Prognosis of Cardiovascular Disease Using Machine Learning Procedures

Md. Shahiduzzaman
Department of Computer Science and Engineering
Bangladesh University of Business and Technology (BUBT)
Dhaka, Bangladesh
shahid@bubt.edu.bd

Nowreen Haque Biswas
Department of Computer Science and Engineering
Bangladesh University of Business and Technology (BUBT)
Dhaka, Bangladesh
nawreenhoque@gmail.com

Md. Momin
Department of Computer Science and Engineering
Bangladesh University of Business and Technology (BUBT)
Dhaka, Bangladesh
mdmuminjap@gmail.com

Raihan Sikdar
Department of Computer Science and Engineering
Bangladesh University of Business and Technology (BUBT)
Dhaka, Bangladesh
raihansikdar10@gmail.com

2022 International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE) | 978-1-6654-6944-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICAEEE54957.2022.9836418

***Abstract*—The topmost crucial muscular body part is our heart, as it pumps blood to all the other organs in our body. Being the essential organ, mortals suffering from heart disease over the last twenty years has been the deadliest disease globally and top number one destroyer for human. Over the recent years, the health industry and technology have worked together to find ways to cut back the risk of cardiac diseases in humans. For early disease prediction, machine learning is a necessity for healthcare as it functions without human interaction. In this paper, a cardiovascular data set with 70,000 data and 12 attributes are analyzed and implemented for the early prognosis of cardiovascular disease. Using the voting ensemble classifier, we combined five different machine learning algorithms to achieve good overall accuracy. K-nearest neighbor classifier gained an accuracy of 75%, which was the best amongst Logistic Regression, Random Forest, Gradient Boosting, and Bernoulli Naive Bayes. This proposal benefits and eases the work for clinicians and doctors and provides appropriate care for heart disease patients.**

***Index Terms*—Cardiovascular disease; Voting ensemble method;Machine learning classifier;**

I. Introduction

Cardiovascular disease became a major issue in the health industry as it set off being the prime reason for fatality worldwide. CVDs (Cardiovascular Diseases) occur due to a bunch of disorders in the heart and clotting of the blood vessels. As claimed by, World Health Organization (WHO) 2019, approximately 17.9 million people died from CVDs, of which 85% of people suffered and died from heart attack and stroke, representing 32% of deaths globally [1]. Like all South Asian countries, Bangladesh is excessively prone to CVDs. Of these mortality rates, 15.23 % died from CVDs in Bangladesh, according to WHO published reports in 2018, making it the chief cause of death [2]. Over the years, based on the number of patients with CVDs, the mortality rate has increased accordingly due to the unhealthy lifestyle of people. There are other possible behavioral factors such as smoking, unhealthy diet, obesity, lack of exercise, alcohol, insomnia, rapid weight gain, which could increase the risk of CVDs. So it became essential to ensure that our cardiovascular system or any other systems in the human body related to diseases must remain healthy [3]. As Prevention is better than cure, maintaining a healthy lifestyle is necessary to lead a happy, healthy, and long life. Early prediction for cardiac diseases is essential with good precision for quick diagnosis, or else a slight mistake can lead to the demise of a person. So in order, to get the correct prognosis, medical advice, tests, and medicines should proceed on short notice, or else it may cause misdiagnosis due to medical tests delay. Therefore, computers may find it hard to complete, quite expensive, and time-consuming for evaluations to carry out [4].

On the other hand, technology has evolved, and artificial intelligence has made advancements in health services. Therefore, to provide a computerized prediction for cardiovascular disease in patients, computer-aided detection (CAD) is being introduced, as examining heart disease is a demanding task. Machine learning is the present time's computer-aided detection method and also popped up technology for analyzing medical information and providing accurate prognosis. [5]. Machine learning appears from pattern recognition, a subfield of AI where computers can learn to perform skills without being programmed can independently adapt when models or different algorithms, mostly being exposed to new data. We used a voting ensemble classifier with multiple machine learning techniques to diagnose the disease and classify or predict the output, guiding them with a nutritional guideline based on their existing symptoms under the doctor's supervision to stop the disease from becoming lethal.

The remains of our study paper are categorized as follows: Second section outlines the related work. In the third section, methodology elaborates about data set description, preprocessing, analysis and modeling. The fourth section includes the experimental setup, which describes the performance for different algorithms. After that, results defines in the fifth

section, including an accurate comparison of the algorithms, and at last, the paper winds in Section VI.

II. Related Works

A vast amount of research was done, linked to coronary disease using different machine learning approaches, and got accuracy and prognosis based on their techniques. Instead of working with a single model, the work completed in [6] attained an accuracy of 90% by applying a voting ensemble classifier using multiple machine learning algorithms. For early detection of heart disease, a hybrid model built by merging the hybrid characteristics of random forest and linear model(HRFLM) into a single machine learning model, where Mohan et al. [7] achieved an accuracy of 88.7%. Using neural networks with an artificial neural network approach, Allahverdi and Kahramanli. [8] has developed a hybrid model for detecting heart disease and gained an accuracy of 82.4%. Using the UCI repository dataset, Rohit et al. [9] beat a score of 94.2% with K-nearest neighbor classifier by coining machine learning and deep learning models, D. Shah, S. Patel, and S. K. Bharti [10] applied data mining techniques pulling off the accuracy of 90%. Therefore, A. Singh and R. Kumar [11] used multiple machine learning algorithms where the accuracy of 87% attained by K-nearest neighbor concluding normalizing the data set before training gives better accuracy. After analyzing various published research papers about coronary illness, the ordering is similar in every research to identify the risk of heart disease. Moreover, the UCI repository is the frequent data set be seen in recent papers, using more or less the same machine learning algorithms with different techniques. Detrano et al. [12] used Cleveland dataset to predict cardiovasular disease and they were one of the first researchers to use this dataset. To reduce the complexity of the machine learning models and for faster training, feature selection is necessary. Yar et al. [13] applied four feature selection techniques and ten classification algorithms from which Gaussian and ET had higher accuracy despite using feature selection procedures, ET had a slight increase from 92.09 to 94.41 percent whereas, the score of GB also increased from 91.34 to 93.36 percent.

Evaluations of multi-level risk of heart diseases proceeded by Aljaaf et al. [14] attained an accuracy of 86.53% using decision tree algorithm.

Many papers working with smaller data sets achieve a high accuracy rate. In some cases, a smaller dataset causes overfitting with deep learning models. Therefore, working with recent larger coronary datasets causes results and accuracies to be more realistic than working with smaller data sets. On the other hand, working with a large data set, an opportunity to learn about the new updates and vast information within developing countries such as Bangladesh, being the least studied for research and helping them treat accordingly. Many researchers are appreciated highly for their work, such as Nowshad et al. [15] who worked with multiple machine learning models, created their dataset of 564 instances and 18 attributes based on the Sylhet region of Bangladesh pulled a score of 91.7% with support vector machine algorithm. K. C. Howlader et al. [16] physically collected data from 3 different hospitals in Bangladesh, and worked with seven algorithms, and concluded that Naive Bayes algorithms performed better with an accuracy of 70.83 percent using their proposed model. With higher proportions of data, PCA can store important information in a new piece, applying various algorithms could set aside. Ratnasari et al. [17] used 150 gray-level thresholds based on principal component analysis (PCA) and ROI to minimize features of the X-ray image. Parthiban et al. [18] conclude that applying Naive Bayes and SVM algorithms withdraws key attributes to detect diabetes as they are the chief cause of heart disease. Krishnan et al. [19] work opens the door to further investigation and concludes that compared to Naive Bayes, the decision tree has a higher accuracy rate in their study. In Sapra et al. [20] paper, they trained with six classifiers and gradient boosting trees achieved the best accuracy of 84 % using Cleveland heart disease and Z-Alizadesh Sani dataset to detect the presence of cardiac diseases.

Due to physical, emotional, and mental behavior, males are more vulnerable to heart disease than females as they are less adaptive. Based on our ongoing research work in clinical data analysis, the prognosis of heart disease became a typical issue as death rates due to CVDs increased rapidly over the years.

III. Methodology

As the main goal of this research is to anticipate the odds of having cardiac disease by automated prognosis which makes it beneficial for health industry as the doctors can diagnosis them early and for patients it will spare the expense of different tests which will not be required to expect the outcome. Well,to fulfil the objective,we have argued about the use of various machine learning techniques by the help of dataset which is mentioned on the paper and propose a model based on the best accuracy.

A. Illustration of the Dataset

The main objective of our research study is to anticipate the odds of having the cardiac disease using an automated prognosis, making it beneficial for the health industry. Doctors can diagnose them early, and for patients, it will spare the expense of different tests which will not be required to expect the outcome. Well, to fulfill the aim with the help of the dataset, in our paper, we argued about the use of various classifier techniques. Therefore, a model is proposed based on the best accuracy.

B. Dataset Illustration

From Kaggle, we used a cardiovascular disease data set [21] for our research study. It consists of data of 70,000 patients with 12 attributes, with one being the predicted attribute. The ”cardio” attribute refers to the risk of cardiac disease in the patient. The values represents 0 = no disease and 1 = having disease. The remaining attributes are defined as follows, and each attribute resemble: label=

1)) age - Patient’s age in years.
2)) gender- sex of the person. (1=male; 0=female)

3)) height-Height of the person in cm.
4)) weight- Weight of the person in kgs.
5)) ap_high-Systolic blood pressure (Each time it beats, the peak number, our heart calculates the force exerted on the linings of the arteries).
6)) ap_lo-Diastolic blood pressure (Between heartbeat, the bottom number, the heart measures the pressure on the walls of our arteries).
7)) cholesterol-Cholesterol is a wax-like substance present in our blood.
8)) gluc-Glucose travels through the bloodstream to all of our body's cells to use for energy and storage. (gluc-1: normal,2: prediabetic,3:diabetic)
9)) smoke - Do they smoke or not? Values are in binary
10)) alco -Consumption of alcohol.
11)) the active- Physical activity of a person.
12)) cardio-The target variable which defines the presence or absence of cardiovascular disease.

C. Pre processing of Data

For any data set, data preprocessing is necessary as it transforms the raw data into a more understandable and efficient format. Moreover, it may contain missing, noisy, or duplicate values, which can be reduced by data preprocessing, and will increase the accuracy score. For this existing dataset, there are no null or duplicate values but may contain some outliers, and applying the data preprocessing techniques model will also train faster. We will be using two data preprocessing techniques. They are feature selection and feature scaling.

1) Feature Scaling: Feature scaling is a method used to alter the scale of independent features of data. It prevents the machine learning algorithm from getting stuck in local minima. It gives a better error surface shape.

2) Feature Selection: While developing a predictive model, feature selection reduces the nonimportant attributes to give a good accuracy score, which reduces the overall training time and overfitting of the model.

D. Proposed Model

After data set analysis of the attributes, we designed a framework of our model using five different machine learning algorithms with a voting ensemble method for prediction. We divided the dataset into testing and training segments in the ratio of 60/40. This paper outlines the voting ensemble classifier with Bernoulli Naive Bayes, Random Forest, Gradient Boosting, Logistic Regression, and K-nearest neighbor classification model, as shown in Fig 1.

E. Voting Ensemble Classifier

A machine learning model that trains on an ensemble several algorithms and predicts its own accuracy based on their highest probability of chosen class as the output which is often very effective.It can be also used for classification or regression problems. Voting works by creating two or more sub-models. Each sub-model makes predictions which are combined in some way, such as by taking the mean or the mode of the predictions, allowing each sub-model to vote on what the outcome should be. [22]

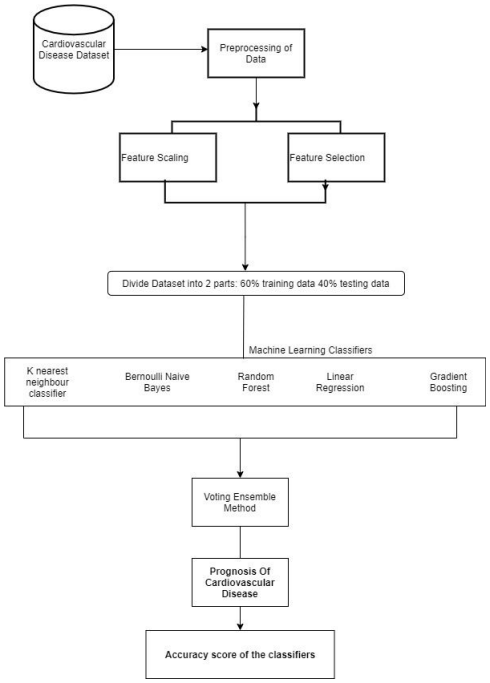

Fig. 1. Proposed model for prognosis of cardiovascular disease.

RESULT ANALYSIS

F. Baseline Dataset

In this study, Heart Disease UCI dataset acts as our baseline dataset.It contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. The "goal" field refers to the presence of heart disease in the patient. Many researchers work with UCI dataset and vast amount of information is gathered related in predicting cardiovascular disease. The table shown below in I illustrates working with this dataset by applying machine learning algorithms and obtained accuracy given as follows:

TABLE I
HEART DISEASE UCI DATASET RESULT SCORE

Alorgithm	Accuracy Score
K-nearest neighbour	**83%**
Random Forest	**77%**
Bernoulli Naive Bayes	**81%**
Gradient boosting classifier	**100%**
Logistic Regression	**81%**
Soft Voting Classifier	**79%**
Hard Voting Classifier	**81%**

In this dataset,gradient boosting algorithm gained an accuracy of 100%.

We compared our analyzed result with mostafa et al [12] paper and concluded that with gradient boosting our result scored an accuracy of 100% .

TABLE II
HEART DISEASE UCI DATASET RESULT SCORE OF MOSTAFA ET AL.

Alorgithm	Accuracy Score
K-nearest neighbour	**97.03%**
Random Forest	**96.70%**
Naive Bayes	**96.04%**
Gradient boosting classifier	**94.06%**
SVM	**97.03%**
Bagging	**92.08%**
ADA Boosting	**91.42%**

G. Collected Dataset

A cardiovascular disease data set from Kaggle [20] is used for this research study.It consists data of 70,000 patients with 12 attributes with one being the predicted attribute.We will be comparing their accuracies as shown based on 2 datasets in table I and table III and concluded that K-nearest neigbour obtained an accuracy of 75% is a good algorithm for predicting cardiovascular disease for cardiovascular disease dataset.Whereas for UCI dataset gradient boosting model obtained an accuracy of 100%.as it has a higher accuracy than other ML approaches.Well data visualization is needed to show the accuracies in order to make the data more understandable.

TABLE III
CADIOVASCULAR DISEASE DATASET RESULT SCORE

Alorgithm	Accuracy Score
K-nearest neighbour	**75%**
Random Forest	**72%**
Bernoulli Naive Bayes	**71%**
Gradient boosting classifier	**74%**
Logistic Regression	**71%**
Soft Voting Classifier	**73%**
Hard Voting Classifier	**71%**

H. Confusion Matrix

Also known as the error matrix, it outlines the performance results of machine learning algorithms. It shows how the classification model is confused when it makes predictions. To calculate the confusion matrix, we need to calculate the accuracy precision, recall, and f1 values, and below, in table IV four mixed combinations of predicted and actual values are illustrated.

- TP - True Positive (Total patients with heart diseases.)
- TN - True Negative (Total patients with heart diseases and no heart diseases.)
- FP - False Positive (Total patients with no heart diseases.)
- FN - False Negative (Total patients with no heart diseases and heart diseases.)

1) Accuracy: Accuracy is the ratio of all accurately predicted cases to the total observations. The model works best when accuracy is higher. In our case, the K-nearest neighbor has the best accuracy than the other algorithms. It has an accuracy of 75%. Figure 2 shows the graph with accuracy comparisons.

$$Accuracy = (TN + TP)/TP + FP + FN + TN \quad (1)$$

TABLE IV
CONFUSION MATRIX

	Actual Values		
3x Predicted Values		Positive	Negative
	Positive	TP	FP
	Negative	FN	TN

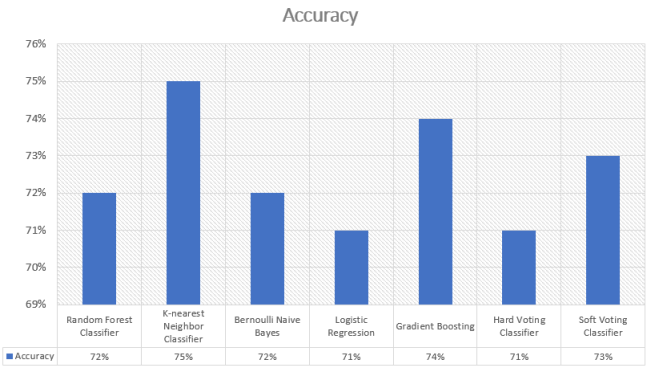

Fig. 2. Accuracy graph

2) Precision: Its represents a ratio of true positive cases from all the positive predictions. Figure 3 shows the precision graph of algorithms.

$$Precision = TP/(TP + FP) \quad (2)$$

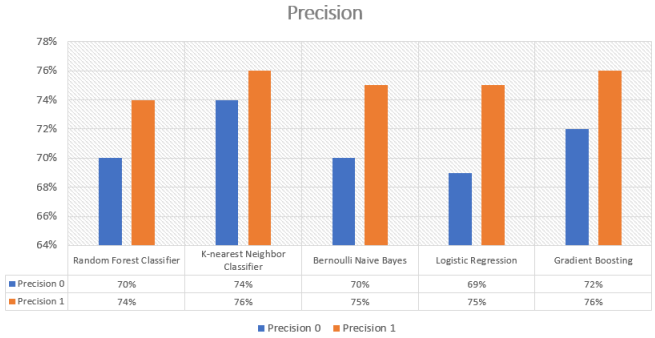

Fig. 3. Precision graph

3) Recall: It illustrates the true positive cases to the total observations in actual class. Figure 3 shows the recall graph of algorithms.

$$Recall = TP/TP + FN \quad (3)$$

4) F1 score: F1 is the average of precision and recall and it takes all the false positive and false negative in count. Moreover its more useful than accuracy. Figure 5 shows the F1 graph of algorithms.

$$F-score = (2\times(Recall\times Precision))/(Recall+Precision) \quad (4)$$

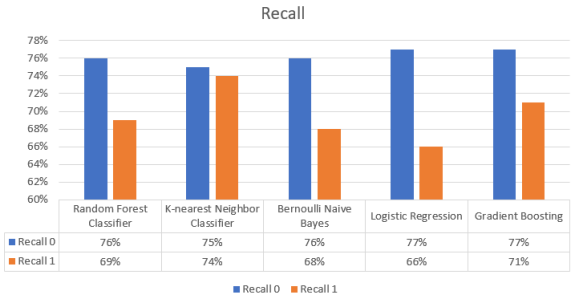

Fig. 4. Recall graph

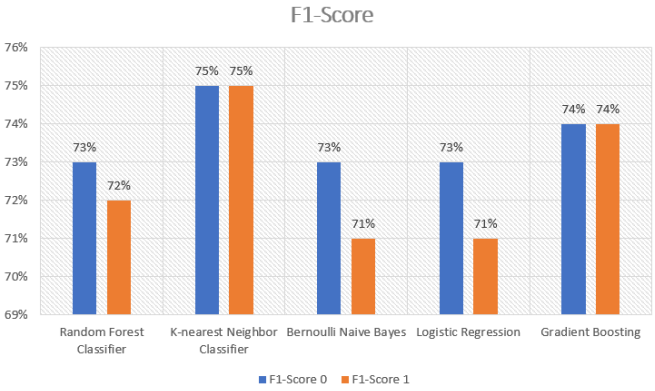

Fig. 5. F1-score graph.

5) ROC curve: The ROC curve shows the trade-off between sensitivity (or TPR) and specificity (1 – FPR). Classifiers that give curves closer to the top-left corner indicate a better performance. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

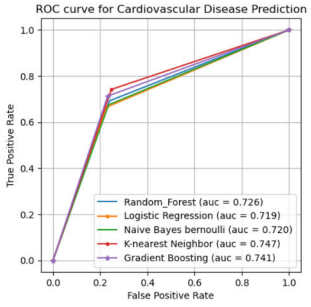

Fig. 6. ROC Curve for cardiovascular disease dataset

OBSERVATIONS

I. Correlation Matrix

From the correlation matrix shown above in fig 7, we can visualize that age,cholesterol,weight,glucose are significantly correlated to the target feature.The green color indicates positive correlation and red indicates a negative correlation.

In fig 8, it shows that from age 55-60 risks of having cardiovascular diseases is at its peak.

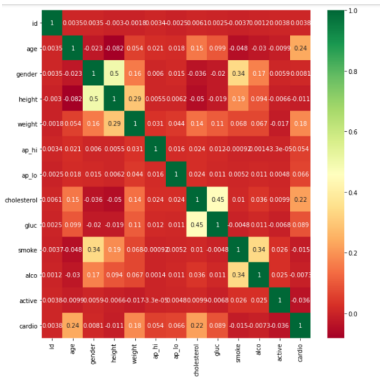

Fig. 7. Correlation for all features in the data set

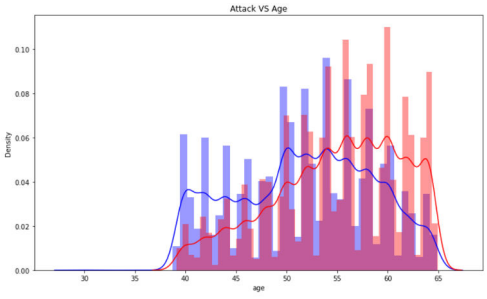

Fig. 8. Attack vs Age for prognosis of CVDs

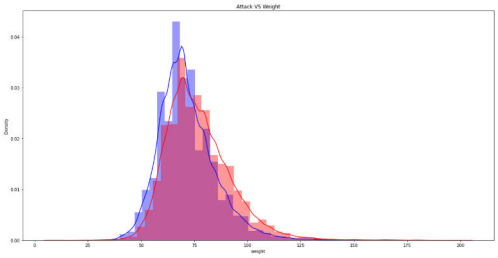

Fig. 9. Attack vs Weight for prognosis of CVDs

From fig 9, it shows that from weight 60-75 kgs having cardiovascular disease is high based on the observations of cardiovascular disease dataset.

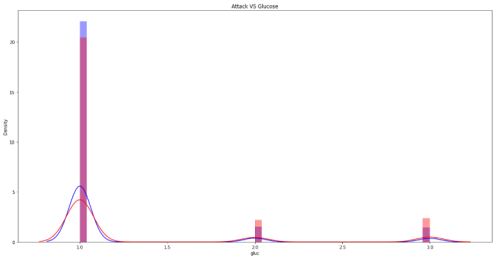

Fig. 10. Attack vs Glucose for prognosis of CVDs

From fig 10, it shows that from glucose on a normal level the risk having cardiovascular disease is more based on the observations of cardiovascular disease dataset.

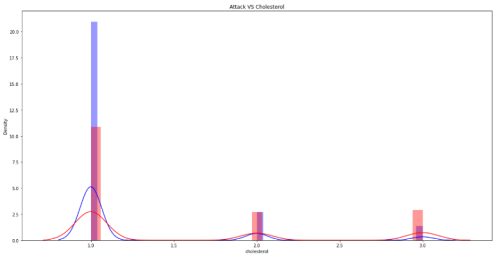

Fig. 11. Attack vs Cholesterol for prognosis of CVDs

From fig 11, it shows that from high cholesterol increases the risk of having cardiovascular disease based on the observations of cardiovascular disease dataset.

IV. Conclusion

Cardiovascular diseases can lessen down death rates if they are diagnosed early with the appropriate precautions. In this research study, with the help of the cardiovascular dataset, we could early detect the patients with cardiovascular disease. K-nearest neighbor had a good prediction accuracy of 75% out-performed the other machine learning algorithms. Therefore,we achieved less accuracy compared to the reviewed papers but worked with the large data set and got vast information about patients and realistic predictions output which will help them diagnose early instead of working with a smaller dataset. The patient with positive results of cardiac diseases can be immediately diagnosed under the supervision of the doctors and take care of all the preventative measures early.This prediction can work with other chronic diseases such as diabetes, cancer, etc. In the future, we will work by collecting all raw data from all the hospitals in Bangladesh and work on improving the accuracy.

References

[1] "World health organization," https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).2021.

[2] "Bangladesh: Coronary heart disease 2018," shorturl.at/hEHW8.

[3] A. Garg, B. Sharma, and R. Khan, "Heart disease prediction using machine learning techniques," in *IOP Conference Series: Materials Science and Engineering*, vol. 1022, no. 1. IOP Publishing, 2021, p. 012046.

[4] X.-Y. Gao, A. Amin Ali, H. Shaban Hassan, and E. M. Anwar, "Improving the accuracy for analyzing heart diseases prediction based on the ensemble method," *Complexity*, vol. 2021, 2021.

[5] K. H. Miao, J. H. Miao, and G. J. Miao, "Diagnosing coronary heart disease using ensemble machine learning," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 10, pp. 1–12, 2016.

[6] R. Atallah and A. Al-Mousa, "Heart disease detection using machine learning majority voting ensemble method," in *2019 2nd International Conference on new Trends in Computing Sciences (ICTCS)*, 2019, pp. 1–6.

[7] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81 542–81 554, 2019.

[8] H. Kahramanli and N. Allahverdi, "Design of a hybrid system for the diabetes and heart diseases," *Expert Syst. Appl.*, vol. 35, pp. 82–89, 07 2008.

[9] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, "Prediction of heart disease using a combination of machine learning and deep learning," *Computational Intelligence and Neuroscience*, vol. 2021, 2021.

[10] D. Shah, S. Patel, and S. K. Bharti, "Heart disease prediction using machine learning techniques," *SN Computer Science*, vol. 1, no. 6, pp. 1–6, 2020.

[11] A. Singh and R. Kumar, "Heart disease prediction using machine learning algorithms," in *2020 international conference on electrical and electronics engineering (ICE3)*. IEEE, 2020, pp. 452–457.

[12] D. R, J. A, P. M. Steinbrunn W, S. JJ, L. S. Sandhu S, Guppy KH, and F. V, "International application of a new probability algorithm for the diagnosis of coronary artery disease," in *The American Journal of Cardiology, vol. 64, no. 5*. PMCID, 1989, p. 304–310.

[13] Y. Muhammad, M. Tahir, M. Hayat, and K. T. Chong, "Early and accurate detection and diagnosis of heart disease using intelligent computational model," *Scientific reports*, vol. 10, no. 1, pp. 1–17, 2020.

[14] A. J. Aljaaf, D. Al-Jumeily, A. J. Hussain, T. Dawson, P. Fergus, and M. Al-Jumaily, "Predicting the likelihood of heart failure with a multi level risk assessment using decision tree," in *2015 Third International Conference on Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE)*, 2015, pp. 101–106.

[15] M. N. R. Chowdhury, E. Ahmed, M. A. D. Siddik, and A. U. Zaman, "Heart disease prognosis using machine learning classification techniques," in *2021 6th International Conference for Convergence in Technology (I2CT)*, 2021, pp. 1–6.

[16] K. C. Howlader, M. S. Satu, and A. Mazumder, "Performance analysis of different classification algorithms that predict heart disease severity in bangladesh," *International Journal of Computer Science and Information Security (IJCSIS)*, vol. 15, no. 5, pp. 332–340, 2017.

[17] N. Ratnasari, A. Susanto, I. Soesanti, and Maesadji, "Thoracic x-ray features extraction using thresholding-based roi template and pca-based features selection for lung tb classification purposes," in *2013 3rd International Conference on Instrumentation, Communications, Information Technology and Biomedical Engineering (ICICI-BME)*, 2013, pp. 65–69.

[18] G. Parthiban, A. Rajesh, and S. Srivatsa, "Diagnosis of heart disease for diabetic patients using naive bayes method," *International Journal of Computer Applications*, vol. 24, no. 3, pp. 7–11, 2011.

[19] S. Krishnan and S. Geetha, "Prediction of heart disease using machine learning algorithms." in *2019 1st international conference on innovations in information and communication technology (ICIICT)*. IEEE, 2019, pp. 1–5.

[20] L. Sapra, J. K. Sandhu, and N. Goyal, "Intelligent method for detection of coronary artery disease with ensemble approach," in *Advances in Communication and Computational Technology*. Springer, 2021, pp. 1033–1042.

[21] S. Ulianova, "Cardiovascular disease dataset," shorturl.at/adjA2.

[22] L. G. Kabari and U. C. Onwuka, "Comparison of bagging and voting ensemble machine learning algorithm as a classifier," *International Journals of Advanced Research in Computer Science and Software Engineering*, vol. 9, no. 3, pp. 19–23, 2019.