IJBPAS, June, 2023, 12(6): 2630-2636

ISSN: 2277–4998

International Journal of Biology, Pharmacy and Allied Sciences (IJBPAS)

'A Bridge Between Laboratory and Reader'

www.ijbpas.com

# COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS FOR CARDIOVASCULAR DISEASE PREDICTION

**PRIYADARSHINEE S[1] AND PANDA M[2*]**

**1:** Research Scholar, G.M. University, Sambalpur, Odisha, India

**2:** Associate Professor, G.M. University, Sambalpur, Odisha, India

[*]**Corresponding Author: Dr. Madhumita Panda: E Mail:** mpanda.gmu@gmail.com

**Received 15th July 2022; Revised 20th Aug. 2022; Accepted 5th Oct. 2022; Available online 1st June 2023**

https://doi.org/10.31032/IJBPAS/2023/12.6.7183

**ABSTRACT**

Today machine learning is playing an important role especially in the healthcare field. Heart disorders, generally referred to as cardiovascular diseases are the main cause of death in the world. The number of tests required for the detection of heart disease is decreasing due to machine learning techniques. This paper looks at heart failure survivors from a group of 299 individuals who were hospitalised to the hospital. The goal is to use of machine learning models that can enhance the predictability of cardiac patient survival. In this paper, we have evaluated the accuracy of seven machine learning methods for cardiac illness prediction, including Nave Bayes (NB), Decision Tree (DT), K Nearest Neighbour (KNN) and Logistic Regression (LR), Random Forest (RF), Extra Tree (ET) and Ridge Classifiers (RC). The comparative study has proven Random Forest (RF) with a maximum accuracy (87.77%) with the lowest error rate.

**Keywords: Cardiovascular diseases, Nave Bayes, Decision Tree, K Nearest Neighbour, Logistic Regression, Random Forest, Extra Tree, Ridge Classifiers**

## I. INTRODUCTION

In the last few years, cardiovascular diseases have emerged as one of the most common causes of deaths worldwide. Heart disease diagnosis is the process of identifying or forecasting heart disease based on patient information. In particular, if the patient has multiple diseases, doctors might not be able to accurately identify the patient in a short amount of time. Because of this, determining the presence of cardiac disease has been a challenging task that requires training and expertise. A wrong

diagnosis could leave the patient dead or disabled **[1]**. Early recognition of the initial signs of cardiovascular diseases and the unceasing medical supervision can help in reducing rising number of patients and eventually the mortality rate. Medical experts and practitioners can forecast cardiac disease with the aid of the disease prediction model. Machine learning techniques can be used in conjunction with the massive quantity of data that can be gathered via digital devices (either by the patient or in the hospital) to diagnose.

Machine learning can be defined as learning from natural phenomena and natural things **[2]**. It is a subset of Artificial Intelligence (AI), which is a broad field of learning in which machines mimic human abilities. It provides an effective testing technology that is founded on training and testing. In this research work, we have employed biological factors as testing data such as smoking, anaemia, sex, age, and so on. On the basis of these, a comparison is done in terms of algorithm correctness. The rest of the paper is organized as follows: Section II presents the related works. The methodology containing a brief discussion on specification of the data set and algorithms used in given in section III. Section IV presents the results and discussion and finally, section V concludes the paper.

## II. RELATED WORKS

Many researches have been carried out in comparing various algorithms which can give best accuracy in predicting the heart disease at an earlier stage.

The machine learning algorithms used in this study **[3]** were Random Forest, Naïve Bayes, KNN and Logistic Regression. The results of comparison showed that Random Forest achieve high classification accuracy of 73 %, specificity of 65% and sensitivity of 80%.

In this paper **[4]** the author predicted the existence of cardiac condition in human body more accurately. Decision trees, Naive Bayes, and KNN were employed as data mining classification methods. According to the results, KNN produced more accurate results than Naive Bayes and Decision Trees.

In this paper **[5]** the authors employed six algorithms for predicting the onset of cardiac disease. According to the findings of the experiments, logistic regression technique was seen to be the best algorithm for predicting cardiac disease since it had a good accuracy of 85%.

The author proposed a hybridization strategy **[6]** in which Artificial Neural Network (ANN) and Decision Tree were used to improve heart disease prediction performance using WEKA tool. In comparison with separate algorithms, hybrid decision tree performed the best

giving a prediction accuracy of 78.14 percent.

To predict cardiac disorders, the author utilised four machine learning methods namely SVM, Decision Tree, Logistic Regression and KNN **[7]**. The Random woodland Machine learning classifier achieved a greater precision of 85 percent, ROC AUC score of 0.8675, and execution time of 1.09 seconds in the prediction of cardiovascular disease.

Several Machine Learning techniques **[8]** including Logistic Regression, Decision Tree, Naive Bayes, Random Forest, SVM, and KNN were used to classify a cardiovascular dataset. With a 73 percent accuracy, the Decision Tree provided the best outcome.

## III. METHODOLOGY

In order to predict cardiovascular illness, we have used seven classification techniques namely Naive Bayes (NB), K Nearest Neighbour (KNN), Decision Tree (DT), Logistic Regression (LR), Random Forest (RF), Extra Tree (ET), and Ridge Classifier (RC). In this study, 30% of the data are utilised for testing, while 70% of the data are used for training. After applying the machine learning algorithms to the dataset, the accuracy rate is computed. The main objective was to find an algorithm that could classify the given dataset most accurately.

### a. Narration of the Dataset

The heart-failure-clinical-records-dataset for the proposed work was taken from Kaggle **[9]**. It includes 299 patients' medical records with heart problems who were accumulated during the time of follow-up, with each profile of the patient containing 13 clinical attributes. There are 194 men and 105 women among the 299 records. All of the patients are beyond the age of 40. In the target class, 1 denotes the deceased and 0 denotes the alive. The **Table 1** provides an overview of the data set.

**Table 1: Specification of the Dataset**

| S. No. | Attributes | Description | Measured In | Range |
|---|---|---|---|---|
| 1 | Age | The patient's age | Years | 40 - 95 |
| 2 | Anaemia | Red blood cell or haemoglobin deficiency | Boolean | 0, 1 |
| 3 | Creatinine phosphokinase (CPK) | CPK enzyme levels in the blood | mcg/L | 23 -7861 |
| 4 | Diabetes | If the patient suffers from diabetes | Boolean | 0, 1 |
| 5 | Ejection fraction | % Of blood leaving | Percentage | 14 - 80 |
| 6 | High_blood_pressure | If a patient has hypertension | Boolean | 0, 1 |
| 7 | Platelets | The number of platelets in the blood | kilo platelets/Ml | 25.01-850.00 |
| 8 | Serum creatinine | Creatinine concentration in the blood | mg/dL | 0.50 -9.40 |
| 9 | Serum sodium | Sodium levels in the blood | mEq/L | 114 -148 |
| 10 | Sex | Woman or man | Binary | 0, 1 |
| 11 | Smoking | If the patient is a smoker | Boolean | 0, 1 |
| 12 | Time | Period of follow-up | Days | 4 - 285 |
| 13 | (Target) death event | If the patient died during the period of follow-up | Boolean | 0, 1 |

**b. Classification Algorithms**

Heart disease prediction is done using seven different Machine Learning algorithms. The datasets are fit to different algorithms to know their accuracy value.

**Naive Bayes**

The Bayes' Theorem is applied to the development of a group of classification methods known as Naive Bayes classifiers. It is a set of algorithms, not a single algorithm, that all share a basic principle. That is, every pair of features being classified is distinct from the others **[10]**.

**K – Nearest Neighbour**

The K-Nearest Neighbour (KNN) approach is one of the most basic but effective categorization strategies. This method finds the first k data points in the training set that are most similar to the data point for which a target value isn't available, and it assigns the average value of those data points to the missing data points **[11]**. In this study we set the value of k=3.

**Logistic regression**

Popular machine learning algorithm logistic regression belongs to the supervised learning methodology. Logistic regression forecasts the outcome of a categorical dependent variable. The result must therefore be a discrete or categorical value. It can be either True or False, 0 or 1, or Yes or No. **[12]**. In our study, for Logistic regression, we have set the hypermeter value n_estimator=100, random_state=0 .

**Decision Tree**

An approach for categorising both categorical and numerical data is a decision tree. A tree-shaped graph's data is straightforward to implement and analyse. The study of the decision tree model is based on three nodes **[13]**. In our work for Decision tree algorithm, we have set the hypermeter value as max_depth=2, random_state=42.

• Root node: The primary node, upon which all other nodes are founded.

• Interior node: manages numerous attributes.

• Leaf node: Display each test's outcome.

**Random Forest Algorithm**

It is a supervised machine learning algorithm that is employed for regression and classification **[14]**. The decision trees are built using data samples, and predictions are obtained from each one. The best answer is then chosen utilising this algorithm's voting mechanism. Given that it is the most common algorithm, it has simplicity and diversity. This method creates several decision trees, which it then blends. It provides more reliable and accurate prediction. It corrects random decision trees' overfitting to their training set. This algorithm was developed using the "bagging" technique. In this study we have set the hypermeter value as

n_estimator=200,random_state=1 in random forest algorithm.

**Extra Trees**

Extra Trees is a method for enhancing accuracy and computing efficiency by combining bagging classifiers with traditional tree-based approaches **[15]**. The primary distinctions from other tree-based algorithms are that it can split the node by randomly selecting cut-points and building the trees utilising all of the learning samples. In this research we have set the hyperparameter value as max_iter=500, Random_state=42 in Extra tree algorithm.

**Ridge Classifier**

Using the Ridge Classifier, which is based on the Ridge Regression Method, the label data is transformed into the range [-1, 1] and uses the Ridge Regression Method to address the issue. For data with many classes, multiple-output regression is used, with the highest prediction value accepted as the target class **[15]**. In the present work we have set the hyperparameter value as max_iter=100, Random_state=123 in Ridge Classifier.

## IV. RESULTS AND DISCUSSION

Python was used as a programming language for doing the comparative study in finding the accuracy of the seven algorithms namely Naïve Bayes (NB), Logistic Regression (LR), Decision Tree (DT), K Nearest Neighbour (KNN), Random Forest (RF), Extra Tree (ET) and Ridge Classifier (RC) on the data set of Kaggle **[9]**. **Table 2** displays the accuracy of various algorithms.

The accuracies obtained from the analysis is shown graphically for better understanding in **Figure 1**.

From the above **Table 2 and Figure 1**, it is clearly seen that the Random Forest performs the best, with an accuracy of 87.77 percent. Decision Tree is second good classifier with 85.55 percent accuracy.

**Table 2: Accuracy Comparison of Algorithms for Heart Disease Prediction**

| Classifier | Accuracy (%) | Inaccuracy (%) |
|---|---|---|
| Naïve Bayes (NB) | 75.55 | 24.45 |
| Decision Tree (DT) | 85.55 | 14.45 |
| Logistic Regression (LR) | 81.11 | 18.89 |
| K Nearest Neighbour (KNN) | 74.44 | 25.56 |
| Extra Tree (ET) | 83.33 | 16.67 |
| Random Forest (RF) | 87.77 | 12.23 |
| Ridge Classifier (RC) | 81.11 | 18.89 |

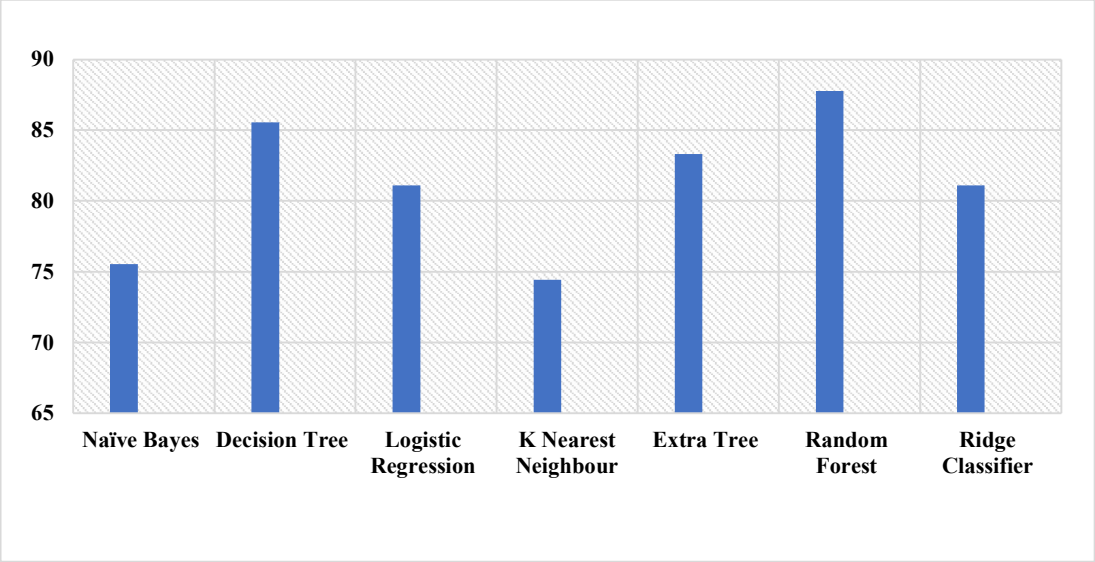


**Figure 1: Comparison of Classifier's Accuracy**

## V. CONCLUSION

Cardiac disease is a main cause of death in India and also in rest of the globe. An important step in the development of medicine is the ability of early identification of cardiac disease to assist high-risk patients in making lifestyle adjustments that will reducing rising number of patients and eventually the mortality rate. The medical community and patients can both greatly be benefitted from the usage of suitable technology support in this field. The use of machine learning algorithms to process raw health data from the heart will help save the lives of cardiac patients.

The proposed work demonstrates the effective use of seven different Machine Learning algorithms for the prediction of Heart Disease. It demonstrates a comparative study of algorithms with the higher amount of accuracy suggesting the algorithm which is having high accuracy is the best algorithm which could be implemented for predicting the Heart disease. With an accuracy rate of 87.77 percent, the Random Forest is the most accurate among the Statistics, 2012.

## REFERENCES

**[1]** Ramalingam, V. V., Ayantan Dandapath, and M. Karthik Raja. "Heart disease prediction using machine learning techniques: a survey." International Journal of Engineering & Technology 7.2.8 (2018): 684-687.

**[2]** Maiga, Jaouja, and Gilbert Gutabaga Hungilo. "Comparison of machine learning models in prediction of cardiovascular disease using health record data." 2019 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS). IEEE, 2019.

**[3]** Rairikar, Abhishek, *et al*. "Heart disease prediction using data mining techniques." 2017 International conference on intelligent computing and control (I2C2). IEEE, 2017.

**[4]** Dwivedi, Ashok Kumar. "Performance evaluation of different machine learning techniques for prediction of heart disease." Neural Computing and Applications 29.10 (2018): 685-693.

**[5]** Maji, Srabanti, and Srishti Arora. "Decision tree algorithms for prediction of heart disease". Information and communication technology for competitive strategies. Springer, Singapore, 2019. 447-454.

**[6]** Kumar, N. Komal, *et al*. "Analysis and prediction of cardio vascular diseaseusing machine learning classifiers." 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS). IEEE, 2020.

**[7]** Princy, R. Jane Preetha, *et al*. "Prediction of cardiac disease using supervised machine learning algorithms." 2020 4th international conference on intelligent computing and control systems (ICICCS). IEEE, 2020.

**[8]** https://www.kaggle.com/andrewmvd/heart-failure-clinical-data/discussion/193109

**[9]** Rennie, Jason & Shih, Lawrence & Teevan, Jaime & Karger, David. (2003). Tackling the Poor Assumptions of Naive Bayes Text Classifiers. Proceedings of the Twentieth International Confer-ence on Machine Learning. 41.

**[10]** Mucherino, Antonio, Petraq J. Papajorgji, and Panos M. Pardalos. "K-nearest neighbor classification". Data mining in agriculture. Springer, New York, NY, 2009. 83-106.

**[11]** https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

**[12]** Breiman, J. Friedman, R. Olshen, and C. Stone, Classification and Regression Trees (Statistics/ Probability Series). 1984.

**[13]** L. Breiman, ‘‘Random forests,’’ Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001

**[14]** A. Sharaff and H. Gupta, ‘‘Extra-tree classifier with metaheuristics approach for email classification,’’ in Proc. Adv. Comput. Commun. Comput. Sci. Singapore: Springer, 20

**[15]** https://www.datatechnotes.com/2020/07/classification-example-with-ridge-classifier-in-python.html