Prediction of Heart Disease Using Machine Learning Algorithms

Md. Julker Nayeem, Sohel Rana, and Md. Rabiul Islam

ABSTRACT

Heart disease has become one of the alarming issues of death. It is accountable for fatty plaques in the arteries. If this fatal condition can be identified early, we can preserve many people's arteries. Different types of supervised machine learning algorithms are applied in our research paper in order to predict heart disease existence in patient body. Besides this, we have focused on an efficient way to improve the performance of our applied classifiers. Imputing mean value technique is applied to handle null values present in our dataset. The features which are unnecessary are removed by using the info-gain feature selection technique. In order to calculate prediction accuracy, K-Nearest Neighbors (KNN), Naive Bayes and Random Forest are applied to the heart disease dataset. Accuracy, precision, recall, F1-score, and ROC are calculated which help us to compare the performance of the classification models. Handling null values on a particular column by imputing mean values of that column and our applied info-gain feature selection technique has aided us in improving the accuracy of our prediction models. Random Forest among all has given the best classification accuracy which is 95.63% with precision, recall, F1-score and ROC are 0.93, 0.92, 0.92 and 0.9, respectively.

Keywords: classification, heart disease, machine learning, supervised algorithms.

Published Online: November 30, 2022

ISSN: 2796-0072

DOI: 10.24018/ejai.2022.1.3.13

M. J. Nayeem*
Department of Computer Science & Engineering, Pundra University of Science & Technology, Bangladesh.
(e-mail: julkernayeemt72@gmail.com)

S. Rana
Department of Computer Science & Engineering, Bangladesh Army University of Science & Technology, Bangladesh.
(e-mail: sohelranacse052@gmail.com)

M. R. Islam
Department of Computer Science & Engineering, Pundra University of Science & Technology, Bangladesh.
(e-mail: mdrabiulislam521@gmail.com)

Corresponding Author

I. INTRODUCTION

Due to increasing the amount of data gradually in the medical industry, it has become difficult work to handle this huge amount of data as well as collect relevant information for accurate decision making. Therefore, at present, it has become highly demandable to apply a decent way that can provide acceptable decisions from a large number of databases.

With the help of data mining, which is a field of machine learning, we will be able to solve this problem properly. In order to solve real-world problems, it is a decent approach to find out hidden patterns and gather relevant information from a large dataset. At present, heart failure symptoms can be expressed at any age of a lifetime in the human body. The possibility to face this type of symptom is comparatively high for old people rather than the young age people. Data mining classification techniques can find previously unknown patterns and strongly linked characteristics that aid in predicting the class label from a huge dataset. With the help of those unseen relationships along with the highly correlated features, it has become easy to detect heart disease patients without any help from medical experts. Then, it will do as a system for separating patients with the presence of heart failure and patients with no heart failure more accurately with less diagnosis time.

Nevertheless, a variety of machine learning techniques are used to make predictions. Finding the ideal method is a difficult challenge for us. In our study, we used KNN, Naive Bayes, and Random Forest to predict whether or not early-stage cardiac disease is present in the patient's body.

There are mainly three contributions in our study. First, we have collected 65535 patient-specific real-world diagnostic datasets for heart disease from the Kaggle cardiovascular dataset. Second, by employing the info-gain feature selection strategy, we were able to identify the features that were not essential, helping us to enhance the performance of our classification model. Finally, we compared the performance of our three methodologies, compared the performance result with the findings of earlier research, and evaluated the prediction outcome based on various risk variables.

Our paper is broken into several sections. Section II presents a literature review. Section III presents the methodology. Results of experiments are presented in Section IV. Section V has finally concentrated on our paper's conclusion.

II. LITERATURE REVIEW

The identification of heart disease and many other major diseases has been the subject of several studies utilizing various data mining techniques. In order to predict coronary heart disease, Yilmaz *et al.* [1] used a variety of machine learning techniques. In comparison to other models, Random Forest (RF) has provided the greatest accuracy, with a score of 92.90%. Heart disease was predicted by Pal *et al*. [2] using the Random Forest (RF) method. The RF-generated accuracy is 86.90%. Heart disease was predicted by Boukhatem *et al*. [3] using several machine learning techniques. When compared to other methods, Support Vector Machine (SVM) has the highest accuracy (91.67%). Using the SVM, Decision Tree (DT), Naive Bayes (NB), K-Nearest Neighbor (KNN) Artificial Neural Network (ANN), etc., Riyaz *et al*. [4] predicted cardiac disease. The ANN provides the best accuracy of 86.91%. Rahman [5] diagnosed heart disease by using the KNN, XgBoost, Logistic Regression (LR), Support Vector Machine (SVM), Ada Boost, Decision Tree (DT), Naive Bayes (NB) and Random Forest (RF) algorithms. In comparison to other algorithms, the DT and RF method has provided the best accuracy at 99%. Heart disease was predicted by Riyaz *et al*. [6] using several machine learning techniques. Gradient Boosting has produced the best accuracy, which is 84.82%. To predict cardiac disease, Jindal *et al*. [7] evaluated various methods.

The KNN model has the highest accuracy (88.52%), compared to other kinds of prediction models. Rajdhan *et al*. [8] predicted heart disease using machine learning algorithms. They achieved the highest accuracy from Random Forest (RF) and is 90.16%. Shah *et al*. [9] predicted heart disease using machine learning algorithms. They achieved the highest accuracy from the KNN model which is 90.789%. Singh *et al*. [10] predicted heart disease from different machine learning algorithms. They obtained an accuracy of 87%, which is the maximum the KNN model could provide. In order to predict cardiac disease, Hasan *et al*. [11] examined several supervised machine learning classification algorithms. A feature selection strategy is used in this study info-gain to increase the classification model's accuracy. Logistic regression has produced the best accuracy, which is 92.76%. Using MLP and SVM, Nahiduzzaman *et al*. [12] predicted cardiac disease. They obtained a two-class classification accuracy from SVM of 92.45%, which is the highest. Nayeem *et al*. [13] used machine learning methods to forecast hepatitis disease. They obtained a Random Forest accuracy of 92.41%, which is the highest.

III. Methodologies

A. K-Nearest Neighbors (KNN)

Three phases are used in this classifier's classification process. The K-value is calculated in step 1. Step 2 computes the distance between all of the training data and ranks it for each test sample. Step 3 will supply the class name to the test sample data by using the majority vote technique [11]. Calculating the Euclidean distance involves:

$$E_d = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \qquad (1)$$

B. Naive Bayes

This method, which is used for classification purposes, is based on the Bayes theorem. It is simple to construct this classification model. With the use of Bayes' Theorem, we may determine the likelihood of an event occurring given the likelihood of an earlier occurrence. Calculating the posterior probability entails:

$$P\,(X|Y) = \frac{P\,(Y|X)P(X)}{P(Y)} \qquad (2)$$

C. Random Forest

This supervised machine learning approach is well-liked for both regression and classification tasks. It has been employed for classification in our research. It functions in three stages. A forest of Decision Trees is created from a number of trees in step 1 of the learning process. In step 2, a class name is predicted for each test set using the trees created in step 1's forest. The test data is given the appropriate class name in step 3, which is the last stage, depending on the results of the majority of votes. Every piece of data in the dataset is subjected to step 3 [11].

D. Working Procedure

We have implemented all steps in the python environment. The following is a list of the essential stages for our research's working method:

- ✓ Step-1: A file with the name dataset heart disease should initially be created after gathering the dataset from the UCI machine learning repository. file type is.csv (comma-separated value).
- ✓ Step-2: Checking the presence of null values in each column.
- ✓ Step-3: Handle null values (if any) by using the mean imputation technique.
- ✓ Step-4: Then, create a new CSV file and give it the name dataset heart update after checking highly correlated features using the info gain feature selection approach to identify highly correlated features.
- ✓ Step-5: In order to determine whether or not heart disease is present, load the dataset heart.csv file (containing all attributes and null values). The dataset will then be classified using our three classifiers.
- ✓ Step-6: In order to determine whether or not heart disease is present, load the dataset heart update.csv (which only contains strongly correlated features and excludes observations with null values). The dataset will then be classified using our three classifiers.
- ✓ Step-7: Finally, contrast the classification model performance obtained in steps 5 and 6 using dataset_heart.csv and dataset_heart_update.csv, respectively.
- ✓ Step-8: We contrast the accuracy of our classification model with that of earlier studies.

The flow chart of our working process has shown in Fig.1 where we have focused all of the necessary steps mentioned in above.

IV. Experimentation

The dataset for this study was obtained from Kaggle [14]. There are 65535 records and 13 total features in our collection. Our experiment has been divided into two phases.

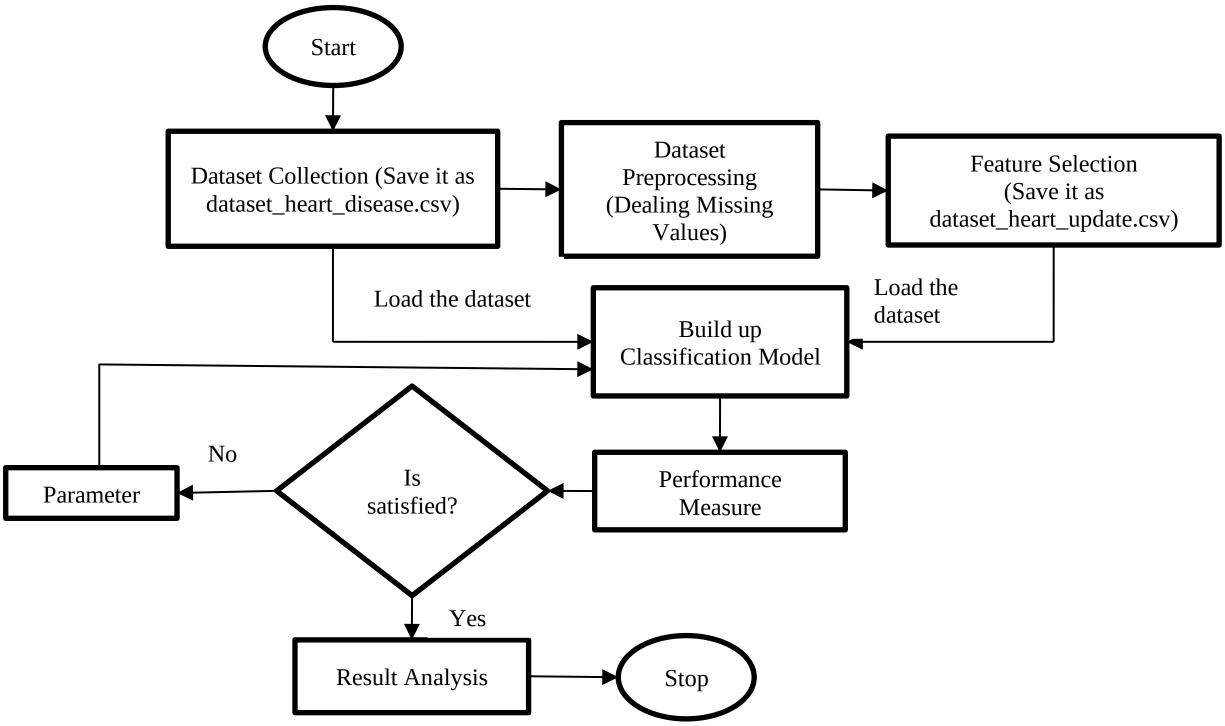
Start
Dataset Collection (Save it as dataset_heart_disease.csv)
Dataset Preprocessing (Dealing Missing Values)
Feature Selection (Save it as dataset_heart_update.csv)
Load the dataset
Load the dataset
Build up Classification Model
No
Parameter
Is satisfied?
Performance Measure
Yes
Result Analysis
Stop

Fig 1. Flow chart of the work process

A. Data Preprocessing

Our dataset has 13 features in total. ID, age, gender, height, weight, ap hi, ap lo, cholesterol, gluc, smoking, alcohol, active, and cardio are some of them. In our dataset, some null values are present. There are several methods available to deal with null values. In our research, we initially searched for records with null values in order to fill them with data using the mean imputation approach. The accuracy of classification can occasionally be negatively impacted by records with null values. Null values might make classification more inaccurate. Steps 2 and 3 of our working process are previously covered in the methodology section's section on how to handle records with null values. In addition, the dataset's features that are highly correlated are identified by using the info-gain attribute selection approach. Low correlation between attributes might make the prediction model less accurate. We have employed feature selection approaches as a result. Out of the 13 qualities, 10 were chosen utilizing this info gain feature selection approach that are highly connected, including age, gender, height, weight, ap hi, ap lo, cholesterol, gluc, active, and cardio. Step 4 of the methodology section's working procedure already discusses the strategy for identifying strongly correlated features.

B. Problem Statement

Due to their objectives are comparable to ours, we used Yilmaz *et al*. [1], Pal *et al*. [2], and Rajdhan *et al*. [8] as our foundation articles in this study. The authors [1], [2] and [8] have used all the attributes present in the dataset and achieve good accuracy from their classification models but they have used small size dataset to compare classification model accuracy as well as did not handle null values present in the dataset. In addition, they did not employ any method of feature selection to identify strongly correlated features that can enhance classification model accuracy.

C. Result and Discussion

Using Jupyter Notebook, three supervised classification methods were implemented (Anaconda 3). To evaluate the effectiveness of the model, we used 10-fold cross-validation and used 70% of the data for training and 30% for testing. Our classification models have been applied to two criteria. First, we utilized the 13 attributes from our dataset_heart_disease.CSV file that had null values at all, and then we used a subset of 10 attributes from dataset_heart_update.CSV. Steps 2, 3, and 4 of our working procedure, which is part of the methodology section, already describe how we obtained the two CSV-formatted files dataset_heart and dataset_heart_update. In our research, K=10 is the ideal number for K-Nearest Neighbors. In our research, numTrees=100 is the ideal number for Random Forest. Table I and Table II, which use our two datasets, show the outcomes for our three classifiers.

We can see from our performance comparison Tables I and II that by employing 10 features rather than 13, we were able to improve the accuracy of our classification models. The mean imputation approach is used to fill in the blanks in records, which improves the performance of our classification models.

TABLE I: THE OUTCOME OF 13-FEATURE CLASSIFICATION MODELS EXISTENCE OF OBSERVATIONS WITH NULL VALUES (DATASET_HEART_DISEASE)

Name of Classification Algorithm	Confusion Matrix		Accuracy

KNN	TP=1641
	FP=259
		FN=300
		TN=361
			86.36%

Naive Bayes	TP=1631
	FP=269
		FN=320
		TN=341
			85.84%

Random Forest	TP=1728
	FP=172
		FN=311
		TN=350
			90.94%

DOI : http://dx.doi.org/10.24018/ejai.2022.1.3.13

TABLE II: THE OUTCOME OF TEN-FEATURE CLASSIFICATION MODELS OBSERVATIONS NOT HAVING NULL VALUES (DATASET_HEART_UPDATE)

Name of Classification Algorithm	Confusion Matrix		Accuracy
KNN	TP=1660	FN=300	87.36%
	FP=240	TN=361	
Naive Bayes	TP=1689	FN=314	88.89%
	FP=211	TN=347	
	FP=13	TN=19	
Random Forest	TP=1817	FN=211	95.63%
	FP=83	TN=450	

When comparing the three classification models, we can see that the Random Forest approach consistently outperforms the other two algorithms in all three instances. Fig. 2 illustrates the precision of our categorization models visually.

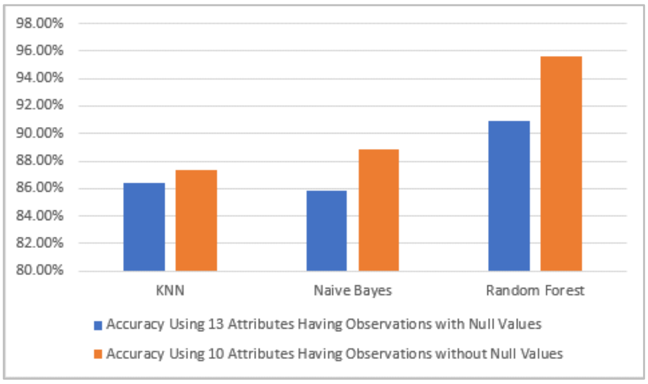

Fig. 2. Bar graph showing our classifiers' accuracy.

TABLE III: THE CLASSIFIER PERFORMANCE REPORT

Name of the Classification Algorithm	Precision	Recall	F1-Score	ROC Area
KNN (with 13 features and null values)	0.85	0.84	0.84	0.8
KNN (with 10 features having and without null values)	0.87	0.85	0.86	0.8
Naive Bayes (with 13 features and null values)	0.85	0.84	0.84	0.85
Naive Bayes (with 10 features and without null values)	0.89	0.88	0.88	0.8
Random Forest (with 13 features and null values)	0.9	0.9	0.91	0.9
Random Forest (with 10 features and without null values)	0.93	0.92	0.92	0.9

According to Table III, 10 of the attributes in dataset_heart_update.csv performed better for all of our classification models than the 13 attributes in dataset heart disease.csv. It is not possible to estimate cardiac condition accurately using all of the features in our dataset. As a result, we used the information-gain attribute selection approach to identify and then eliminate unnecessary features. We have improved the performance of our classification models in this way. When comparing the three classification models, we can see that the accuracy of Random Forest algorithm is 95.63% which is higher than the accuracy of the other two classification techniques. In addition, if we have contrasted the results with past research, as was done in part IV of the problem statement portion, we attained the best accuracy using Random Forest, which is greater than [1] and is 95.63%. Yilmaz *et al*. [1] obtained the best accuracy from Random Forest, which is 92.90%. We attained the best accuracy from Random Forest, which is better than [2] and is 95.63%. Pal *et al*. [2] obtained the best accuracy from Random Forest, which is 86.90%. The best accuracy from Random Forest obtained by Rajdhan *et al*. [8] was 90.16%, but the best accuracy from Random Forest that we have obtained 95.63%, which is greater than [8]. Besides this, we can observe that though our dataset is comparatively large than [1], [2] and [8] in number of records, we have succeeded to keep our model accuracy comparatively high than [1], [2] and [8].

V. CONCLUSION AND FUTURE WORKS

We have demonstrated in this study how managing null values and feature selection are crucial for increasing classification model accuracy. We have compared our classification models in an effort to find the best classifier. Each of our classification methods performs admirably when handling observations from the dataset that have null values using mean imputation and using the info gain feature selection strategy to our dataset. The dataset's lower contribution features and null values might be to blame for the poor classification accuracy. The Random Forest classifier performed the best out of our three classifiers, scoring 95.63% overall with accuracy, recall, F1-score, and ROC values of 0.93, 0.92, 0.92, and 0.9, respectively. It is advised to utilize different classification algorithms in the future that employ better feature selection methods.

REFERENCES

[1] Yilmaz R, Yagin FH. Early detection of coronary heart disease based on machine learning methods. *International Medical Journal*. 2022 Jan 1; 4(1): 1–6. doi: 10.37990/medr.1011924.

[2] Pal M, Parija S. Prediction of heart diseases using random forest. *Journal of Physics: Conference Series*. 2021 Mar 15; 1817(1): 1–9. doi: 10.1088/1742-6596/1817/1/012009.

[3] Boukhatem C, Youssef HY, Nassif AB. Heart disease prediction using machine learning. *IEEE Advances in Science and Engineering Technology International Conferences (ASET)*. 2022 Feb 21–24, Dubai, United Arab Emirates.

[4] Riyaz L, Butt MA, Zaman M, Ayob O. Heart disease prediction using machine learning techniques: a quantitative review. *International Conference on Innovative Computing and Communications*, pp. 81–94, vol. 1394, Singapore: Springer; 2022.

[5] Rahman MM, Rana MR, Alam MNA, Khan MSI, Uddin KMM. A web-based heart disease prediction system using machine learning algorithms. *Network Biology*. 2022 Jun 1; 12(2): 64–80.

[6] Riyaz L, Butt MA, Zaman M. Improving coronary heart disease prediction by outlier elimination. *Applied Computer Science*. 2022 Mar 28; 18(1): 70–88. doi: 10.35784/acs-2022-6.

[7] Jindal H, Agrawal S, Khera R, Jain R, Nagrath P. Heart disease prediction using machine learning algorithms. *IOP conference series: materials science and engineering*. 2021 Jan 18; 1022(1): 1–11. doi: 10.1088/1757-899X/1022/1/012072.

[8] Rajdhan A, Sai M, Agarwal A, Ravi D, Ghuli DP. Heart disease prediction using machine learning. *International Journal of Research and Technology*. 2020; 9(4): 659–662.

[9] Shah D, Patel S, Bharti SK. Heart disease prediction using machine learning Techniques. *SN Computer Science*. 2020 Oct 16; 1(6): 1–6. doi: https://doi.org/10.1007/s42979-020-00365-y.

[10] Sing A, Kumar R. Heart disease prediction using machine learning algorithms. *IEEE international conference on electrical and electronics engineering (ICE3)*. 2020 Feb 14–15, pp. 452–457, Gorakhpur, India.

[11] Hasan SMM, Mamun MA, Uddin MP, Hossain MA. Comparative analysis of classification approaches for heart disease prediction. *IEEE International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)*. 2018 Feb 8–9, Rajshahi, Bangladesh.

[12] Nahiduzzaman M, Nayeem MJ, Ahmed MT, Zaman MSU. Prediction of heart disease using multi-layer perceptron neural network and support vector machine. *IEEE 4th International conference on electrical information and communication technology (EICT)*. 2019 Dec 20-22, Khulna, Bangladesh.

[13] Nayeem MJ, Rana S, Alam F, Rahman MA. Prediction of hepatitis disease using k-nearest neighbors, naive bayes, support vector machine, multi-layer perceptron and random forest. *IEEE International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*. 2021 Feb 27–28, pp. 280–284, Dhaka, Bangladesh.

[14] Kaggle.com. *Kaggle Cardiovascular Disease Dataset*. [Internet]. 2019 [updated 2019 Jan 20; cited 2022 Sep 01]; Available from: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset.

Md. Julker Nayeem is currently pursuing his M.Sc. Engineering degree in Computer Science & Engineering (CSE) from the Department of Computer Science & Engineering, Rajshahi University of Engineering Technology (RUET), Bangladesh. He obtained his B.Sc. Engineering degree in CSE from the Department of Computer Science & Engineering, Pabna University of Science & Technology (PUST), Pabna, Bangladesh in 2017. He is currently working as Lecturer in Department of Computer Science & Engineering, Pundra University of Science & Technology, Bogura, Bangladesh since 12th July 2019 to Date. His research focuses on data mining and machine learning.

Sohel Rana received his M. Sc. Engineering degree in CSE from the Department of Computer Science & Engineering, Rajshahi University of Engineering & Technology (RUET), Bangladesh in 2021. He obtained his B. Sc. Engineering degree in CSE from the Department of Computer Science & Engineering, Bangabandhu Sheikh Mujibur Rahman Science & Technology University, Gopalganj, Bangladesh in 2016. He is currently working as Assistant Professor in Department of Computer Science & Engineering, Bangladesh Army University of Science & Technology, Saidpur, Bangladesh. His research focuses on data mining, pattern recognition and machines learning. He has over five years of teaching and research experience.

Md. Rabiul Islam obtained his B. Sc. Engineering degree in CSE from the Department of Computer Science & Engineering, Rajshahi University, Bangladesh in 2021. He is currently working as Lecturer in Department of Computer Science & Engineering, Pundra University of Science & Technology, Bogura, Bangladesh since 01st January, 2022 to Date. His research focuses on data mining and machine learning.

DOI : http://dx.doi.org/10.24018/ejai.2022.1.3.13