KNN, SVM | 81.75% |

 

## 3. Theory/Calculation

Let D is the dataset which is the input of our system. We are classifying using different ML algorithms into H1 and H2.
D €{ H1,H2},where H1-->presence of heart disease and H2--->absence of heart disease

## 4. Experimental Method/Procedure/Design

The heart disease data set from Kaggle [9] is from 1988 and contains four databases: Hungary, Cleveland, Long Beach V, and Switzerland. It originally contains 76 parameters, including the prediction, but all experiments use a subset of 14 attributes. The target field determines the existence of cardiovascular disease. It gives integer values, 0 for absence and 1 for the presence of heart disease.
Details of the attributes: [9]

1. sex
2. age
3. chest pain type
4. number of major vessels
5. resting blood pressure
6. serum cholesterol
7. fasting blood sugar
8. resting ECG results
9. that, 0- normal, 1 = fixed defect, 2 = reversible defect
10. exercise-induced angina
11. old peak = ST depression induced by exercise relative to rest
12. slope of the peak exercise ST segment
13. maximum heart rate.

After feature extraction, the dataset is split into training data (80%) and test data (20%). Then feature scaling is done to normalize the range of independent variables, in the pre-processing stage.

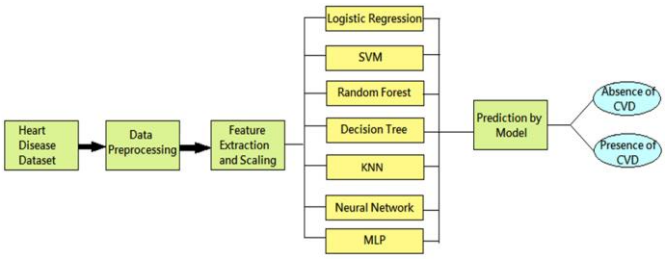


**Figure 1**: Block diagram of the proposed approach for heart disease prediction

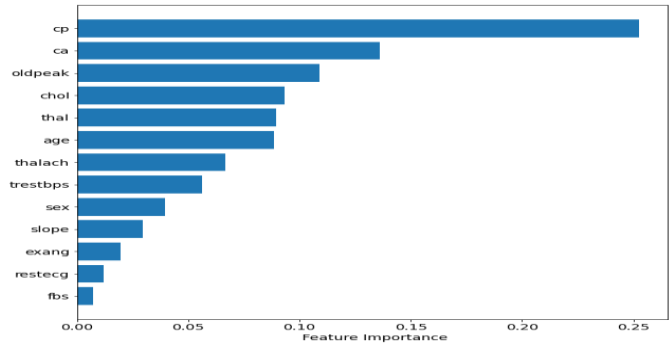


**Figure 2**. Shows result of feature importance

### 4.1 Classification

Two binary classes are involved, 0 and 1, to determine the absence and presence of heart disease respectively. Various classification algorithms are used, as follows:

**Logistic Regression** is a statistical method for developing machine learning models where the dependent variable has a binary value (0 or 1). It has been used to classify the presence or absence of heart disease.
Train Accuracy of Logistic Regression=0.8719512195121951
Test Accuracy of Logistic Regression= 0.7951219512195122
Best result was obtained with these values of hyperparameters 'C': 0.5, 'penalty': 'l2', 'random_state': 0
**Support Vector Machine (SVM)**
Kernel= 'poly', degree=3, coef0=2, C=8
Train Accuracy of SVM Classifier= 1.0
Test Accuracy of SVM Classifier= 0.9853658536585366
**Random Forest** – n_estimators=39, max depth=9, random_state=42
Train Accuracy of Random Forest= 1.0
Test Accuracy of Random Forest= 1.0
**K-Nearest Neighbours (KNN)** K-neighbours=1
Train Accuracy of KNN Classifier= 1.0
Test Accuracy of KNN Classifier= 0.9853658536585366
**Decision Tree Classifier (DT)**
max_depth=10, min_samples_leaf=1, min_samples_split=2, random_state=42
Train Accuracy of DT= 1.0
Test Accuracy of DT= 0.9853658536585366
**Multilayer Perceptron (MLP)**
Train Accuracy of MLP= 0.9109756097560976
Test Accuracy of MLP 0.8390243902439024

**Neural Network (NN)**

**Model summary**

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 100)               1400

 dense_1 (Dense)             (None, 64)                6464

 dropout (Dropout)           (None, 64)                0

 dense_2 (Dense)             (None, 128)               8320

 dropout_1 (Dropout)         (None, 128)               0

 dense_3 (Dense)             (None, 1)                 129

=================================================================
Total params: 16,313
Trainable params: 16,313
Non-trainable params: 0
_________________________________________________________________
```

## 5. Results and Discussion

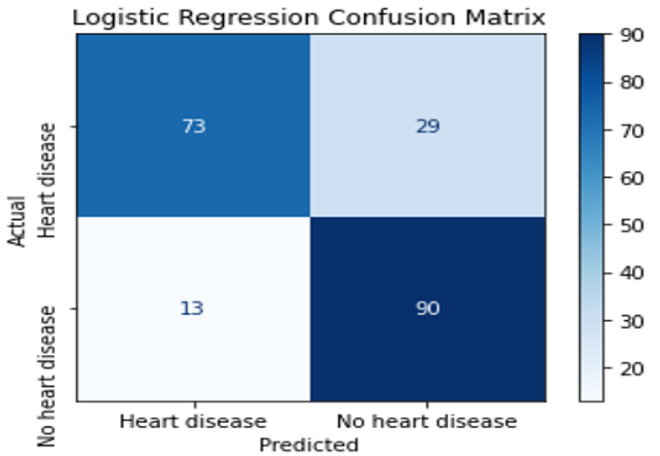


**Figure 3**. Confusion Matrix of Logistic Regression

 

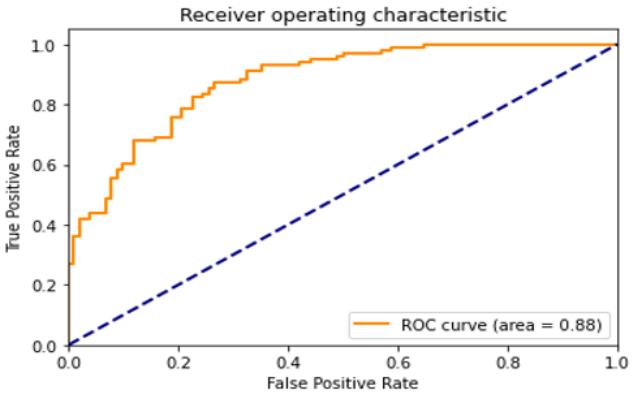


**Figure 4**. ROC graph of Logistic Regression

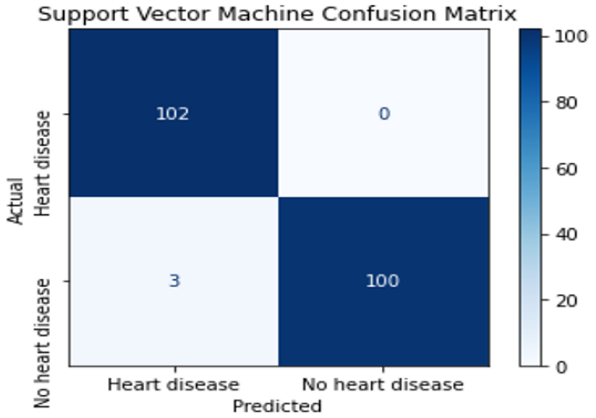


**Figure 5**. Confusion Matrix of SVM

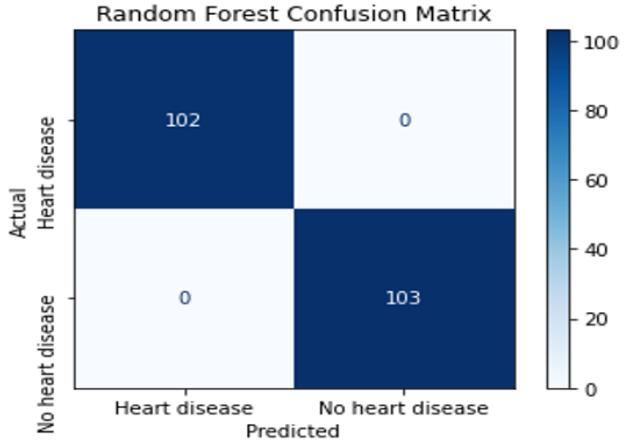


**Figure 6**. Confusion matrix of Random Forest

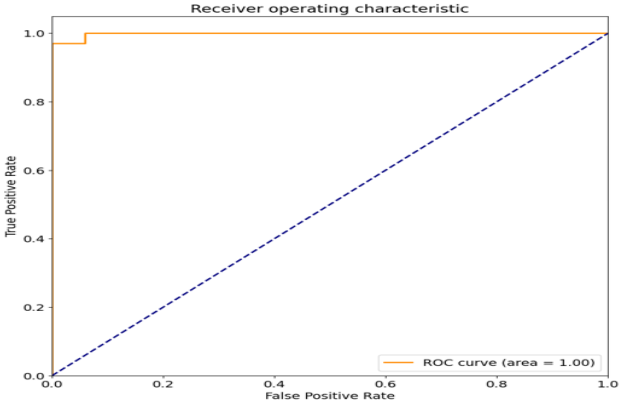


**Figure 7**. ROC graph of Random Forest

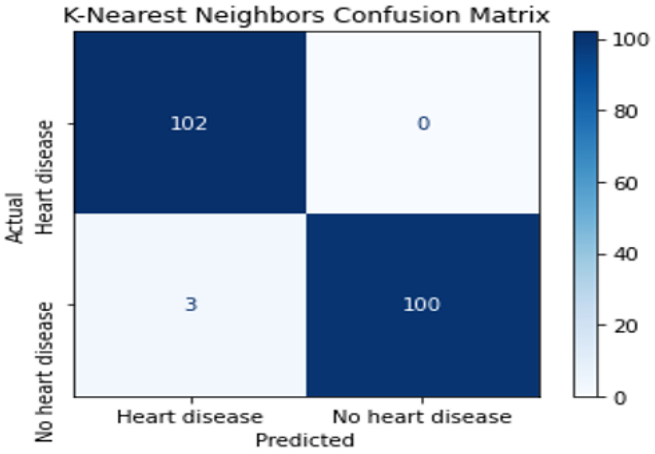


**Figure 8**. Confusion Matrix of K-Nearest Neighbor

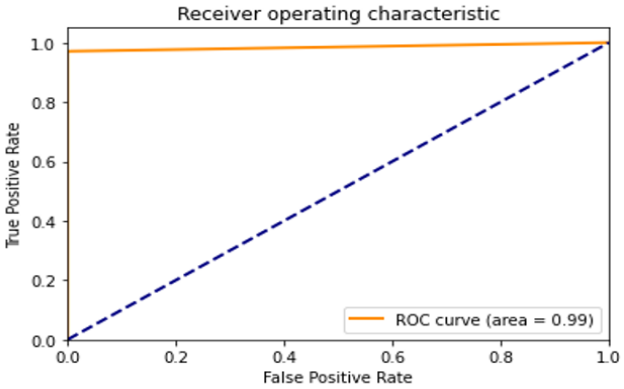


**Figure 9**. ROC graph of K-Nearest Neighbor

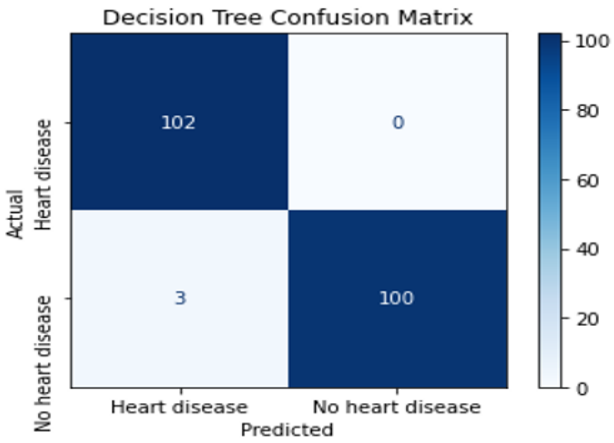


**Figure 10**. Confusion Matrix of Decision Tree

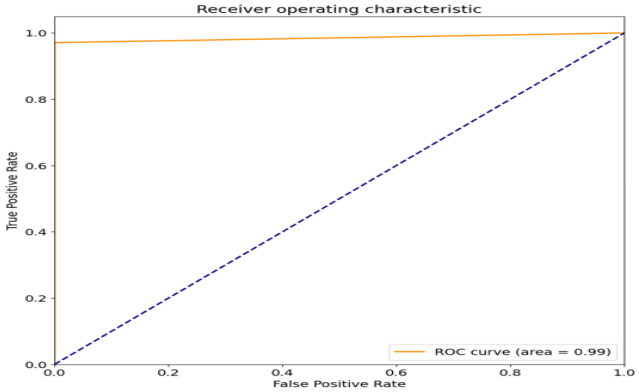


**Figure 11**. ROC graph of Decision Tree

 

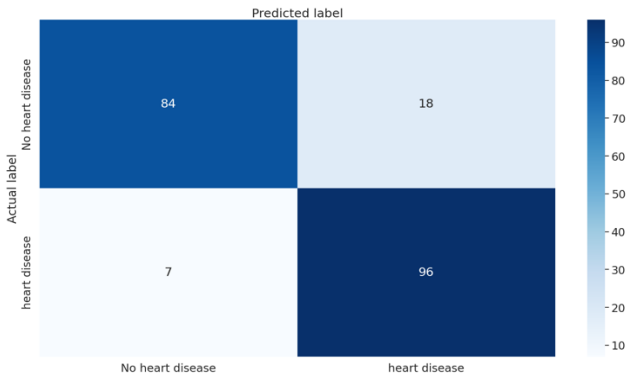


**Figure 12**. Confusion matrix of Neural Network

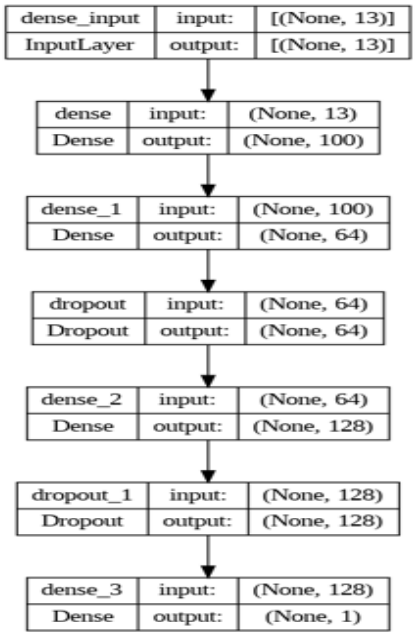


**Figure 13**. Details of each layer of Neural Network

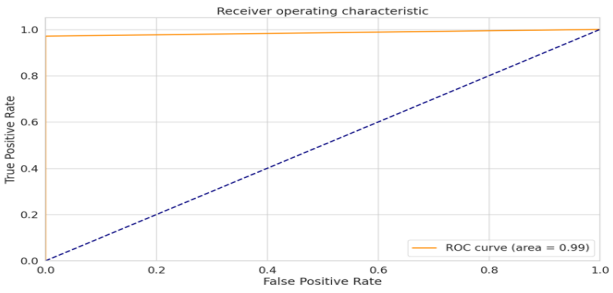


**Figure 14**. ROC graph of Neural Network

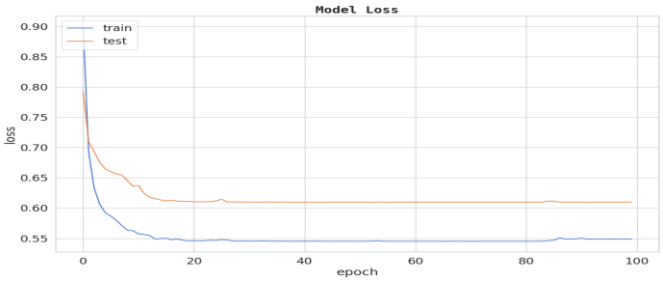


**Figure 15**. Loss vs Epochs for NN

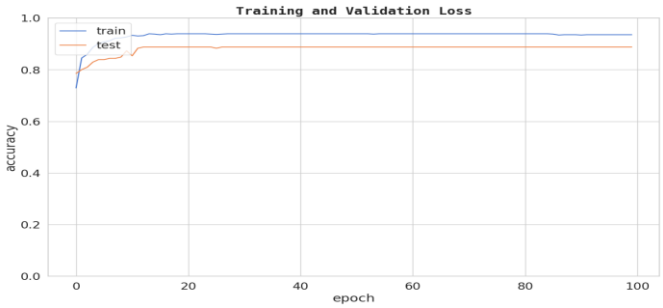


**Figure 16**. Accuracy vs Epochs graph for NN

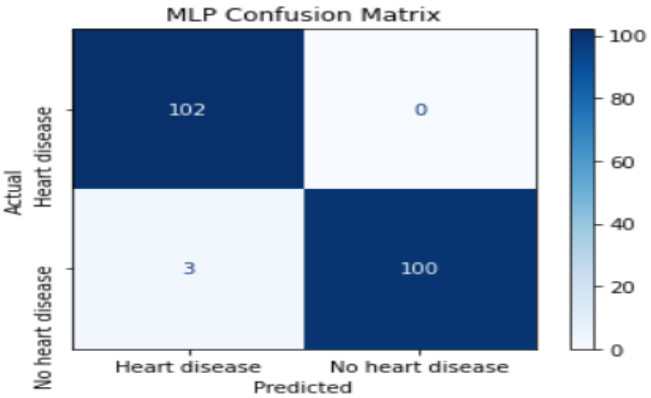


**Figure 17.** Confusion Matrix of MLP

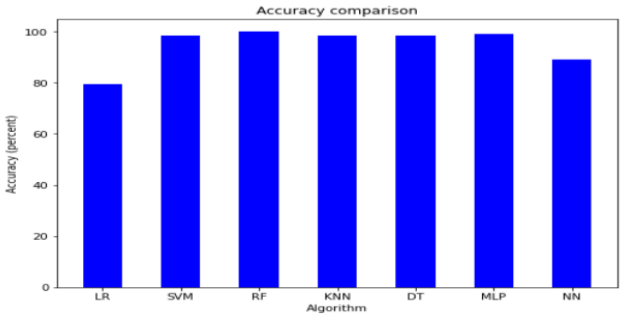


**Figure 18.** Comparison of accuracy of different ML algorithms used

**Table 2.** Comparison of classification reports of the algorithms

| | Logistic Regression | | | | Support Vector Machine | | | | Random Forest | | | | KNN | | | | Decision Tree | | | | Neural Network | | | | Multilayer Perceptron | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Cl | P | R | F | Cl | P | R | F | Cl | P | R | F | Cl | P | R | F | Cl | P | R | F | Cl | P | R | F | Cl | P | R | F | |
| 0 | 0.85 | 0.72 | 0.78 | 0 | 0.97 | 1 | 0.99 | 0 | 1 | 1 | 1 | 0 | 0.97 | 1 | 0.99 | 0 | 0.97 | 1 | 0.99 | 0 | 0.92 | 0.82 | 0.87 | 0 | 0.97 | 1 | 0.99 | |
| 1 | 0.76 | 0.87 | 0.81 | 1 | 1 | 0.97 | 0.99 | 1 | 1 | 1 | 1 | 1 | 1 | 0.97 | 0.99 | 1 | 1 | 0.97 | 0.99 | 1 | 0.84 | 0.93 | 0.88 | 1 | 1 | 0.97 | 0.99 | |
| Acc | | | 0.8 | Acc | | | 0.99 | Acc | | | 1 | Acc | | | 0.99 | Acc | | | 0.99 | Acc | | | 0.88 | Acc | | | 0.99 | |

Cl – Class
P – Precision
R – Recall
F- F1 score

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (2)$$

 

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3)$$

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

TP – True Positive
TN – True Negative
FP - False Positive
FN - False Negative

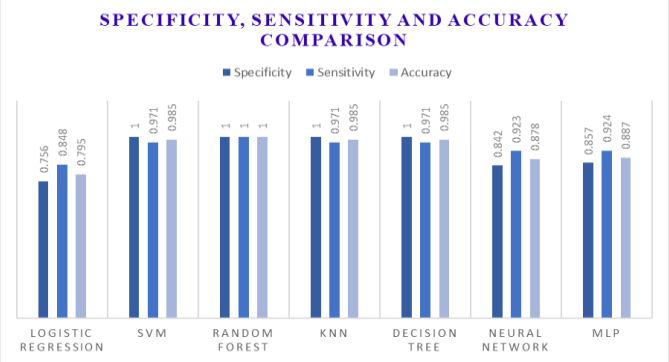


**Figure 19.** Comparison between Specificity, Sensitivity, and Accuracy of ML algorithms

**Table 3.** Shows accuracy percentage of all the used algorithms

| Algorithm | Accuracy |
|---|---|
| Logistic Regression | 79.5% |
| Support Vector Machine | 98.5% |
| Random Forest | 100% |
| K-Nearest Neighbour | 98.5% |
| Decision Tree | 98.5% |
| Neural Network | 89% |
| Multilayer Perceptron | 99% |

## 6. Conclusion and Future Scope

In this paper, we have proposed a methodical prediction of cardiovascular disease (CVD) also known as heart disease. The novelty lies in the accuracy percentage. We have seen precision values for all classifiers and found that Random Forest gives us the highest accuracy of 100%, followed by 99% for MLP, 98.5% by KNN, SVM, and Decision Tree, 89% by Neural Network and 79.5% in Logistic Regression. Therefore, the best algorithm we can use to predict heart disease is Random Forest. This model can be implemented to diagnose patients at a much earlier stage and thereby saving them from life threats. This can be a boon for the medical science field, whereby doctors can be assured about the presence of any kind of cardiovascular disease and start the required treatment at the earliest. In the future, more work can be done to detect the type and severity of the disease, if present. It would aid the medical team to diagnose and give immediate medication, hence the mortality rate can be curbed to a great extent.

**Data Availability**
The heart disease data set was acquired from Kaggle [9].

**Conflict of Interest**
Do not have any conflict of interest.

**Funding Source**
None

**Authors' Contributions**
Author-1 Detail Analysis of various approaches on this topic.
Author-2 Researched literature and conceived the study, wrote the first draft of the manuscript.

**Acknowledgments** We appreciate my friends and colleagues for their encouragement, assistance, and insightful talks that have greatly aided in the completion of this work.

## References

[1] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," IEEE Access, Vol.**7**, pp.**81542–81554, 2019**, doi: 10.1109/ACCESS.2019.2923707.
[2] Dinesh Kumar G, Arumugaraj K, Santhosh Kumar D and Mareeswari V, "Prediction of Cardiovascular Disease Using Machine Learning Algorithms", Proceeding of 2018 IEEE International Conference on Current Trends toward Converging Technologies, Coimbatore, India.
[3] D. Shah, S. Patel, and S. K. Bharti, "Heart Disease Prediction using Machine Learning Techniques," SN Comput Sci, Vol.1, no.6, pp.**345**, Nov. **2020**, doi: 10.1007/s42979-020-00365-y.
[4] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, "Prediction of Heart DiseaseUsing a Combination of Machine Learning and Deep Learning," Comput Intell Neurosci, Vol.2021, 2021, doi: 10.1155/2021/8387680.
[5] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, R. Sun, and I. Garciá-Magarinõ, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms," Mobile Information Systems, Vol.2018, 2018, doi: 10.1155/2018/3860146.
[6] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. W. Quinn, and M. A. Moni, "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison," Comput BiolMed, vol. 136, Sep. 2021, doi: 10.1016/j.compbiomed.2021.104672.
[7] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," IEEE Access, Vol.**8**, pp.**107562–107582, 2020**, doi: 10.1109/ACCESS.2020.3001149.
[8] A. K. Dwivedi, "Performance evaluation of different machine learning techniques for prediction of heart disease," Neural Comput Appl, Vol.**29**, no.**10**, pp.**685–693**, May **2018**, doi: 10.1007/s00521-016-2604-1.
[9] https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset
[10]Archana Singh, Rakesh Kumar, "Heart Disease Prediction Using Machine Learning Algorithms", 2020 International Conference on Electrical and Electronics Engineering (ICE3-2020)
[11] https://www.cdc.gov/heartdisease/facts.htm

## AUTHOR'S PROFILE

Aparna Datta holds an MCA and MTech degree from Maulana Abul Kalam Azad University of Technology, West Bengal. She is currently an Assistant Professor in the Department of Computer Application of Meghnad Saha Institute of Technology. Her research interests include medical image processing and machine learning.

Sreeja Ghosh is a final year BTech student in the department of Computer Science and Engineering from Maulana Abul Kalam Azad University of Technology. She is currently doing internship at Intel Corporation, Bangalore. She has keen interest in research in the fields of machine learning and artificial intelligence.

 