Cardiovascular Disease Prediction Using Machine Learning Approaches

Taminul Islam
Department of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
taminul@ieee.org

Adifa Vuyia
Department of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
vuyia25-132@diu.edu.bd

Mahadi Hasan
Department of Data Analytics
University of Tennessee
Chattanooga, Tennessee, United States
kqq544@mocs.utc.edu

Md Masum Rana
Department of Computer Science
University of South Dakota
Vermillion, SD, United States
mdmasum.rana@coyotes.usd.edu

2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES) | 979-8-3503-2391-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/CISES58720.2023.10183490

***Abstract*— As of the release of COVID-19, cardiovascular disease has surpassed all other causes of mortality among both sexes. In most cases, this condition is associated with atherosclerosis and the formation of blood clots. Heart disease, stroke, and other CVDs are major causes of mortality worldwide. Because even a small inaccuracy might lead to exhaustion or death, increased precision, perfection, and accuracy are required for diagnosing and predicting heart-related disorders. In this paper, data was collected from 1189 patients with some attributes related to heart disease and kept 80% data for training, and 20% data for testing and determining their accuracy using different models. In this study, enhanced preprocessing steps were used to increase the accuracy of cardiovascular disease prediction. It aids in determining whether a patient has heart disease and helps a doctor to determine whether or not a patient has cardiovascular disease. The applied models are Extreme Gradient Boosting, Random Forest, CART, Extra Tree Classifier, and Gradient Boosting Machine. This work compared the performance of several Machine Learning algorithms that make use of the accuracy of the metrics, F_1-score, recall, and precision to demonstrate the validity of our findings. The Extreme Gradient Boosting model has achieved the best 91.9% accuracy in this research.**

Keywords—cardiovascular disease, heart attack, cardiovascular prediction, machine learning, xgb

I. Introduction

Cardiovascular disease has been one of the most common causes of death in the medical sector. The number of persons suffering from cardiovascular illnesses is growing annually as a result of an improvement in people's living conditions and an increase in life pressure. Acute cardiovascular disorders and chronic cardiovascular diseases make up the majority of cardiovascular diseases. The computer technique is crucial in the treatment of cardiovascular disorders since traditional wet-lab studies used to diagnose cardiovascular problems frequently prove to be ineffective and time-consuming. Cardiovascular disease, often known as heart disease, is regarded as a lethal condition that is becoming worse faster in our contemporary society. In 2030, the World Health Organization (WHO) anticipates 24.5 additional deaths [1]. In addition, the main causes of this fatal disease, which is now widely prevalent among individuals of all ages, include obesity, high blood pressure, high cholesterol, smoking, family history, and drinking. However, given that cardiovascular illness is accompanied by a variety of symptoms, a rapid and precise diagnosis of the condition seems to be fairly difficult for medical specialists. As a result, a sizable quantity of data is being gathered internationally by the healthcare sector in order to learn more about cardiac disorders and uncover information that will help specialists better comprehend the condition and guarantee that patients receive effective therapy. However, gathering data requires extensive screening and processing in order to efficiently extract information. These analyses on huge datasets, nevertheless, were previously impractical using conventional statistics. As a result, machine learning (ML) has become the most effective technology available today for processing data and using that data to advance the healthcare industry [2]. The human body's most important organ is the heart due to its critical function in blood pumping. The mortality rate associated with cardiac illnesses can be significantly decreased by using machine learning to forecast heart health and anticipate disease [3]. The causes of heart disease might vary widely. The evolution of lifestyle variables such as smoking, physical activity, eating habits, diabetes, and obesity, as well as biochemical elements like glycemia or blood pressure. Because of this, it is necessary to document key cardiac behavior specific to each form of heart illness and to develop a system that aids clinicians in establishing accurate and effective diagnoses. In reality, a medical diagnosis is a categorization mission in which a doctor attempts to locate the flaw by examining the values of several qualities. As a means of transporting oxygen throughout the body, the human heart pumps blood to the lungs. Worldwide, cardiovascular diseases (CVDs) account for the majority of deaths [4]. Predicting the onset of cardiac disease and making a prompt diagnosis is crucial in maximizing the prognosis for patients. Factors such as high body mass index (BMI), cholesterol levels, the use of alcohol and tobacco products, and a sedentary lifestyle are all contributors to the development of cardiovascular disease.

Artificial intelligence includes the field of machine learning. This field aims to develop and improve algorithms that enable computers to continuously improve their performance in response to data [5], [6]. In order for a computer to learn, it must examine its prior knowledge in order to uncover helpful patterns and regularities, even those that a person may overlook. Creating models, like patterns and rules, automatically from data is a primary goal of machine

Authorized licensed use limited to: Southern Illinois University Carbondale. Downloaded on March 08,2025 at 17:28:28 UTC from IEEE Xplore. Restrictions apply.

learning research. Data warehouses and subjects like data mining, statistics, inductive reasoning, pattern recognition, and theoretical computer science" are closely connected to machine learning. Algorithms for machine learning have been used in the past to diagnose cardiac problems. Sadly, the specificity, sensitivity, and accuracy were extremely low. The art of machine learning involves controlling a system without using explicit computation [7]. They are used to examine the analytical setup in large-scale, varied data sets, such as those pertaining to heart ailments. They are utilized in the identification of arrangements (patterns) that enable forecasting and control mechanisms for analysis and medicine.

The paper aims to Prognostic CVDs using the top Five machine learning algorithms. Here is further divided into the following sub-sections: the next section describes an extensive literature review of previous works. A broad overview of the use of machine learning models for cardiovascular diseases and data overview in the third section, explains the number of features and descriptions of each aspect. This study makes use of data preprocessing techniques, as shown in Section III. The rest of the paper was designed with the experimental result, discussion, and conclusions.

II. Literature Review

Following feature standardization and feature extraction using PCA, R. Perumal et al. [8] used seven principal features to train the ML classifiers on the Cleveland dataset of 303 data samples to predict the presence of heart disease. After comparing the accuracy produced by k-NN (69%) and the other methods (LR (87%), SVM (85%), they found that the accuracy provided by the other two methods was quite close.

Ensemble approaches were used by C. B. C. Latha et al. [9] to conduct a comparative study of the Cleveland dataset of 303 trials in order to enhance the predicting accuracy of heart disease risk. They used a brute-force approach to get every conceivable combination of attribute sets and train the classifiers. Using an attribute set of nine features, they were able to boost the accuracy of a weak classifier by a maximum of 7.26 % due to the ensemble approach and to obtain an accuracy of 85.48 % by majority vote employing NB, BN, RF, and MLP classifiers.

Three classification models, logistic regression (LR), decision tree (DT), and Gaussian naive Bayes (GNB) were constructed by D. Ananey-Obiri et al. [10] for predicting cardiovascular disease using the Cleveland dataset. The number of features was cut in half, from 13 to 4, by a process called single-value decomposition. Both LR and GNB were found to have an AUC of 0.87 and a predictive score of 82.75. Other models were proposed, including a support vector machine, k-nearest neighbors, and random forest.

With the help of a fuzzy analytical hierarchy approach for feature weighting and a synthetic neural network for classification tasks, an integrated decision-making system for heart failure risk prediction was developed [11]. In order to improve the accuracy of cardiac illness diagnoses, we present an intelligent system that uses a deep neural network for feature refinement and a 2 statistical model for classification. The model achieves an accuracy of 91.57 percent, a specificity of 93.12%, and a sensitivity of 89.58% [12].

Using a set of weighted fuzzy rules, an adaptive system is described for determining the probability of heart disease. The accuracy of this genetic algorithm-based fuzzy model auto-diagnostic system is 92.3% [13], and it was achieved by deploying an aim of influencing multi-swarm particle optimization techniques A decision tree for classification and algorithms for univariate and multivariate feature selection result in a 92.8% accurate heart disease detection model.

Heart disease is a leading cause of mortality, especially in poor nations. Diagnosis and prognosis of cardiac disease have been the subject of several research looking to introduce bias. Khalil et al.[14] proposed a new deep-learning algorithm for identifying cardiac conditions from a single ECG channel. Chowdhury et al. [15] created a digital stethoscope concept that allows for real-time monitoring of a patient's heart sounds and the detection of any anomalies.

Fast-correlation-based feature selection by Khourdifi et al. [16] enhanced the categorization of heart disease. In the wake of this, we tried out various other classification strategies, such as Naive Bayes, support vector machine (SVM), K-nearest neighbor (KNN), random forest (RF), and an artificial neural network with multilayer perception that was optimized with ant colony optimization (ACO) and particle swarm optimization (PSO) (PSO). In their study [17], Latha et al. incorporated many cardiovascular disease classifiers. Similarly, we improved the performance of insufficient algorithms by combining numerous classifiers. An ML-based method was proposed to detect cardiovascular disease other categorization approaches include decision trees, neural networks, K-nearest neighbors, and support vector machines.

III. Methodology

The methodology of research refers to the specific steps that are done to find, select, process, and analyze data that is relevant to the study. The reader of a research report will have the chance to examine the general validity and dependability of a study in the section devoted to the technique. This section is broken down into three parts: the description of the dataset, the preparation of the data, and the explanation of the model. The methodology that was used for this study is depicted in Figure 1.

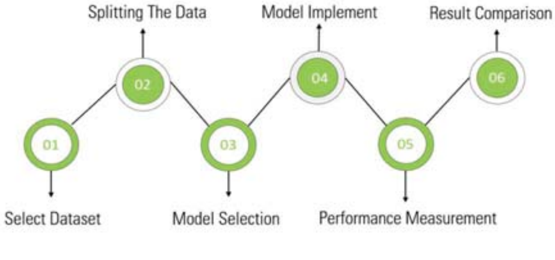

Fig. 1. Working methodology of this research work

A. Data Description

The dataset utilized in this study, also known as the Dataset for Heart Disease [18], was gathered through the Kaggle Platform. This is an initial dataset with several variables, but only twenty attributes have been chosen to predict illness in a single patient. To determine the cause of heart failure, this dataset is essential. There are also 1190 patient records in database files. Table I below shows each entire description, the number of features, and the values for each attribute:

Authorized licensed use limited to: Southern Illinois University Carbondale. Downloaded on March 08,2025 at 17:28:28 UTC from IEEE Xplore. Restrictions apply.

TABLE I. DESCRIPTION OF THE DATA WITH ITS ATTRIBUTES AND VALUES

Description of Attribute	*Values*
age: Years of age for patients in numeric variables. [Minimum Age: 28, Maximum Age: 77]	different values among 28 to 77
sex: patient's gender. 0 means Female and 1 means Male	0,1
Kind of chest pain or discomfort experienced by the patient, ranging from 1 (normal) to 4 (asymptomatic)	Multiple values
resting_blood_pressure:: Blood pressure value in the relaxed position in mm/HG	Multiple values
If your fasting blood sugar is over 120 mg/dl, then (1) is true, and (0) is false, according to the fasting blood sugar enumeration.	0,1
rest_ecg: Electrocardiogram findings are displayed when at rest. It is represented in 3 different values where 0 understands for Normal, 1 stand for Abnormality in ST-T wave, and 2 stands for Left ventricular hypertrophy	0,1,2
max_heart_rate_achieved: reached maximum heart rate	Multiple values
Angina, or chest pain, brought on by exertion; (One depiction yes, zero depictions no)	0,1
st_depression: Exercise-induced ST-depression versus the resting state	Multiple values
Calculations of the peak slope of the ST segment in relation to exercise intensity (0: Normal 1: Upsloping 2: Flat 3: Downsloping)	0,1,2,3
Target: We have to predict the target variable. [0 indicates that the patient is healthy, whereas 1 indicates that they are at risk for heart disease.]	0,1

There are 561 healthy people and 629 people with heart disease in the dataset. Fig. 2. shows the pie chart of patients and normal patients.

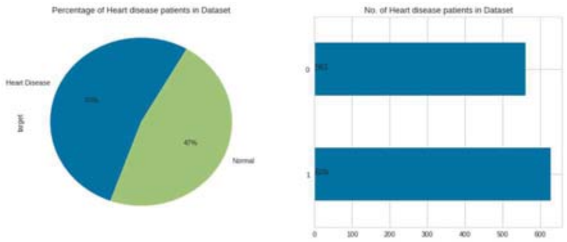

Fig. 2. Represent of illness and normal patient's data in a pie chart

B. Data Pre-Processing

Preprocessing is a crucial stage in the categorization of data. Preprocessing of data is a necessary step in making testable and data clean for any machine learning. In this investigation, the chosen dataset underwent a number of preparation processes. First off, the dataset size was insufficient for this application of machine learning techniques. As previously mentioned, the num of data used for the use of machine learning might introduce bias and affect the outcomes of machine learning models. Data preprocessing tasks including data cleaning, exploratory data analysis, filling in missing values, outliner identification, and deleting redundant data are the main responsibilities of this phase since the dataset contains both redundant and missing information.

The data set was extracted by Google Colab [19] using Python. The convenience of cloud sharing results in decreased data security, yet Google Colab is a need for anybody wishing to save their work to the cloud and sync their notebooks across various devices. Here we divide the total dataset into testing and training dataset. 80% data is used as training data and 20% dataset is used as testing data. Fig. 3. represents the process of data preprocessing.

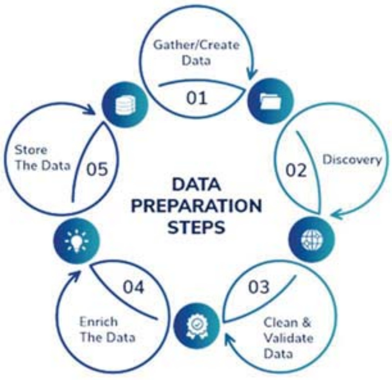

Fig. 3. Data preprocessing steps

C. Machine Learning Model

Machine learning algorithms are increasingly being used in cardiovascular disease prediction because they can process large amounts of data and identify patterns that may not be visible to human experts. In this work, five machine-learning algorithms has been used to predict cardiovascular disease.

1) Gradient Boosting (XGB)

XGBoost is a popular and powerful open-source gradient-boosted trees solution. Gradient boosting is a supervised learning technique that attempts to improve prediction accuracy by combining the outputs of many less sophisticated models [20]. It's a version of gradient-boosted decision trees optimized for speed. Quite a bit faster than competing gradient-boosting methods. The model predominates in tabular or structured information for regression and classification predictive modeling concerns. This model's framework is depicted in Fig. 4.

Authorized licensed use limited to: Southern Illinois University Carbondale. Downloaded on March 08,2025 at 17:28:28 UTC from IEEE Xplore. Restrictions apply.

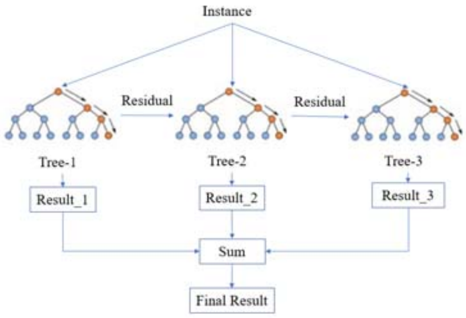

Fig. 4. The primary structure of the XGB model [21]

2) Extra Tree Classifier

The Extra Trees Classifier is an example of an ensemble learning approach, which generates a classification result by averaging the output of a group of "de-correlated" decision trees. Only in the details of its decision trees is it different in principle from a Random Forest Classifier. The first training sample is used to build each decision tree in the Extra Trees Forest. Then, each tree is fed a different subset of k features from the whole feature set at each test node. The tree then uses a predetermined set of mathematical criteria to determine which of these features is most suited to partition the data. By picking features at random, we can generate many decision trees with little correlation. Fig. 5 shows how the supplemental tree classifier performs.

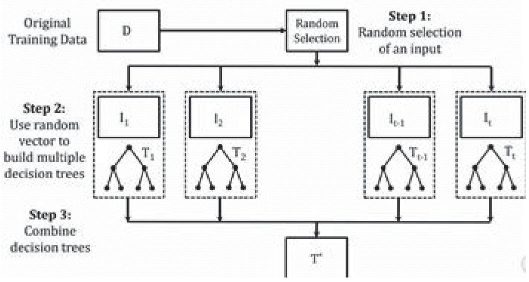
Original Training Data
D
Random Selection
Step 1: Random selection of an input
Step 2: Use random vector to build multiple decision trees
I_1
I_2
I_{t-1}
I_t
T_1
T_2
T_{t-1}
T_t
Step 3: Combine decision trees
T^*

Fig. 5. The basic process of extra tree classifier model [22]

A deep learning model is used to extract a random sample of 'n' leukocyte characteristics and deliver them to each tree. The associated decision tree then uses the most pertinent characteristic determined by the Gini index [23] to divide the data. According to the demand, the number of top features may be chosen, and the feature selection is carried out by arranging the random features in decreasing order based on the Gini relevance of each characteristic. When a feature is randomly selected, the Gini index, which reflects the likelihood that it will be erroneously identified, is shown in equation 1:

$$Gini = 1- \sum_{i}^{n}(\mu_i)^2 \qquad (1)$$

where μ_i is the likelihood that a feature will be categorized as a certain class and "n" is the number of data points [16].

3) Random Forest (RF)

L. Breiman created the random forest algorithm in 2001 [24], and it has since become widely used as a powerful tool for both classification and regression. The method has been found to be highly effective in situations when there are more variables than observations, as it mixes many randomized decision trees and averages their predictions. In addition, it may be scaled up to deal with complex issues, altered to fit a wide range of impromptu learning projects, and produce metrics of varying significance. The accuracy of the random forest algorithm's result predictions is higher than that of the decision tree method. Using many decision trees, the random forest can accurately categorize data [25]. It uses bagging and feature randomization to produce each tree in an attempt to produce a forest of trees whose collective prediction is more accurate than that of any individual tree. A diagram of the Random Forest method is presented in Fig. 6.

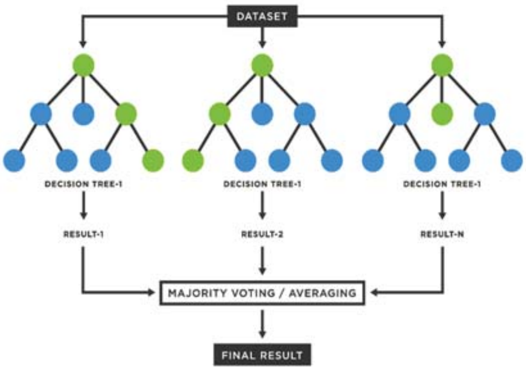

Fig. 6. Basic working procedure of random forest [26]

4) Classification and Regression Trees (CART)

Classification and Regression Trees (CARTs) [27] are a type of predictive model that describe how the values of one outcome variable may be predicted based on the values of other variables. The results of a CART are presented in the form of a decision tree, where each terminal node represents a prediction for the outcome variable and each fork represents a division in the predictor variables. It only supports category variables and utilizes binary division to deliver "success" or "failure" outcomes, and it only works with Boolean variables. The Gini index can be anywhere from 0 to 1, with 0 denoting a perfectly unified set of components, and 1 denoting an abundance of distinct classes. If the Gini index is 0.5, then the items are distributed evenly within certain classes, and if it is 1, then the elements are dispersed randomly across all the classes. The formula for the Gini Impurity is as follows:

$$Gini = 1- \sum_{i=1}^{n}(\rho i)^2 \qquad (2)$$

where p_i is the likelihood that an object belongs to a specific class. Fig 7. Shows the working procedure of CART model.

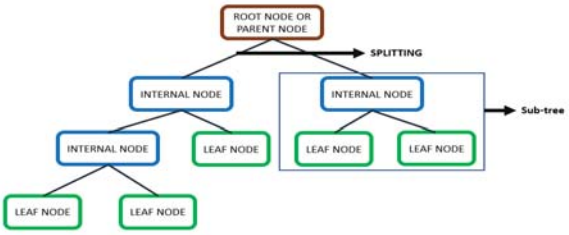

Fig. 7. The primary steps of the CART model [28]

5) Gradient Boosting Machine (GBM)

Using Gradient Boosting [29], several models are trained gradually, cumulatively, and sequentially. When comparing AdaBoost with Gradient Boosting Algorithm, the primary distinction is in the respective algorithms' methods for

Authorized licensed use limited to: Southern Illinois University Carbondale. Downloaded on March 08,2025 at 17:28:28 UTC from IEEE Xplore. Restrictions apply.

determining where weak learners fall short (eg. decision trees). Gradient boosting is similar to the AdaBoost model in that it uses high-weight data points to spot problems, but it also incorporates gradients into the loss function ($y = ax + b + e$) where $e$is the error term) [29] to do so. A model's coefficients are only as good as their loss function, which measures how well the model fits the data. What we're optimizing towards can inform our rational interpretation of the loss function. If we want to use regression to forecast sales prices, for instance, the loss function would be the deviation from the actual home price. Fig. 8 shows the basic organization of GBM.

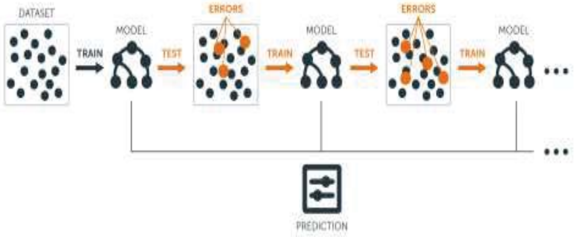

Fig. 8. The basic organization of GBM model [30]

IV. Result & Discussion

This study applied five machine learning models to predictably assess a patient's cardiac condition. As a consequence, the dataset came from Kaggle [18], as was already mentioned. The dataset has 12 characteristics. Data from 937 patients, or 80% of our training dataset, were utilized. Data from 235 patients, or 20% of our testing dataset, were used. The model's performance is displayed in Table III as precision [31], sensitivity [32], accuracy [33], and f_1- score [34]. In this work, five machine-learning models were applied to estimate the likelihood of a heartbeat in a hospitalized patient. In light of the foregoing, the dataset was retrieved from Kaggle and contains medical records that were gathered from various sources. In this work, the Google Colab tool was employed for experimental purposes. We can train our machine learning algorithms using Google Colab which is a free, cloud-based environment for Jupiter notebooks. This study included five machine learning (ML) models: RF, Extra tree classifier, XGB, GBM, and CART in addition to the suggested technique described in the preceding section. The initial stage of model learning made use of the training dataset. The training data was used for the first stage of model learning. In Google Colab, the "Read CSV" operator was utilized to import the dataset for this purpose. Training was used to prevent the selection of comparable values throughout the learning and testing phases of the model. Table II demonstrates the model's performance in the form of precision, sensitivity, specificity, and f_1- score

TABLE II. The Model's Performance In The Form Of Precision, Sensitivity, Specificity, And F1-Score

Model	*Accuracy (%)*	*Precision*	*Sensitivity*	*Specificity*	*$F_1 -$ Score*
XGB	91.9%	0.906	0.943	0.892	0.924
RF	91.0%	0.886	0.951	0.866	0.917
Extra Tree	90.2%	0.878	0.943	0.857	0.909
GBM	85.1%	0.828	0.902	0.794	0.863
CART	84.2%	0.835	0.869	0.812	0.852

The statistical analysis and model performance of a learning and testing phase are computed by this operator. All Models were connected to the same dataset and edited simultaneously during maintenance to determine the test's quality and precise accuracy. Fig 9. shows the comparison ROC curve of the five models.

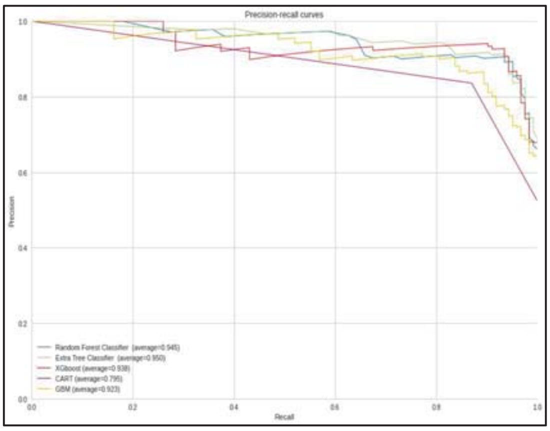

Fig. 9. The comparison ROC curve of the five models

The accuracy value of applied classifiers is shown in Fig. 10. The best performance in this research is XGB which is 91.9%, and the lowest accuracy rate is CART which is 84.2%. RF achieved 91% accuracy in this work. GBM achieved 85.1% model accuracy and the Extra tree classifier got 90.2% accuracy.

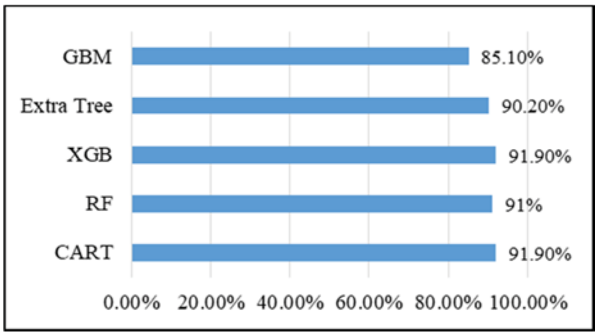

Fig. 10. Accuracy comparison between five algorithms

TABLE III. Comparison Between Previous Work

Ref	Contributions	Algorithms used	Best Accuracy
This work	Applied top machine learning algorithms to predict early-stage cardiovascular disease	XGB, RF, Extra Tree, GBM, CART	91.9% (XGB)
[35]	Authors improved cardiovascular disease prediction. It enabled doctors	NB, SVM, and KNN	86.8% (SVM)

Authorized licensed use limited to: Southern Illinois University Carbondale. Downloaded on March 08,2025 at 17:28:28 UTC from IEEE Xplore. Restrictions apply.

	diagnose cardiovascular illness and identify the patient's heart status.		
[36]	Developed several machine learning algorithms for forecasting cardiovascular disease uncertainty based on various criteria.	RF, SVM, NB, GB, and LR	86.5% (LR)
[37]	SVM, RF, and LR—machine learning methods—were used to predict cardiovascular disease.	SVM, LR, and RF	78.84% (SVM)
[38]	Cloud-based machine learning algorithms were used to forecast heart disorders. An Arduino-based monitoring device detects temperature, blood pressure, and heartbeat every ten seconds.	KNN, DT, NB, LR, SVM, NN and Vote (a hybrid technique with Naïve Bayes and Logistic Regression)	87.4% (Vote)

From Table III, we can understand that this work performs better than other existing works. In this work, we have used five machine learning XGB, RF, Extra Tree, GBM, and CART. Among the other algorithms, XGB achieved the best accuracy 91.9%. In [35] works, authors applied three algorithms to get better accuracy and SVM achieved the best 86.8% accuracy. In [36] work, LR achieved 86.5% best model accuracy whereas in [37] SVM achieved 78.84% accuracy. In [38], Vote (a hybrid technique with Naïve Bayes and Logistic Regression) has got 87.4% accuracy to predict the heat disorders.

V. Conclusions and Future Work

Patients with heart failure are becoming more prevalent every day. A system that can be used to create or classify data rules is required to get out of this dangerous situation and reduce the likelihood of heart disease. Early detection and treatment of heart disease can help to prevent serious complications, such as heart attack, stroke, and death. There are a number of risk factors for heart disease, including high blood pressure, high cholesterol, smoking, diabetes, and obesity. These risk factors can be modified through lifestyle changes, such as exercise, diet, and weight loss. This research can help doctors to identify patients who are at high risk for the disease. This information can be used to encourage patients to make lifestyle changes and to start preventive treatment. This research can also help to improve the quality of life for patients and to prevent serious complications. As a result, this study of machine learning techniques discusses, proposes, and implements a machine learning algorithm that combines five different techniques. When compared to prior research, this study has demonstrated a considerable improvement and a high level of accuracy. In machine learning, preprocessing is a vital step that promotes improved outcomes. In order to increase their accuracy, this paper compared machine learning algorithms with several performance criteria. In our method, missing data from the preprocessing stage is replaced with the mean value. Outcomes demonstrate how well the mean performs when used to replace missing variables. Using the XGB model, the best accuracy of 91.9% was attained. The limitation of this paper is the dataset. We will work more organized and bulk dataset in the future. This research also aims to use deep learning with additional datasets in the future so that the findings show that the system may be efficient and useful for doctors.

References

[1] M. Hassan *et al.*, "Mean Temperature and Drought Projections in Central Africa: A Population-Based Study of Food Insecurity, Childhood Malnutrition and Mortality, and Infectious Disease," *International Journal of Environmental Research and Public Health 2023, Vol. 20, Page 2697*, vol. 20, no. 3, p. 2697, Feb. 2023, doi: 10.3390/IJERPH20032697.

[2] K. Kumar, P. Kumar, D. Deb, M.-L. Unguresan, and V. Muresan, "Artificial Intelligence and Machine Learning Based Intervention in Medical Infrastructure: A Review and Future Trends," *Healthcare 2023, Vol. 11, Page 207*, vol. 11, no. 2, p. 207, Jan. 2023, doi: 10.3390/HEALTHCARE11020207.

[3] S. H. Bani Hani and M. M. Ahmad, "Machine-learning Algorithms for Ischemic Heart Disease Prediction: A Systematic Review," *Curr Cardiol Rev*, vol. 19, no. 1, Jun. 2022, doi: 10.2174/1573403X18666220609123053.

[4] T. Islam, A. Kundu, T. Ahmed, and N. I. Khan, "Analysis of Arrhythmia Classification on ECG Dataset," *2022 IEEE 7th International conference for Convergence in Technology, I2CT 2022*, 2022, doi: 10.1109/I2CT54291.2022.9825052.

[5] D. Mhlanga, "Artificial Intelligence and Machine Learning for Energy Consumption and Production in Emerging Markets: A Review," *Energies 2023, Vol. 16, Page 745*, vol. 16, no. 2, p. 745, Jan. 2023, doi: 10.3390/EN16020745.

[6] A. Holzinger, K. Keiblinger, P. Holub, K. Zatloukal, and H. Müller, "AI for life: Trends in artificial intelligence for biotechnology," *N Biotechnol*, vol. 74, pp. 16–24, May 2023, doi: 10.1016/J.NBT.2023.02.001.

[7] T. Islam, A. Kundu, N. Islam Khan, C. Chandra Bonik, F. Akter, and M. Jihadul Islam, "Machine Learning Approaches to Predict Breast Cancer: Bangladesh Perspective," *Smart Innovation, Systems and Technologies*, vol. 302, pp. 291–305, 2022, doi: 10.1007/978-981-19-2541-2_23/COVER.

[8] K. A. Ramya Perumal, "Early Prediction of Coronary Heart Disease from Cleveland Dataset using Machine Learning Techniques," *International Journal of Advanced Science and Technology*, vol. 29, no. 06, pp. 4225–4234, May 2020, Accessed: Mar. 08, 2023. [Online]. Available: http://sersc.org/journals/index.php/IJAST/article/view/16428

[9] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Inform Med Unlocked*, vol. 16, p. 100203, Jan. 2019, doi: 10.1016/J.IMU.2019.100203.

[10] D. Ananey-Obiri and E. Sarku, "Predicting the Presence of Heart Diseases using Comparative Data Mining and Machine Learning Algorithms Current trends on Superfoods View project Bioinformatics and machine learning View project Predicting the Presence of Heart Diseases using Comparative Data Mining and Machine Learning Algorithms," *Article in International Journal of Computer Applications*, vol. 176, no. 11, pp. 975–8887, 2020, doi: 10.5120/ijca2020920034.

Authorized licensed use limited to: Southern Illinois University Carbondale. Downloaded on March 08,2025 at 17:28:28 UTC from IEEE Xplore. Restrictions apply.

[11] D. Deepika and N. Balaji, “Effective heart disease prediction using novel MLP-EBMDA approach,” *Biomed Signal Process Control*, vol. 72, p. 103318, Feb. 2022, doi: 10.1016/J.BSPC.2021.103318.

[12] L. Ali, A. Rahman, A. Khan, M. Zhou, A. Javeed, and J. A. Khan, “An Automated Diagnostic System for Heart Disease Prediction Based on $\chi 2$ Statistical Model and Optimally Configured Deep Neural Network,” *IEEE Access*, vol. 7, pp. 34938–34945, 2019, doi: 10.1109/ACCESS.2019.2904800.

[13] A. K. Paul, P. C. Shill, M. R. I. Rabin, and K. Murase, “Adaptive weighted fuzzy rule-based system for the risk level assessment of heart disease,” *Applied Intelligence*, vol. 48, no. 7, pp. 1739–1756, Jul. 2018, doi: 10.1007/S10489-017-1037-6/METRICS.

[14] L. El bouny, M. Khalil, and A. Adib, “An End-to-End Multi-Level Wavelet Convolutional Neural Networks for heart diseases diagnosis,” *Neurocomputing*, vol. 417, pp. 187–201, Dec. 2020, doi: 10.1016/J.NEUCOM.2020.07.056.

[15] M. E. H. Chowdhury *et al.*, “Real-Time Smart-Digital Stethoscope System for Heart Diseases Monitoring,” *Sensors 2019, Vol. 19, Page 2781*, vol. 19, no. 12, p. 2781, Jun. 2019, doi: 10.3390/S19122781.

[16] Y. Khourdifi and M. Bahaj, “Heart Disease Prediction and Classification Using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization,” *International Journal of Intelligent Engineering and Systems*, vol. 12, no. 1, 2019, doi: 10.22266/ijies2019.0228.24.

[17] C. B. C. Latha and S. C. Jeeva, “Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques,” *Inform Med Unlocked*, vol. 16, p. 100203, Jan. 2019, doi: 10.1016/J.IMU.2019.100203.

[18] “Heart Disease Dataset | Kaggle.” https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset (accessed Mar. 09, 2023).

[19] “Welcome To Colaboratory - Colaboratory.” https://colab.research.google.com/ (accessed Mar. 09, 2023).

[20] A. Dezhkam and M. T. Manzuri, “Forecasting stock market for an efficient portfolio by combining XGBoost and Hilbert–Huang transform,” *Eng Appl Artif Intell*, vol. 118, p. 105626, Feb. 2023, doi: 10.1016/J.ENGAPPAI.2022.105626.

[21] W. Wang, G. Chakraborty, and B. Chakraborty, “Predicting the risk of chronic kidney disease (Ckd) using machine learning algorithm,” *Applied Sciences (Switzerland)*, vol. 11, no. 1, pp. 1–17, Jan. 2021, doi: 10.3390/APP11010202.

[22] N. Chakrabarty and S. Biswas, “Navo Minority Over-sampling Technique (NMOTe): A Consistent Performance Booster on Imbalanced Datasets,” *Journal of Electronics and Informatics*, vol. 2, no. 2, pp. 96–136, Jun. 2020, doi: 10.36548/JEI.2020.2.004.

[23] B. Chen *et al.*, “A full generalization of the Gini index for bearing condition monitoring,” *Mech Syst Signal Process*, vol. 188, p. 109998, Apr. 2023, doi: 10.1016/J.YMSSP.2022.109998.

[24] L. Breiman, “Random forests,” *Mach Learn*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324/METRICS.

[25] “What is Random Forest? | IBM.” https://www.ibm.com/topics/random-forest (accessed Mar. 09, 2023).

[26] “What is a Random Forest? | TIBCO Software.” https://www.tibco.com/reference-center/what-is-a-random-forest (accessed Mar. 09, 2023).

[27] G. S. Kori and M. S. Kakkasageri, “Classification And Regression Tree (CART) based resource allocation scheme for Wireless Sensor Networks,” *Comput Commun*, vol. 197, pp. 242–254, Jan. 2023, doi: 10.1016/J.COMCOM.2022.11.003.

[28] “CART (Classification And Regression Tree) in Machine Learning - GeeksforGeeks.” https://www.geeksforgeeks.org/cart-classification-and-regression-tree-in-machine-learning/ (accessed Mar. 09, 2023).

[29] L. Delong, M. Lindholm, and H. Zakrisson, “A Note on Multi-Parametric Gradient Boosting Machines with Non-Life Insurance Applications,” *SSRN Electronic Journal*, Feb. 2023, doi: 10.2139/SSRN.4352505.

[30] A. Natekin and A. Knoll, “Gradient boosting machines, a tutorial,” *Front Neurorobot*, vol. 7, no. DEC, 2013, doi: 10.3389/FNBOT.2013.00021/FULL.

[31] M. Kane, “The Precision of Measurements,” *http://dx.doi.org/10.1207/s15324818ame0904_4*, vol. 9, no. 4, pp. 355–379, 2009, doi: 10.1207/S15324818AME0904_4.

[32] W. J. Dixon and A. M. Mood, “A Method for Obtaining and Analyzing Sensitivity Data,” *J Am Stat Assoc*, vol. 43, no. 241, pp. 109–126, 1948, doi: 10.1080/01621459.1948.10483254.

[33] T. Islam, M. A. Hosen, A. Mony, M. T. Hasan, I. Jahan, and A. Kundu, “A Proposed Bi-LSTM Method to Fake News Detection,” *2022 International Conference for Advancement in Technology, ICONAT 2022*, 2022, doi: 10.1109/ICONAT53423.2022.9725937.

[34] O. Sharif, M. Z. Hasan, M. A. Halim, Md. M. Hasan, and M. Z. Sirdari, “Regression Analysis to Rank: Evidence from Profit Risk and Efficiency,” *Communications in Computational and Applied Mathematics*, vol. 2, no. 2, Oct. 2020, Accessed: Mar. 09, 2023. [Online]. Available: http://fazpublishing.com/ccam/index.php/ccam/article/view/45

[35] N. Louridi, M. Amar, and B. El Ouahidi, “Identification of Cardiovascular Diseases Using Machine Learning,” *7th Mediterranean Congress of Telecommunications 2019, CMT 2019*, Oct. 2019, doi: 10.1109/CMT.2019.8931411.

[36] K. G. Dinesh, K. Arumugaraj, K. D. Santhosh, and V. Mareeswari, “Prediction of Cardiovascular Disease Using Machine Learning Algorithms,” *Proceedings of the 2018 International Conference on Current Trends towards Converging Technologies, ICCTCT 2018*, Nov. 2018, doi: 10.1109/ICCTCT.2018.8550857.

[37] W. Sun, P. Zhang, Z. Wang, and D. Li, “Prediction of Cardiovascular Diseases based on Machine Learning,” *ASP Transactions on Internet of Things*, vol. 1, no. 1, pp. 30–35, May 2021, doi: 10.52810/TIOT.2021.100035.

[38] M. S. Amin, Y. K. Chiam, and K. D. Varathan, “Identification of significant features and data mining techniques in predicting heart disease,” *Telematics and Informatics*, vol. 36, pp. 82–93, Mar. 2019, doi: 10.1016/J.TELE.2018.11.007.

Authorized licensed use limited to: Southern Illinois University Carbondale. Downloaded on March 08,2025 at 17:28:28 UTC from IEEE Xplore. Restrictions apply.