

Heart Disease Prediction Using Machine Learning: A Comparative Analysis of Classification Algorithms

Sayama Alam

Lecturer, Department of Computer Science and Engineering, Dhaka Commerce College, Bangladesh

Abstract

Heart disease is a general term that includes many types of heart problems. It is also called cardiovascular disease which means heart and blood vessel disease. Now a day's heart disease is one of the most prominent causes of mortality worldwide. According to estimates from the World Health Organization, heart disease causes 12 million deaths annually worldwide. Cardiovascular diseases account for half of all deaths in the US and other affluent nations. With the advent of new web technology, analysis and prediction of heart disease has paved the way of invoking an efficient and effective research era. Improving patient outcomes and lowering death rates depend on early identification of cardiac disease. To automatically predict, diagnose, and treat heart disease, machine learning (ML) algorithms and techniques have been applied to a variety of available heart disease datasets during the past several years. By using massive volumes of patient data to find patterns that might not be immediately noticeable to medical practitioners, machine learning (ML) offers potential methods for the early identification and prediction of cardiac disease. Logistic Regression, Decision Trees, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Random Forest, Gradient Boosting, Naive Bayes, XGBoost and AdaBoost are the nine machine learning models that have been chosen for evaluation. The models' effectiveness was evaluated based on accuracy, F1-score, and area under the ROC curve (AUC). According to the results, most of the models gave satisfactory results. These results demonstrate how machine learning-based methods can increase the precision of diagnoses, enable early treatments, and support medical practitioners in making well-informed decisions about the management and treatment of cardiac disease.

Keywords: Heart disease, heart disease prediction, machine learning models, models evaluation, classification

1. Introduction

Cardiovascular disease (CVD) represents a broad classification encompassing a diverse array of heart-related ailments. This nomenclature is synonymous with disorders affecting both the heart and vascular systems [1]. On a global scale, CVDs constitute the primary contributor to mortality, accounting for an estimated 17.9 million fatalities annually—approximately 32% of total global deaths. Among these fatalities, 85% are attributable to myocardial infarctions and cerebrovascular accidents [2]. A significant majority, exceeding three-quarters of deaths associated with CVD, transpire in low- and middle-income nations. In the year 2019, noncommunicable diseases were responsible for 17 million premature deaths (occurring in individuals under the age of 70), with CVDs constituting 38% of this total [3]. This category

of disorders encompasses coronary heart disease, cerebrovascular disease, rheumatic heart disease, along with various other pathologies. Over four out of five fatalities attributed to CVD are the result of myocardial infarctions and cerebrovascular accidents, with one-third of these incidences occurring prematurely in individuals younger than 70 [4]. Heart disease manifests in numerous forms. Certain varieties, such as congenital heart disease, are present from birth, whereas other types emerge progressively over time. The most prevalent variant is coronary artery disease (also referred to as coronary heart disease), which advances gradually as a lipid-rich and viscous substance known as plaque accumulates within the arteries supplying blood to the cardiac muscle. This accumulation results in the narrowing or occlusion of blood flow, potentially culminating in severe complications and additional cardiac disorders [5]

Angina- chest pain from lack of blood flow.

Heart attacks- when part of the heart muscle dies from loss of blood flow.

Heart failure- when your heart can't pump enough blood to meet your body's needs.

Arrhythmia- a problem with the rate or rhythm of your heartbeat The causes of heart disease depend on the type of disease. Some possible causes include lifestyle, genetics, infections, medicines, and other diseases [6]. There are many different factors that can make us more likely to develop heart disease. Some of these factors we can change, but others we cannot. The most important behavioral risk factors of heart disease and stroke are unhealthy diet, physical inactivity, tobacco use and harmful use of alcohol [7] Amongst environmental risk factors, air pollution is an important factor. The effects of behavioral risk factors may show up in individuals as raised blood pressure, raised blood glucose, raised blood lipids, and overweight and obesity. These “intermediate risks factors” can be measured in primary care facilities and indicate an increased risk of heart attack, stroke, heart failure and other complications [8]

Identifying those at highest risk of CVDs and ensuring they receive appropriate treatment can prevent premature deaths. Access to noncommunicable disease medicines and basic health technologies in all primary health care facilities is essential to ensure that those in need receive treatment and counseling [9]. Given the prevalence of cardiovascular diseases, which continue to rank among the world's top causes of mortality, detecting heart disease is one of the most important problems facing modern healthcare.

A massive amount of health-related data is stored in databases by medical institutions worldwide because of the quick development of digital technologies [10]. Several studies on the prediction of heart disease using automated methods such as data mining, machine learning, deep learning, etc. have been conducted in recent years [11]. However, identifying heart diseases has historically required expensive and time-consuming specialized procedures including physical examinations, blood tests, stress testing, and electrocardiograms [12]. The progression of machine learning (ML) techniques has created a chance to use automated methods to increase the effectiveness, precision, and affordability of heart disease diagnostics perspectives [13]. Because of allowing systems to learn from data and make predictions or judgments without explicit programming, machine learning—a type of artificial intelligence (AI)—has completely changed several industries [14]. Huge volumes of medical data can be analyzed by machine learning models to find patterns and correlations in the diagnosis of cardiac disease that may not be immediately obvious to human practitioners [15]. These models can categorize a patient's likelihood of having heart disease based on several characteristics, including lifestyle choices, clinical test findings, and demographic data [16]

Nine different machine learning models—each with distinctive advantages, disadvantages, and applicability for various data types—are applied in this work to the identification of heart disease. The

models examined in this paper cover a wide range of algorithmic techniques, from more complicated, non-linear methods to conventional statistical models, enabling a comprehensive evaluation of their efficiency. The aim of this research is to find out the patient's likelihood of receiving a diagnosis of any cardiovascular heart illness based on their medical features, including gender, age, chest pain, fasting blood sugar level, etc. A dataset including the medical history and characteristics of the patient is chosen from the UCI repository. We make predictions about the patient's likelihood of having heart disease using this dataset. Nine algorithms are used to train these medical attributes: Logistic Regression, Decision Trees, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Random Forest, Gradient Boosting, Naive Bayes, XGBoost and AdaBoost.

2. Literature Review

Because machine learning algorithms can evaluate complex datasets and produce predictions, their use in the identification of cardiac disease has grown expressively. Numerous research has shown how successful various machine learning models are, underscoring their potential for early diagnosis and intervention in clinical contexts.

In [17], the authors employed a Decision Tree Classifier Algorithm, achieving a prediction accuracy of approximately 98.83%. Another noteworthy study by [18] utilized the Cleveland dataset from the UCI Machine Learning Repository, which contained 70,000 patient records and 12 distinct features. The performance of Support Vector Machine (SVM), Random Forest, and Logistic Regression models was evaluated using precision, accuracy, recall, F1-score, and ROC-AUC metrics. Among these, SVM achieved the highest accuracy, although it performed slightly below Logistic Regression in some aspects. It also delivered the best ROC-AUC score, effectively balancing true positives and false positives compared to the other models.

In [19] the researchers analyzed 13 clinical features from 1,025 samples using a Random Forest model to detect cardiac disease. This approach achieved an accuracy of 98.54%, successfully identifying key variables and providing dependable predictions to support early diagnosis and personalized treatment planning.

Similarly, the work in [20] focused on developing a highly accurate machine learning framework for early heart disease prediction by applying various feature selection techniques. Using both private and public datasets, along with multiple cross-validation methods, the authors evaluated the effectiveness of their proposed prediction system.

3. About Dataset

This type of dataset is classified as multivariate, meaning it involves the use or analysis of multiple distinct statistical or mathematical variables

in numerical data analysis. In this case, the dataset comprises 14 key attributes: age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiogram results, maximum heart rate achieved, exercise-induced angina, oldpeak (ST depression induced by exercise relative to rest), the slope of the peak exercise ST segment, number of major vessels, and thalassemia. Although the original database contains a total of 76 attributes, only these 14 are consistently used in published studies. To date, machine learning researchers have focused exclusively on these features when conducting predictive modeling and analysis. The dataset includes records collected from various sources, including the Cleveland Clinic Foundation, the Hungarian Institute of Cardiology, V.A. Medical Center, Long Beach,

and the University Hospital Zurich. These sources make the dataset useful for model comparison, but they also pose challenges because missing values and differences in feature distributions may appear in the data.

4. Dataset handling

id	age	sex	dataset	cp	trestbps	chol	fst	restecg	
0	1	63	Male	Cleveland	typical angina	145.0	233.0	True	lv hypertrophy
1	2	67	Male	Cleveland	asymptomatic	160.0	286.0	False	lv hypertrophy
2	3	67	Male	Cleveland	asymptomatic	120.0	229.0	False	lv hypertrophy
3	4	37	Male	Cleveland	non-anginal	130.0	250.0	False	normal
4	5	41	Female	Cleveland	atypical angina	130.0	204.0	False	lv hypertrophy

iol	fst	restecg	thalch	exang	oldpeak	slope	ca	thal	num
13.0	True	lv hypertrophy	150.0	False	2.3	downsloping	0.0	fixed defect	0
16.0	False	lv hypertrophy	108.0	True	1.5	flat	3.0	normal	2
19.0	False	lv hypertrophy	129.0	True	2.6	flat	2.0	reversible defect	1
10.0	False	normal	187.0	False	3.5	downsloping	0.0	normal	0
14.0	False	lv hypertrophy	172.0	False	1.4	upsloping	0.0	normal	0

Figure 1: Output of loading dataset with various attributes

4.1 Dataset visualization

Data visualization refers to the use of charts, graphs, maps, and other visual tools to present information and data in a clear and understandable way. For the purpose of visualization, the following outputs are presented:

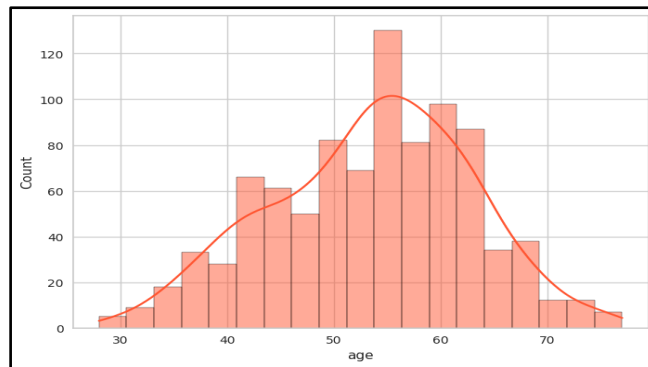


Figure 2: Plot the histogram with custom colors

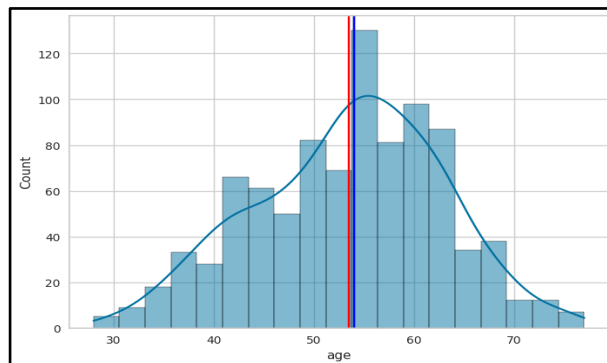


Figure 3: plot of mean, median and mode of age column

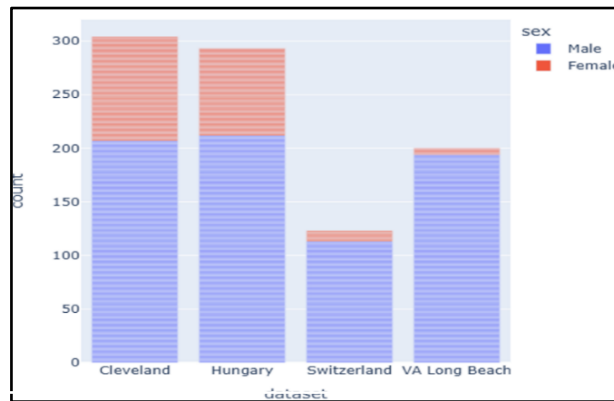


Figure 5: plot of the count plot of dataset column

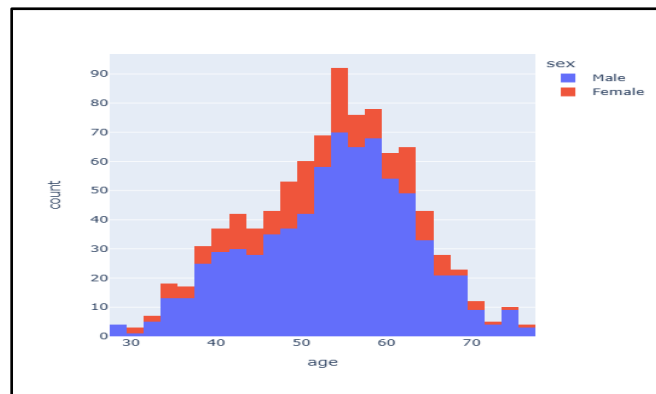


Figure 4: Histogram plot of age column using plotly and coloring this by sex column

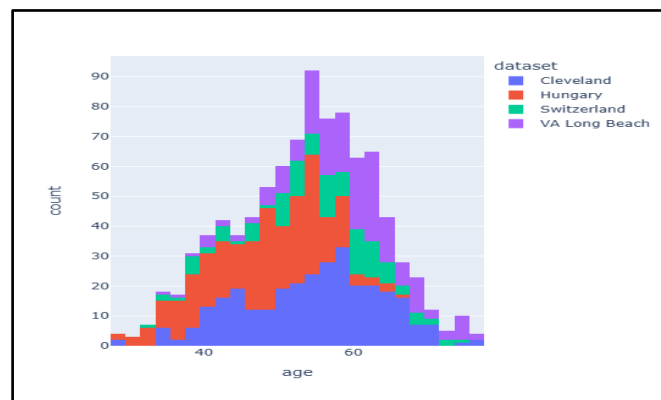


Figure 6: Histogram plot of age column using plotly and coloring by dataset

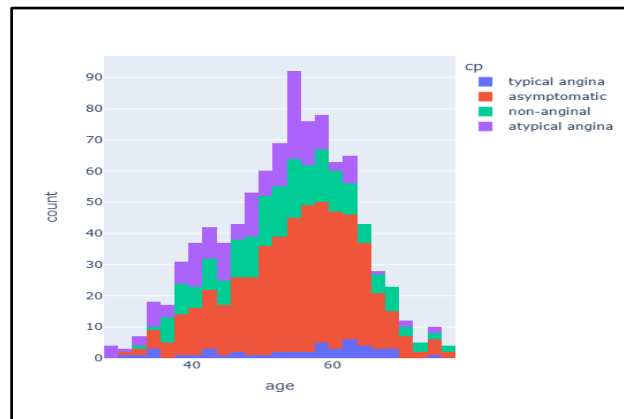


Figure 7: Histogram plot of age column group by cp column

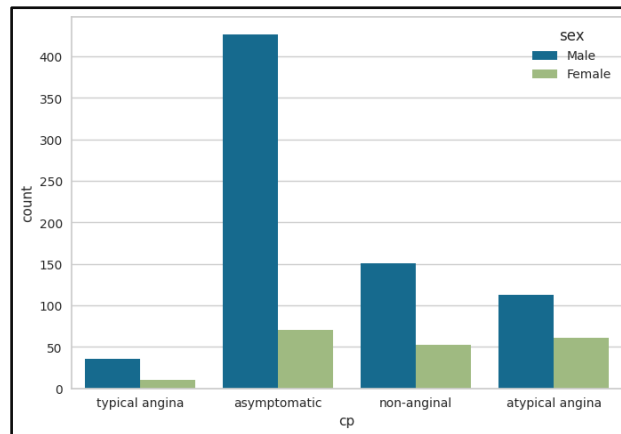


Figure 8: count plot of cp column by sex

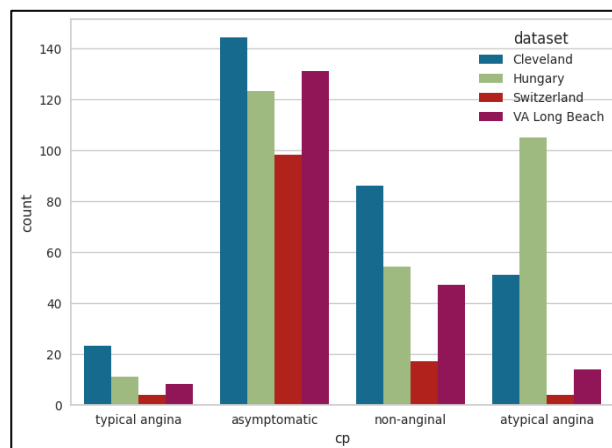


Figure 9: count plot of cp column by dataset column

4.2 Dataset preprocessing

One crucial phase in the data mining process is data preparation. It describes the process of preparing data for analysis by cleaning, converting, and integrating it. We have follow some steps in this stage which are given below:

- a. Handling missing values in trestbps column: There are some missing values because total values are 920 but here, we have 861. After implementation, we have found the Percentage of missing values in trestbps column: 6.41%
- b. Removing missing values using Iterative imputer: A multivariate imputation procedure called Iterative Imputer fills in missing values in a dataset using a round-robin method. When variables in the dataset are related to one another, iterative imputer might be helpful in addressing missing data. By using iterative imputers, we impute the missing values of trestbps column using iterative imputers and as a result we have found that missing values in trestbps column: 0. After that we have imputed other columns with missing values.
- c. Dealing with Missing values (Categorical with Machine Learning Models): Steps:
 Find the columns with missing values and store them in an object
 Find the Columns based on data type
 Numeric columns.
 Categorical Columns.
 Boolean.
 Define the function to impute missing values.
 Apply the function to our dataset with missing values.
 Check the missing values after imputation.
- d. Impute missing values using our functions: At that stage we have used some functions such as Label encoder, Random Forest classifier, Fit transform, Random forest repressors.

4.3 Dealing with Outliers

In machine learning, an outlier is a data point that is considerably distinct from the rest of the data in a data set. Outliers can be caused by several variables, including Data input errors, Sensor errors, rare events, Bad data collection, and Variables that weren't addressed when collecting the data. We have done this using box plot- A box plot, also known as a box-and-whisker plot, is a graph that summarizes a data set and shows how the data is distributed. After implementation, we have found some results which have been given below:

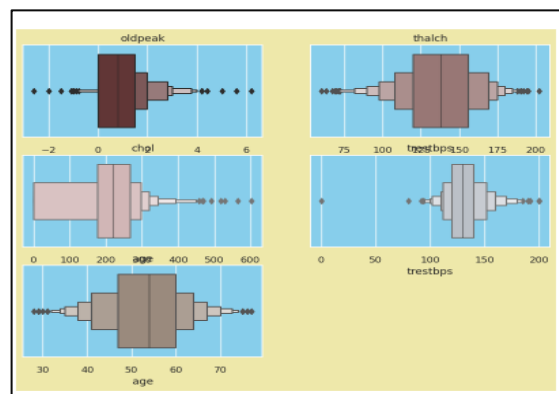


Figure 10: Box plot for all the numeric

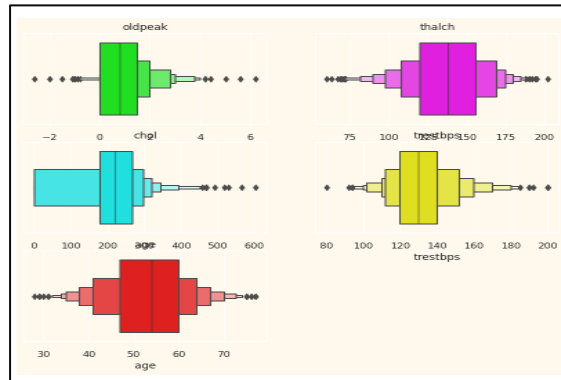


Figure 11: Boxplot of handling old peak

From figure 10 boxplot, it seems there are outliers in the columns, but we will examine whether it's really the outliers or not because sometime these values come in outliers' range but it not. The outliers totally depend on the range which is different for every column in the dataset. For Example: When we investigate the plot and the column of age, we can see that there are outliers. Below 32 is all consider here is outliers but the age of 30 is possible and cannot be an outlier. From figure 11 boxplot, we can observe in the plot that -2 have been removed.

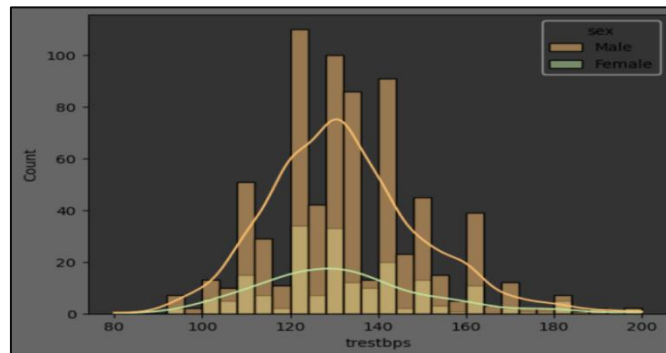


Figure 12: Histogram trestbps column to analyze with sex

From figure 12, there are no outliers in the columns. Columns are cleaned from outliers and missing values.

5. Methodology

This investigation aims to forecast the probability of cardiovascular disease by employing an advanced computerized predictive model, thereby providing substantial assistance to both healthcare practitioners and patients. In order to refine the methodology, the dataset will undergo a cleansing process to eliminate extraneous noise and irrelevant data, while also integrating additional pertinent features. The subsequently enhanced dataset will be utilized to train the predictive model. This improved methodology is anticipated to produce superior accuracy and enhanced overall model efficacy, as depicted in Figure 13 [21].

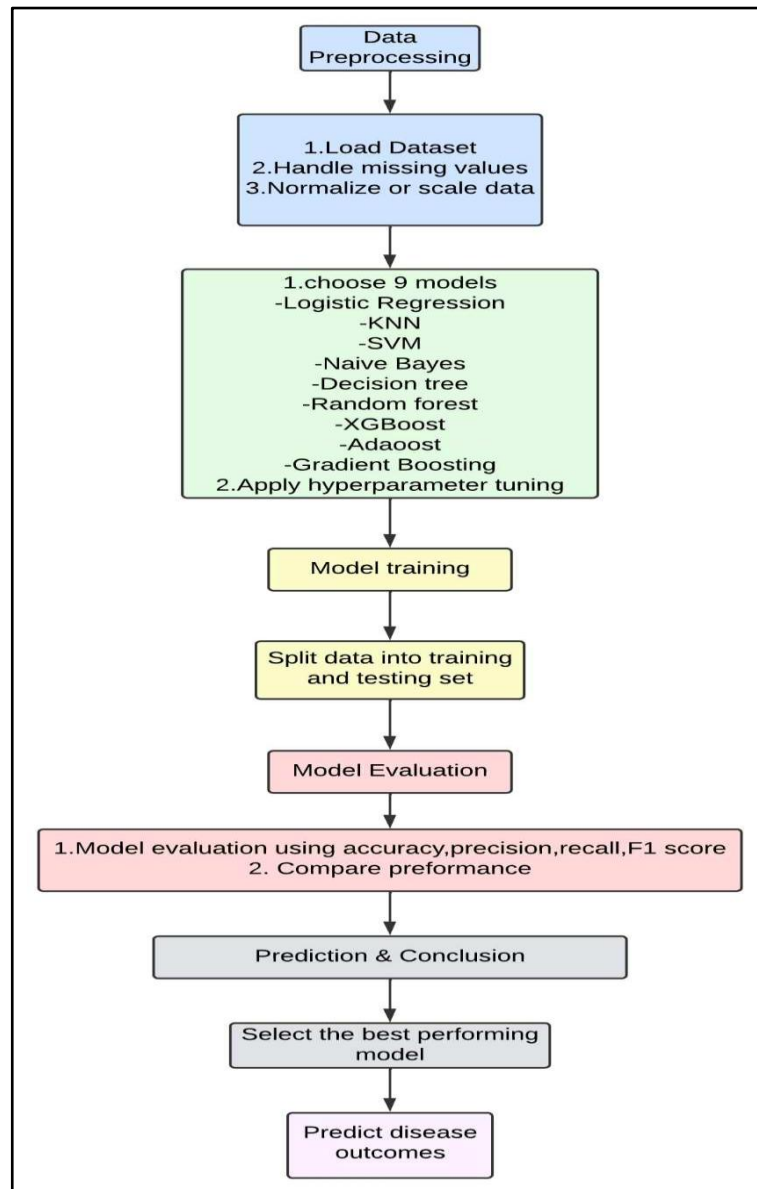


Figure 13: Proposed Machine Learning Framework for Heart Disease

Pseudocode

Disease Prediction Using Machine Learning Models

Input: Dataset D

Output: Best model M_{best} and predicted labels Y_{pred}

Load dataset D

if missing values exist in D then

Handle missing values using imputation or removal

end if

Normalize or scale feature set

Split D into X (features) and Y (target)

Split X, Y into training set (X_{train} , Y_{train}) and testing set (X_{test} , Y_{test})

```
Define model set M = {
Logistic Regression,
KNN,
SVM,
Naive Bayes,
Decision Tree,
Random Forest,
XGBoost,
AdaBoost,
Gradient Boosting
}
for each model m ∈ M do
Perform hyperparameter tuning on m
Train m using (X_train, Y_train)
Predict Y_pred using X_test
Compute performance metrics:
Accuracy, Precision, Recall, F1-score
Store performance results of m
end for
Compare all models based on evaluation metrics
Select model M_best with highest performance score
Use M_best to predict final disease outcomes
return M_best, Y_pred
```

5.1 Normalize or scale data

Normalization constitutes a particular methodology of feature scaling that reconfigures the range of features to conform to a standardized scale. The implementation of normalization, and indeed any technique for data scaling, is necessitated solely in instances where the dataset encompasses features exhibiting disparate ranges. The application of normalized data significantly enhances the performance of models and augments their accuracy. When all features within a dataset are aligned on a uniform scale, it facilitates the identification and visualization of interrelationships among various features, thereby enabling meaningful comparisons [22]

5.2 Nine machine learning models

A machine learning model is defined as a mathematical representation of the output of the training process. A machine learning model is like computer software designed to recognize patterns or behaviors based on previous experience or data [23]. These models are represented as a mathematical function that takes requests in the form of input data, makes predictions on input data, and then provides an output in response. Once these models get trained, they can be used to predict the unseen dataset [24]. For the purpose of heart disease prediction, we have chosen 9 machine learning models which are given below:

Logistic Regression: This statistical method is employed for binary classification tasks (e.g., forecasting yes/no outcomes). It utilizes a logistic function to estimate probabilities.

Decision Trees: These algorithms predict the value of a dependent variable by deriving straightforward decision rules based on the characteristics of the data.

Random Forests: This ensemble method comprises multiple decision trees, commonly applied in both classification and regression contexts, which enhances model accuracy and mitigates overfitting.

Support Vector Machines (SVM): Demonstrating efficacy in high-dimensional spaces, SVM is predominantly utilized for classification purposes, although it is also applicable in regression analyses.

Naive Bayes Algorithm: The Naive Bayes classifier operates under the assumption that the existence of a specific attribute within a class is independent of the existence of any other attributes.

KNN (K- Nearest Neighbors): This algorithm is applicable to both classification and regression tasks. Notably, within the realm of Data Science, it is predominantly utilized for addressing classification challenges. It is characterized as a straightforward algorithm that retains all existing instances and categorizes any new instances by employing a majority voting mechanism among its k nearest neighbors.

Gradient Boosting: Algorithm and AdaBoosting Algorithm: The Gradient Boosting Algorithm and AdaBoosting Algorithm are classified as boosting methodologies utilized when extensive datasets must be processed to generate predictions with elevated accuracy. Boosting constitutes an ensemble learning technique that amalgamates the predictive capabilities of multiple foundational estimators to enhance overall robustness. [25].

XGBoost Algorithm: XGBoost is recognized as a gradient boosting algorithm tailored for supervised learning applications. It is distinguished by its remarkable efficiency and scalability, presenting a performance that is often on par with that of other leading-edge machine learning algorithms in various scenarios [26].

5.3 Split data into training and testing set

Data partitioning refers to the process of segmenting a dataset into two or more distinct subsets. Typically, in a bifurcated division, one subset is utilized for the evaluation or testing of the data while the other is employed for the training of the model. For instance, data may be partitioned according to an 80-20 or a 70-30 ratio between training and testing datasets. The precise ratio is contingent upon the characteristics of the dataset; however, a 70-20-10 distribution for training, development, and testing subsets is considered optimal for smaller datasets [27].

5.4 Train and evaluate models

The dataset designated for training is employed for the purpose of model development, whereas the evaluation dataset is utilized for the assessment of the trained model’s performance. It is imperative that these datasets remain distinct from one another and that the data contained within the evaluation datasets has not been previously encountered during the training phase, in order to guarantee an unbiased estimation of performance [28]. Numerous metrics exist for performance assessment, including Accuracy, Precision, Recall, F1 Score, and Area under the Curve, Confusion Matrix, and Mean Squared Error [29]. Cross-validation represents a methodology that is implemented during the training phase and simultaneously serves as a model evaluation technique [30]. This approach utilizes the training data to adjust the model's parameters, and it is routinely assessed by applying the trained model to an alternative "validation" dataset to evaluate its efficacy on data that has not been previously observed.

Table 1: Evaluation of models after cross validation

Model Name	Cross Validation Accuracy	Test Accuracy
Model Logistic Regression	0.515115005	0.47826087
Model Gradient Boosting	0.625118656	0.625

Model KNNs Classifier	0.576736035	0.586956522
Model Decision Tree Classifier	0.582833151	0.565217391
Model AdaBoost Classifier	0.591266886	0.5978260869565217
Model Random Forest	0.649353779	0.663043478
Model XGboost Classifier	0.632362176	0.663043478
Model Support Vector Machine	0.587703541	0.586956522
Model Naive bayes Classifier	0.570719241	0.532608696

Table 2: The evaluation of metrics resulting from different classifiers

Model Name	Test Accuracy
Model Logistic Regression	0.47826087
Model Gradient Boosting	0.625
Model KNN Classifier	0.586956522
Model Decision Tree Classifier	0.565217391
Model AdaBoost Classifier	0.5978260869565217
Model Random Forest	0.663043478
Model XGboost Classifier	0.663043478
Model Support Vector Machine	0.586956522
Model Naive bayes Classifier	0.532608696

5.5 Hyper parameter tuning

In the realm of machine learning, hyperparameters represent the configuration variables that are established prior to the commencement of the model training process [31]. Hyper parameters are frequently employed to optimize the performance of a model, and they can exert a considerable influence on the model’s accuracy, generalization capabilities, and various other performance metrics [32]. Hyperparameters encompass settings that govern the learning dynamics of the model, including but not limited to, the learning rate, the quantity of neurons within a neural network, or the kernel dimensions in a support vector machine [33]. The objective of hyperparameter tuning is to identify the specific values that yield optimal performance for a designated task [34]. After hyperparameter tuning, the result is Best Model: Random Forest and XGboost Classifier.

6. Descriptive Analysis of Heart Disease Cases by Age, Gender, Origin, and Chest Pain Type

The minimum age to have heart disease starts from 28 years old. Most people get heart disease at the age of 53 to 54 years. Most of the males and females get are with heart disease at the age of 54 to 55 years. Males represent 78.91% of the total heart disease cases, while females represent 21.09%. This means that the number of male patients is much higher than the number of female patients. We have the highest number of people from Cleveland (304) and lowest from Switzerland (123). The highest number of females in this dataset is from Cleveland (97) and lowest are from VA Long Beach (6). The highest number of male are from Hungary (212) and lowest from Switzerland (113). The high number of Typical anginas, Asymptomatic and Non anginal chest pain is in the Cleveland while Atypical angina is highly occurred in Hungary. Lowest number of chest pain (Typical angina, Asymptomatic, non anginal and Atypical angina) is happened in Switzerland as compare to other origins. The highest chest pain that happened is Asymptomatic Angina and the lowest number of cp is that happened is Typical angina. The age in which

highest number of Typical Angina happened is 62 to 63 years. The age in which the highest number of Asymptomatic Angina happened is 56 to 57 years. The age in which highest number of Non Anginal happened is 54 to 55 years. The age at which the highest number of Atypical Angina happens is 54 to 55 years.

7. Result analysis and model performance comparison

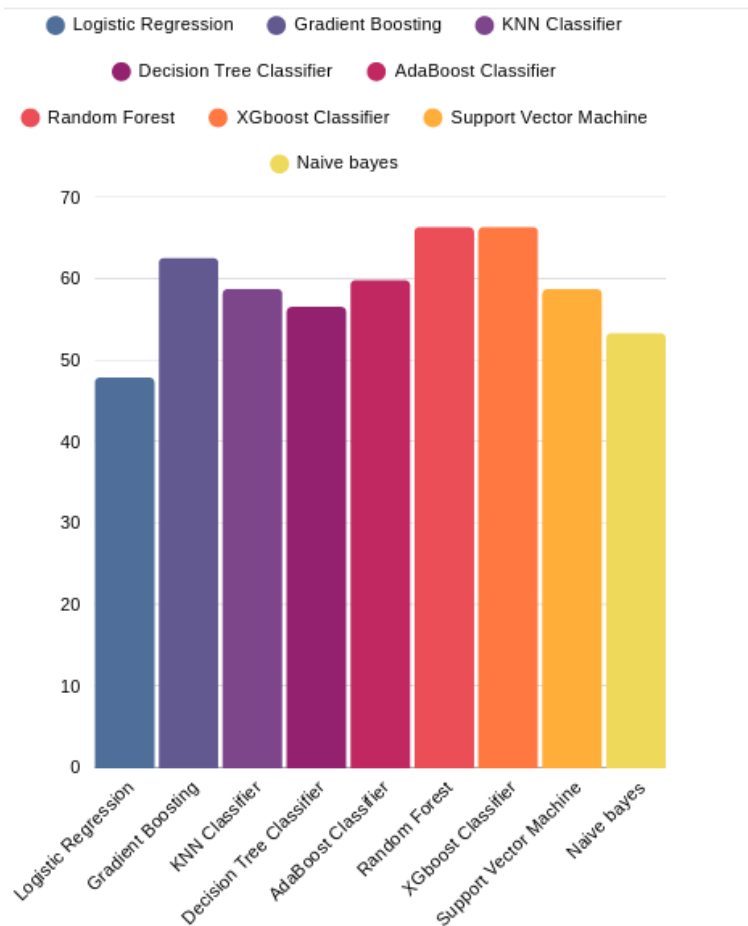


Figure 14: The accuracy results of the nine ML

Following testing and training using the machine learning approach, we discover that XGBoost and Random Forest Classifier’s accuracy is significantly more efficient than that of other algorithms. It is necessary to compute accuracy using the confusion matrix of each algorithm. The accuracy value is computed using accuracy, F1-score and it is determined that XGBoost and Random Forest is the best of them all with an accuracy of 66.30%.

8. Conclusion and future work

One of the most significant and crucial organs in the human body is the heart, and since heart disease prediction is a major concern for people, algorithm accuracy is one of the parameters used to analyze algorithm performance [[35] [36]]. The dataset used for training and testing determines how accurate machine learning algorithms are. We find that XGboost and Random forest is the best method when we analyze it using a dataset whose properties are displayed in Figure.1 and a confusion matrix. To further

improve accuracy, future studies should concentrate on combining deep learning methods such as CNNs (Convolutional Neural Networks) with bigger datasets. Furthermore, creating intuitive diagnostic tools based on these models can help medical personnel make data-driven choices, which will eventually improve patient outcomes and lower the death rates associated with cardiac disease. In this study, Random Forest and XGBoost gave better results than the other models. Both methods are based on ensemble learning. Random Forest builds several decision trees and combines their outputs to make the final prediction. This makes the result more stable and helps reduce overfitting. XGBoost also uses several small models, but it improves the prediction step by step by learning from previous mistakes. These two models can work well with different medical features, such as age, blood pressure, cholesterol, chest pain type, and heart rate. They can also find complex patterns in the dataset. For this reason, Random Forest and XGBoost achieved the highest test accuracy in the final evaluation and were selected as the best models for this dataset.

References

1. b. J. H. P. c. A. G. c. M. B. f. P. S. W. c. J. M. S. a. Joost Boeckmans a, "Clinical utility of the Fibrosis-4 index for predicting mortality in patients with heart failure with or without metabolic dysfunction-associated steatotic liver disease: a prospective cohort study," *ELSEVIER*, 2024 Nov 30.
2. T. A. Gaziano, "Cardiovascular diseases worldwide," *Public Health Approach Cardiovasc*, pp. 8-18, 2022.
3. G. A. M. G. A. J. C. O. A. G. A. E. B. L. M. .. & G.-N.-J. G. B. o. C. D. W. G. Roth, "Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the GBD 2019 study," *Journal of the American College of Cardiology*, p. 76(25), 2020.
4. C. & V. R. S. Andersson, "Epidemiology of cardiovascular disease in young individuals," *Nature Reviews Cardiology*, pp. 230-240., 2018.
5. C. W. J. T. J. & T. Y. D. Shao, "Coronary artery disease: from mechanism to clinical practice," *Coronary Artery Disease: Therapeutics and Drug Discovery*, pp. 1-36, 2020.
6. P. A. S. D. A. K. F. C. M. R. F. T. L. H. S. .. & F. G. Ponikowski, "Heart failure: preventing disease and death worldwide," *ESC heart failure*, pp. 4-25., 2014.
7. B. & S. R. L. Boden-Albala, "Lifestyle factors and stroke risk: exercise, alcohol, diet, obesity, smoking, drug use, and stress.," *Current atherosclerosis reports*, pp. 160-166, 2000.
8. D. W. P. W. & K. W. B. Mozaffarian, " Beyond established and novel risk factors: lifestyle risk factors for cardiovascular disease.," *Circulation*, pp. ,3031-3038, 2008.
9. H. M. D. & Y. S. J. D. M. M., "Resource effective strategies to prevent and treat cardiovascular disease," *Circulation*, pp. 742-755., 2016.
10. Q. S. M. Z. M. & L. R. Wang, "Integrating digital technologies and public health to fight Covid-19 pandemic: key technologies, applications, challenges and outlook of digital healthcare.," *International Journal of Environmental Research* , 2021.
11. L. O. N. D. & G. E. J. Yahaya, "A comprehensive review on heart disease prediction using data mining and machine learning techniques.," *American Journal of Artificial Intelligence* , pp. 20-29., 2020.
12. S. M. & S. V. Srinivasan, " Applications of AI in cardiovascular disease detection—A review of the specific ways in which AI is being used to detect and diagnose cardiovascular diseases. AI in Disease Detection:," *Advancements and Applicatio*, 2025.

13. N. S. K. & S. A. Rana, "Diagnostic strategies using AI and ML in cardiovascular diseases: Challenges and future perspectives.," *In Deep Learning and Computer Vision: Models and Biomedical Applications*., vol. Volume 1, pp. 135-165, 2025.
14. I. J. G. & P. F. Ahmed, "From artificial intelligence to explainable artificial intelligence in industry 4.0," : *a survey on what, how, and where. IEEE transactions on industrial informatics*., vol. 18(8), pp. 5031-5042, 2022.
15. J. A. M. N. M. T. A. M. A. T. S. & W. G. Azmi, "A systematic review on machine learning approaches for cardiovascular disease prediction using medical big data.," *Medical engineering & physics*., pp. 05, 103825., 2022.
16. I. .. & M. K. G. D. J. A. H. L. S. E. D. T. P. C. G. S. Tzoulaki, "Prediction models for cardiovascular disease risk in the general population: systematic review," *bmj*, p. 353., 2016.
17. O. E. E. P. S. & D.-O. F. B. Taylor, ".A model to detect heart disease using machine learning algorithm.," *International Journal of Computer Sciences and Engineering*, 7(11), 1-5., vol. 7(11), pp. 1-5., 2019.
18. P. K. M. M. N. I. U. M. K. S. M. M. H. P. L. I. M. R. & G. N. Bhowmik, "Advancing Heart Disease Prediction through Machine Learning: Techniques and Insights for Improved Cardiovascular Health.," *British Journal of Nur*, (2024)..
19. "Wenyu, Wang. (2024). Heart disease prediction using machine learning models.," *Theoretical and natural science*., vol. 51(1), pp. 9-17., 2024.
20. F. E.-S. B. B. Y. M. A. E.-L. Hosam, " A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method.," *Dental science reports*., vol. 14(1) , 2024.
21. C. M. P. P. G. T. & M. P. L. Bhatt, "Effective heart disease prediction using machine learning techniques. Algorithms," vol. 16(2), p. 88., 2023.
22. D. A. & K. H. P. Keim, "Visualization techniques for mining large databases: A comparison.," *IEEE Transactions on knowledge and data engineering*., vol. 8(6), pp. 923-938, (2002).
23. S. K. M. A. M. R. & K. G. Dargan, " A survey of deep learning and its applications: a new paradigm to machine learning.," *Archives of computational methods in engineering*., vol. 27(4), pp. 1071-1092, 2020.
24. I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions.," *SN computer science* , Vols. 2(3),, p. 160. , 2021.
25. A. J. & F. M. A. Ferreira, "Boosting algorithms: A review of methods, theory, and applications.," *Ensemble machine learning*., pp. 35-85, 2012.
26. P. & Z. A. Florek, "enchmarking state-of-the-art gradient boosting algorithms for classification," *B. arXiv preprint arXiv:2305.17094*., (2023).
27. H. & W. J. T. Abdulnabi, "Novel binning-based methods for model fitting and data splitting improved machine learning imbalanced data.," *.bioRxiv*., pp. 2025-06, (2025).
28. C. K. T. & S. K. Noyunsan, "Performance evaluation of supervised learning algorithms with various training data sizes and missing attributes.," *Engineering and Applied Science Research*, vol. 45(3), pp. 221-229., 2018.
29. G. Z. T. & S. E. M. Naidu, "A review of evaluation metrics in machine learning algorithms.," *In Computer science on-line conference . Cham: Springer International Publishing*., pp. 15-25, (2023, April).

30. C. & B. J. M. Bergmeir, "On the use of cross-validation for time series predictor evaluation.," *Information Sciences*, pp. 191, 192-213., (2012).
31. S. Raschka, " Model evaluation, model selection, and algorithm selection in machine learning.," *.arXiv preprint arXiv:1808.08448*, p. 811.12808., (2018)..
32. P. B. A. L. & B. B. Probst, " Tunability: Importance of hyperparameters of machine learning algorithms.," *Journal of Machine Learning Research*, , Vols. 20(53),, pp. 1-32., (2019)..
33. L. L. H. S. W. & M. L. Liao, " An empirical study of the impact of hyperparameter tuning and model optimization on the performance properties of deep neural networks.," *ACM Transactions on Software Engineering and Methodology (TOSEM)*, vol. 31, (2022)..
34. L. R. A. & P. J. P. Dioşan, ". Improving classification performance of support vector machine by genetically optimising kernel shape and hyper-parameters.," *Applied Intelligence*, vol. 36(2), pp. 280-294, (2012).
35. P. M. J. I. E. R. J. & B. A. Schratz, "Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data.," *Ecological Modelling*, , Vols. 406,, pp. 109-120., (2019).
36. M. M. P. B. K. A. K. B. F. M. Q. J. M. & M. M. A. Ali, " Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison," *.Computers in Biology and Medicine*, , . , vol. 136, p. 104672, (2021).