Comprehensive Analysis of Machine Learning Models for Cardiovascular Disease Detection and Diagnosis

Dr. D P Singh

Amity University Uttar Pradesh Greater Noida Campus

Abstract

Cardiovascular disease (CVD) is a leading cause of mortality globally, making early detection and diagnosis crucial for improving patient outcomes and reducing healthcare costs. Machine learning (ML) models offer promising capabilities for predicting the likelihood of cardiovascular disease, thus assisting in timely diagnosis and treatment. This study conducts an extensive analysis of various ML models, including decision trees, logistic regression, support vector machines, and ensemble methods, to evaluate their effectiveness in predicting cardiovascular diseases. Performance metrics such as accuracy, precision, recall, F1 score, and cross-validation accuracy are utilized to evaluate and compare the effectiveness of models. The findings highlight the potential of machine learning (ML) to improve early prediction and diagnosis of cardiovascular diseases. Through the comparison and analysis of the applied algorithms on the Cleveland and Stat log heart datasets, this research furthers the development of ML techniques in healthcare. The developed machine learning system acts as a valuable resource for healthcare professionals, aiding in the early diagnosis and prediction of cardiovascular diseases, while also offering potential applications for identifying and diagnosing other medical conditions.

Keywords

Diagnosis, Healthcare, Cardiovascular Disease, Predictive Models, Machine Learning, Accuracy, Precision, Recall, F1 Score.

1. Introduction

Cardiovascular disease (CVD) remains a primary health concern worldwide, accounting for a significant proportion of annual mortality rates. Globally, heart disease remains the leading cause of death [39], as highlighted by the World Health Organization, with heart disease and stroke contributing to 17.5 million deaths annually. Over 75% of these fatalities occur predominantly in low- and middle-income countries. Additionally, heart attacks and strokes account for 80% of all deaths caused by cardiovascular diseases (CVDs)[27]. The buildup of plaques in the arteries can block blood flow, potentially leading to a heart attack or stroke. Diagnosis of heart disease often involves observing patient symptoms and performing a physical examination. Common risk factors for CVDs include smoking, aging, family history of heart disease, high cholesterol, high blood pressure, obesity, diabetes, and stressincluding lack of physical activity, poor diet, and excessive use of alcohol and tobacco[10,11,36].

Lifestyle changes, including quitting smoking, losing weight, engaging in regular exercise, and managing stress, can help address certain risk factors. Heart disease is identified by assessing a patient's medical history, conducting a physical exam, and utilizing diagnostic tools such as electrocardiograms, echocardiograms, cardiac MRIs, and blood tests. Treatment options include lifestyle modifications, medications, medical procedures such as angioplasty, coronary artery bypass surgery, or the use of devices like pacemakers and defibrillators[37]. Additionally, the abundance of patient data available through modern healthcare systems enables the development of heart disease prediction models. Machine learning, a method for analyzing and organizing large datasets from multiple perspectives, plays a crucial role in transforming this data into actionable insights [28].

Early detection and effective diagnosis of CVD are essential for preventing severe health complications, reducing treatment costs, and improving patient outcomes. Despite advances in medical diagnostics, the increasing complexity of health data and varying risk factors associated with cardiovascular disease present significant challenges in accurate and timely diagnosis.

In this context, Machine Learning (ML) and Deep Learning (DL), as subsets of Artificial Intelligence (AI), have gained prominence as valuable tools for researchers and healthcare professionals in predicting and diagnosing CVDs. [40].

Artificial Intelligence is a broad concept with various interpretations, and its significance continues to evolve depending on the field of application. In essence, AI can be described as the utilization of machines capable of learning and performing tasks that resemble human cognitive functions [29].

Machine learning (ML) plays a crucial role within the broader field of artificial intelligence (AI). In supervised learning, ML involves using trained algorithms to enable machines to learn, perform tasks, and solve problems independently, based on known input-output relationships [5]. On the other hand, unsupervised learning deals with situations where outputs are unknown. ML and deep learning (DL) are widely applied across various domains, image analysis [6,7], urban traffic management [1,38], digital marketing [2], fraud detection [12], handwritten recognition [13], autonomous driving [30],including data science [41],voice and noise processing [42], and more.

By integrating a large dataset of electronic health records with socio-demographic information, the researchers effectively stratified cardiovascular disease (CVD) risks, achieving high accuracy in their predictions. Similarly, another study [3] implemented a deep learning (DL) algorithm to forecast coronary artery disease (CAD).

A decision support system (DSS) is employed to obtain results and seek second opinions from experienced doctors. This approach helps avoid unnecessary diagnostic tests, saving both time and money [16, 31].

The main priority is to ensure adequate follow-up for this group, as hospitalizations resulting from acute heart failure (HFD) decompensation are the primary driver of healthcare expenses. Research and statistics indicate that heart diseases, particularly HFD, are a major health issue[17,32]. HFD is an increasingly common condition associated with various health issues, including hypertension, insomnia, and heart disease, among others.

Detection of HFD in Electrocardiograms (ECG) involves identifying changes in heartbeat durations, measured as the time intervals between consecutive PQRST waves. MCG (Magnetocardiography) is gaining attention as a promising non-invasive tool for the early detection of ischemic heart disease (IHD). Unlike ECG, MCG is less influenced by issues like electrode-skin contact interference and demonstrates high sensitivity to vortex currents and tangential effects in ischemic cardiac tissue. However, despite its superior signal quality, the interpretation of MCG remains time-intensive, depends heavily on the interpreter's expertise, and is not yet widely used in clinical practice. Therefore, clinicians would greatly benefit from an autonomous system capable of detecting and localizing ischemia at an early stage[18].

In the field of applied medicine, prior studies have demonstrated that machine learning (ML) and deep learning (DL) techniques can be utilized to predict a range of diseases, such as cardiovascular diseases (CVDs) [19,33,34], breast cancer [20,43], and diabetic retinopathy, among others. These risks can be mitigated by adopting a healthy lifestyle, such as reducing salt intake, eating more fruits and vegetables, engaging in regular exercise, and quitting alcohol and tobacco use[21]. Identifying heart disease early, combined with better diagnostic methods and the use of predictive models to pinpoint high-risk individuals, is widely advised to lower fatality rates and improve decisions regarding treatment and prevention strategies.

The growing adoption of machine learning (ML) in medicine underscores its potential in tackling complex health challenges. Specifically for cardiovascular diseases (CVDs), ML can provide doctors with critical insights to aid in early diagnosis and prediction, ultimately enhancing treatment outcomes and public health.

In recent years, machine learning (ML) models have shown substantial promise in assisting clinicians by providing predictive insights into patient health. Through the analysis of patient data, including demographics, clinical test results, and lifestyle factors, ML algorithms can identify patterns and correlations indicative of cardiovascular conditions. These predictive capabilities have the potential to support early diagnosis, enabling healthcare professionals to implement preventive or therapeutic measures more effectively.

This study presents an extensive evaluation of various ML models for cardiovascular disease prediction, comparing traditional algorithms, such as decision trees and logistic regression, with more complex models, including neural networks and ensemble methods. The study aims to determine which models best suit cardiovascular disease prediction in clinical contexts by assessing performance on key metrics such as accuracy, precision, recall, F1 score, and cross-validation accuracy. The results of this analysis will offer insights into the model selection process for cardiovascular disease prediction, assisting healthcare providers and researchers in choosing optimal ML techniques for enhanced diagnostic precision and patient care.

2. Literature Review:

Diagnosing heart disease is the initial step in a patient's treatment within the cardiovascular department. However, predicting these diseases has become a major concern for physicians and researchers in the field. Recently, several studies have focused on enhancing systems to predict heart failures and cardiovascular conditions, leveraging the capabilities of AI techniques and machine learning tools.

The cardiovascular system is formed by the integration of the lymphatic and vascular systems [14]. The heartbeat is a sequence of actions that make up a heart cycle. Typically, the cycle involves both atria, followed by synchronized contractions of each ventricle shortly afterward. The heart is made up of heart muscle cells that are interconnected, so when one cell contracts, it causes adjacent cells to contract as well. The muscles relax between beats during the cardiac cycle, allowing aerobic respiration to occur. This study delves deeper into these two components.

2.1. The first phase is called 'Systole,' which refers to contraction. This occurs when the ventricles contract, forcing blood into the heart's vessels, with the A–V valves closing and the semilunar valves opening

2.2. The second phase is called 'Diastole,' which refers to relaxation. This happens when the ventricles relax, creating a back-pressure that causes the semilunar valves to close and the A–V valves to open.

Previous research has demonstrated encouraging outcomes for the automatic detection of cardiovascular diseases (CVD). However, some challenges remain. One major issue is that studies using private datasets often face problems such as database variability and small sample sizes, particularly in magnetocardiogram (MCG) research, where large public datasets are scarce. On the other hand, for electrocardiogram (ECG) studies using public datasets, performance may not transfer equally well from benchmark datasets to clinical settings. Public ECG datasets are typically composed of representative samples, but they may be biased toward identifying abnormal cases, which could influence early diagnosis.

Several studies have been carried out to evaluate the significance of various features. To tackle this, an extensive set of features has been created, including the two previously mentioned categories. The following conclusions are drawn from the feature importance analysis:

(i) The synchronization of T wave repolarization is recognized as an important characteristic for identifying individuals with IHD.

(ii) The features suggest that the properties of the magnetic pole are associated with the locations of coronary stenosis.

Recently, machine learning (ML)-based clinical decision-making has been applied in healthcare. Recent developments in ML have highlighted the advantages of discriminative classifiers for the automatic detection of cardiac diseases. Earlier research has shown that machine learning algorithms such as SVM, RF (Random Forest), LR (Logistic Regression), BPNN (Back Propagation Neural Network), and MLP (Multilayer Perceptron) are effective decision-making tools for predicting heart disease using individual data. Additionally, several studies have highlighted the advantages of hybrid models, which have demonstrated impressive results in heart disease prediction. Notable examples include RF combined with a linear model, MLP, Bayes Net (BN), majority voting of NB, and RF with two stacked SVMs[22].

The accuracy of risk prediction improves with fewer attributes, achieved by utilizing methods such as the K-nearest neighbor algorithm, Naïve Bayes, and neural networks. The authors found that accuracy increases when fewer attributes are used, thanks to the application of different techniques[23].

Meanwhile, Shadab et al. applied Naive Bayes (NB) data mining to assist users in finding answers to predefined questions in a web-based application. Doctors utilize intelligent decision-making tools, and the accuracy of the NB algorithm for heart disease diagnosis can be enhanced using various techniques[35].

O. Terrada et al,[24] introduced a medical diagnostic support system designed for the early detection of atherosclerosis. They assessed two supervised machine learning algorithms, artificial neural networks (ANN) and K-nearest neighbors (KNN), on four distinct heart disease datasets: Cleveland, Hungarian, Switzerland, and Long Beach. The performance metrics applied in this study were sensitivity, specificity, and accuracy. The findings indicated that the ANN algorithm outperformed KNN across all datasets. Similarly, O. Terrada et al.[25] also proposed a system for the early prediction of atherosclerosis disease.

M.S. Amin and others [9],created a model to predict heart disease by employing various machine learning classification methods, such as KNN, Decision Tree, Logistic Regression, Support Vector Machine, Naive Bayes (NB), a hybrid approach called Vote, and Neural Network.

A.K. Dwivedi [8],applied various machine learning algorithms for classification on the Statlog heart disease dataset, including ANN, SVM, NB, LR, KNN, and Classification Trees. They employed ten-fold cross-validation for assessment and calculated eight performance metrics, including accuracy, sensitivity, specificity, precision, negative predictive value, and false positive rate. The results showed that Logistic Regression achieved the highest accuracy and sensitivity, SVM was the most precise and specific, while Classification Trees exhibited the highest rates of misclassification and false positives.

A. Bhatt and others[4], aimed to predict heart disease using two machine learning algorithms on separate datasets. The algorithm was implemented on the Hungarian dataset, while the Naive Bayes algorithm was applied to the echocardiogram dataset. To evaluate the performance of both models, various metrics were used, including the confusion matrix, accuracy, true positive rate, precision, F-measure, and ROC area. The experiments were performed using the Weka data mining tool, with two tests conducted for each dataset—one considering all attributes and another with a selection of attributes. The findings revealed that accuracy improved when all attributes were included. The accuracy on the Hungarian dataset was 65.64% using selected attributes and 82.3% with all attributes. On the echocardiogram dataset, the accuracy was 93.24% with selected attributes and 98.64% with all attributes.

In [15], the findings indicated that the hybrid algorithm achieved the highest accuracy, suggesting that combining multiple algorithms can sometimes enhance performance. However, this result may not apply universally to other datasets or tasks, and additional experimentation is required to assess the hybrid algorithm's effectiveness in different contexts. Furthermore, the study in [26] emphasizes the importance of thorough experimentation and evaluation when selecting a machine learning algorithm or data mining tool for a specific task.

In summary, these studies emphasize the need for careful experimentation and assessment when selecting a machine learning (ML) algorithm or data mining tool for a specific task. However, despite progress in developing ML algorithms, several challenges remain. One key issue is the absence of standardized datasets, with each study using a different dataset, making it hard to compare results across studies. Lastly, many studies overlook the interpretability of the models, which is crucial for healthcare professionals to understand how predictions are made. This lack of interpretability could hinder the adoption of these models in practical, real-world scenarios.

3. Proposed methodology:

Our proposed method aims to identify the most suitable algorithm for predicting heart disease by evaluating the performance of twelve machine-learning models[44,45,46,47].

3.1.Gaussian Naive Bayes (GNB)

Gaussian Naive Bayes assumes that the features follow a Gaussian (normal) distribution.

Model Assumption: Features are conditionally independent given the class label.

Probability Model:

$$P(C|X_1,X_2,X_3 \dots\dots\dots X_n) = \frac{P(C)\prod_{i=1}^{n}P(X_i|C)}{P(X_1,X_2,X_3 \dots\dots\dots X_n)}$$

Where C is the class, X_1,X_2,…,Xn are the feature values.

Gaussian Assumption: The conditional probability for feature X_i given class C is modeled as:

$$P(X_i|C) = \frac{1}{\sqrt{2\pi\sigma_C^2}}e^{\left(-\frac{(X_i-\mu_C)^2}{2\,\sigma_C^2}\right)}$$

Where μ_C and $\sigma_C$are the mean and standard deviation of the feature for class C

3.2. Bernoulli Naive Bayes (BNB): Bernoulli Naive Bayes is used for binary/Boolean features (0/1 values).

Model Assumption: Like GNB, Bernoulli Naive Bayes assumes conditional independence between features.

Probability Model:

$$P(C|X_1,X_2,X_3 \dots\dots\dots X_n) = \frac{P(C)\prod_{i=1}^{n}P(X_i|C)}{P(X_1,X_2,X_3 \dots\dots\dots X_n)}$$

$$P(X_1|C) = P_C^{X_i}(1-p_c)^{(1-X_i)}$$

where p_c is the probability that feature X_i =1 in class C.

3.3.Support Vector Machine (SVM): SVM is a supervised learning model that tries to find a hyperplane that best separates the data into classes.

Objective: Find a hyperplane that maximizes the margin between two classes.

Model: The decision boundary is defined by: $w^T x + b = 0$

where w is the weight vector, x is the input vector, and b is the bias.

Optimization: The goal is to maximize the margin between the two classes, subject to the constraint that all points are classified correctly: $\min_{w,b}\frac{1}{2}\|w\|^2$. Subject to $y_i(w^T x + b) \geq 1$ for all i , where $y_i\{-1,1\}$is the class label.

3.4. Decision Trees: A decision tree is a highly effective tool in supervised learning, used for both classification and regression tasks. It is structured as a tree-shaped flowchart where each internal node represents a test on an attribute, each branch represents an outcome, and each leaf node represents a class label. The tree is built by iteratively splitting the training data into subsets based on attribute values, stopping according to criteria such as maximum tree depth or minimum samples required for node splitting.

Mathematically, it's structured as:

Entropy (used in classification trees): Entropy measures the impurity in a split. For a node with binary classification (0 or 1), entropy E is defined as:

$$E(S) = -p_1 log_2(p_1) - p_0 log_2(p_0)$$

Where $p_1$the proportion of is class 1 in set S and p_0 is the proportion of class 0.

Gini Index (another impurity measure): $Gini(S) = 1 - p_1^2 - p_0^2$

Recursive Splitting: The tree recursively splits data at each node by selecting a feature and threshold that minimizes impurity (e.g., entropy or Gini) at the child nodes.

Cost Function : For a regression tree, the cost function to minimize is usually the mean squared error (MSE):

$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_1)^2$,where y_i is the true label, and $\hat{y}_1$ is the predicted label for i-th the sample.

3.5. Random Forest: Random Forest, a widely-used ensemble learning method involving decision trees, creates a 'forest' of multiple trees. These trees are usually trained with the 'bagging' technique, which merges multiple models to improve the overall result. Random Forest boosts the performance of Decision Trees by reducing variance, achieved by growing more trees and introducing more randomness into the model. Rather than always choosing the most significant feature for splitting nodes, it selects the best feature from a random subset of features, leading to a more robust model.

Random Forest Regression is an ensemble machine learning technique that handles both regression and classification problems by using several decision trees, along with Bootstrap and Aggregation, a process commonly referred to as bagging. Rather than relying on a single decision tree, this approach combines the outputs of several trees to produce the final result. In Random Forest, numerous decision trees act as the core learning models.

3.6. Gradient Boosting

Gradient Boosting is an iterative process where weak learners (usually decision trees) are added sequentially to minimize a loss function. Its mathematics is based on functional gradient descent.

Loss Function: Let $L(y, \hat{y})$be the loss function to be minimized (e.g., log loss for classification, MSE for regression). The goal is to find a function F(x) such that the predictions F(x) minimize this loss: $F(x) = arg \min_{F(x)} \sum_{i=1}^{n} L(y_i, F(x_i))$

Additive Model: Gradient Boosting builds an additive model: $F_m(x) = F_{m-1}(x) + \alpha . h_m(x)$

where $F_{m-1}(x)$ is the prediction from the previous iteration, $h_m(x)$ is the new decision tree (or weak learner), and α is the learning rate.

Gradient Descent: The new learner $h_m(x)$ is fitted to the negative gradient of the loss function with respect to the current predictions: $h_m(x) = arg \min_{F(x)} \sum_{i=1}^{n} \left[-\frac{\partial L(y_i, F_{m-1}(x))}{\partial F_{m-1}(x)} - h(x_i)\right]^2$

Final Prediction: After M iterations, the final prediction is: $\hat{y} = F_m(x) = \sum_{m=1}^{M} \alpha h_m(x)$.

3.7.K-Nearest Neighbors (KNN)

KNN is a non-parametric, instance-based learning algorithm used for both classification and regression.

Distance Calculation:

To predict the class or value of a new data point, KNN determines the distance between the new point and all points in the training data. The Euclidean Distance is the most frequently used distance metric: $d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$

where x and y are two points in an n-dimensional space.

Prediction: For classification, the algorithm finds the k-nearest neighbours to the new data point and assigns the class based on a majority vote of the neighbours' labels.

For regression, it predicts the value by taking the mean (or weighted average) of the nearest neighbours' values.

3.8. Extreme Gradient Boosting (XGBoost)

XGBoost is a method that implements gradient-boosted decision trees, optimized for high performance and speed.

In boosting, XGBoost builds a series of decision trees, with each new tree working to fix the mistakes made by the earlier ones by placing more emphasis on the misclassified samples.

Gradient Descent: XGBoost minimizes the loss function by gradient descent. For a loss function $L(y, \hat{y})$, where y is the true label and $\hat{y}$ is the prediction, the next tree tries to minimize: $L = \sum_{i=1}^{n} Loss(\ y, \hat{y}) + \sum_{k=1}^{K} \Omega(f_k)$

where: $\Omega(f_k)$ is a regularization term to prevent overfitting.

f_k is a weak learner, typically a decision tree.

The loss function is often chosen as logistic loss for classification and mean squared error for regression.

Regularization: To avoid overfitting, XGBoost uses both L_1 (Lasso) and L_2 (Ridge) regularization: $\Omega(f) = \Upsilon T + \frac{1}{2} \lambda \sum_{j=1}^{T} \omega_j^2$

where T is the number of leaves, ω_j is the weight of leaf j and Υ λ are regularization parameters.

3.9.AdaBoost (Adaptive Boosting)

AdaBoost is a technique that merges several weak classifiers to form a more powerful classifier. These weak classifiers are typically decision trees with only one split (known as decision stumps). The method works by giving more weight to instances that were misclassified in previous iterations, prompting the next classifiers to concentrate on the more challenging cases.

Objective: Minimize the classification error by combining weak learners.

Steps:

1. Initialize weights $w_i = \frac{1}{N}$, where N is the number of samples.
2. For each classifier t:

Train a weak learner $h_t(x)$ with the weighted dataset.

Compute error $\varepsilon_t = \frac{\sum \omega_i \ I(y_i \neq h_t(x_i))}{\sum \omega_i}$ where I(·) is the indicator function.

Compute classifier weight $\alpha_t = \frac{1}{2} \ln(\frac{1-\varepsilon_t}{\varepsilon_t})$.

Update weights for misclassified points: $w_i \leftarrow w_i . e^{\alpha_t I(y_i \neq h_t(x_i))}$, and normalize the weights.

Final classifier is a weighted vote of weak learners: $H(x) = sign(\sum_t \alpha_t h_t(x))$

3.10. CatBoost (Categorical Boosting)

CatBoost is a gradient boosting algorithm developed to handle categorical features natively. It is based on boosting decision trees and is highly efficient in dealing with high-cardinality categorical data. CatBoost reduces prediction shifts and applies an ordered boosting strategy.

Objective: Minimize a loss function (usually log-loss for classification or RMSE for regression) by iteratively building an ensemble of trees.

Mathematical steps: At each iteration, fit a decision tree that predicts the gradient of the loss function with respect to the predictions made by the current ensemble of trees.

Update the model by adding the newly fitted tree to the ensemble: $F_m(x) = F_{m-1}(x) + \eta. h_m(x)$

where $F_m(x)$ is the model after m-th iteration, $h_m(x)$ is the decision tree at iteration , and η is the learning rate.

CatBoost applies a permutation-driven strategy to avoid overfitting on categorical data by reducing the effect of target leakage. The ordered boosting formula can be written as: $F_m(x) = F_{m-1}(x) - \eta . \frac{\partial L}{\partial F_{m-1}(x)}$

where L is the loss function and η is the learning rate.

3.11. Logistic Regression:

Logistic regression is a statistical approach mainly used for binary classification, aiming to predict the likelihood of a two-outcome event (e.g., yes/no, 0/1, true/false). This model can be mathematically represented in the following way:

a. Model Representation:

The logistic regression model predicts the probability that a given input x belongs to a particular class. This probability is modeled using the logistic (sigmoid) function.

Probability of Class 1: $P\left(y = \frac{1}{x} = \sigma(z)\right) = \frac{1}{1+e^{-z}}$

Where:

$\sigma(z)$ is the sigmoid function.

z is the linear combination of input features and model parameters:$z = w^T + b$.

$w = [w_1, w_2, w_3, \ldots . w_n]$ are the weights (coefficients) associated with each feature.

b is the bias (intercept) term.

$x = [x_1, x_2, x_3, \ldots \ldots x_n]$ are the input features.

Probability of Class 0: $P\left(y = \frac{0}{x}\right) = 1 - P\left(y = \frac{1}{x}\right) = 1 - \sigma(z)$

b. Decision Boundary: The model classifies the input x as class 1 if $\left(y = \frac{1}{x}\right) \geq 0.5$, and as class 0 otherwise. The decision boundary occurs where the probability is exactly 0.5,which corresponds to z=0: $w^T + b = 0$

c.Cost Function: To train the logistic regression model, we minimize a cost function. The most commonly used cost function for logistic regression is the log-loss or binary cross-entropy loss:

$$J(w, b) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^i log\left(P(y^i = 1/x^i)\right) + (1 - y^i) log\left(1 - P(y^i = 1/x^i)\right)\right]$$

Where:

m is the number of training examples.

y^i is the actual label of the i-th example.

$P(y^i = 1/x^i)$ is the predicted probability that the i-th example belongs to class 1.

d. Gradient Descent Optimization:

The model parameters w and b are updated iteratively using gradient descent to minimize the cost function: $w \coloneqq w - \alpha \nabla_w J(w, b)$

$$b \coloneqq b - \alpha \nabla_b J(w, b)$$

Where: α is the learning rate. $\nabla_w J(w, b)$ and $\nabla_b J(w, b)$ are the gradients of the cost function with respect to the weights and bias, respectively.

e. Prediction:

Finally, once the model is trained, predictions are made by computing the probability P(y=1|x) for a new input x and assigning the class label based on the decision boundary:

$$\check{y} = \begin{cases} 1 & if\ \sigma(z) \geq 0.5 \\ 0 & if\ \sigma(z) < 0.5 \end{cases}$$

This is the basic mathematical framework behind logistic regression.

Confusion Matrix in Machine Learning:

A confusion matrix summarizes a machine learning model's performance on a test dataset, visually displaying both accurate and inaccurate predictions. It is commonly used to evaluate classification models, which assign categorical labels to input data. This matrix is essential for assessing a classification model's performance, providing detailed counts of true positives, true negatives, false positives, and false negatives. It enables a deeper understanding of the model's recall, accuracy, precision, and overall ability to distinguish between classes by showing the frequency of predicted outcomes on the test dataset[44-47]..

4.1 Accuracy: Accuracy measures a model's effectiveness by calculating the ratio of correctly classified instances to the total number of instances.

$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$,

where TP= True positives, TN= True negatives, FP= False positives and FN= False negatives.

4.2 Precision: Precision refers to the accuracy of a model's positive predictions. It is measured by the ratio of true positive predictions to the total number of positive predictions made by the model.

$$Precision = \frac{TP}{TP + FP}$$

4.3 Recall: Recall measures how well a classification model can identify all the relevant instances within a dataset. It is calculated by dividing the number of true positive (TP) cases by the total number of true positives and false negatives (FN).

$$Recall = \frac{TP}{TP + FN}$$

4.4 Specificity: Specificity, an essential metric for evaluating classification models, particularly in binary cases, measures how accurately a model identifies negative instances, also known as the True Negative Rate.

$$Specificity = \frac{TN}{TP + FP}$$

5. Data Cleaning and Feature Engineering: To evaluate cardiovascular disease, we utilized data from (https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset/code), comprising 14 clinical features. These features were employed to develop predictive models for heart disease detection. As shown in Table 1, the dataset is complete, with no missing values.

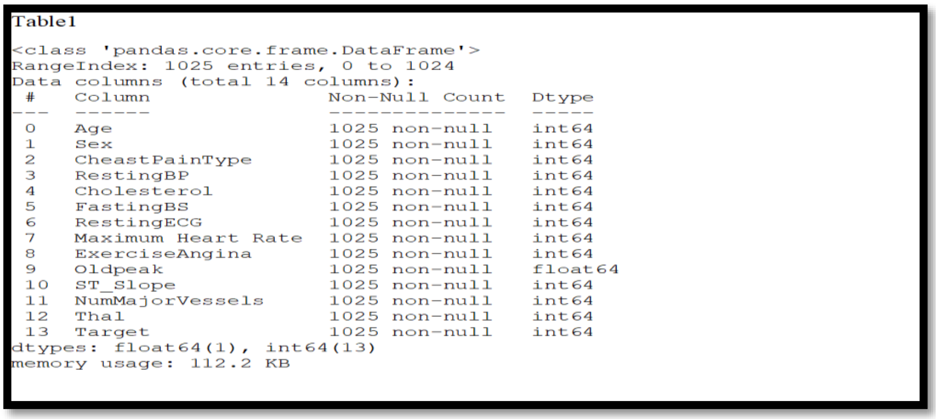
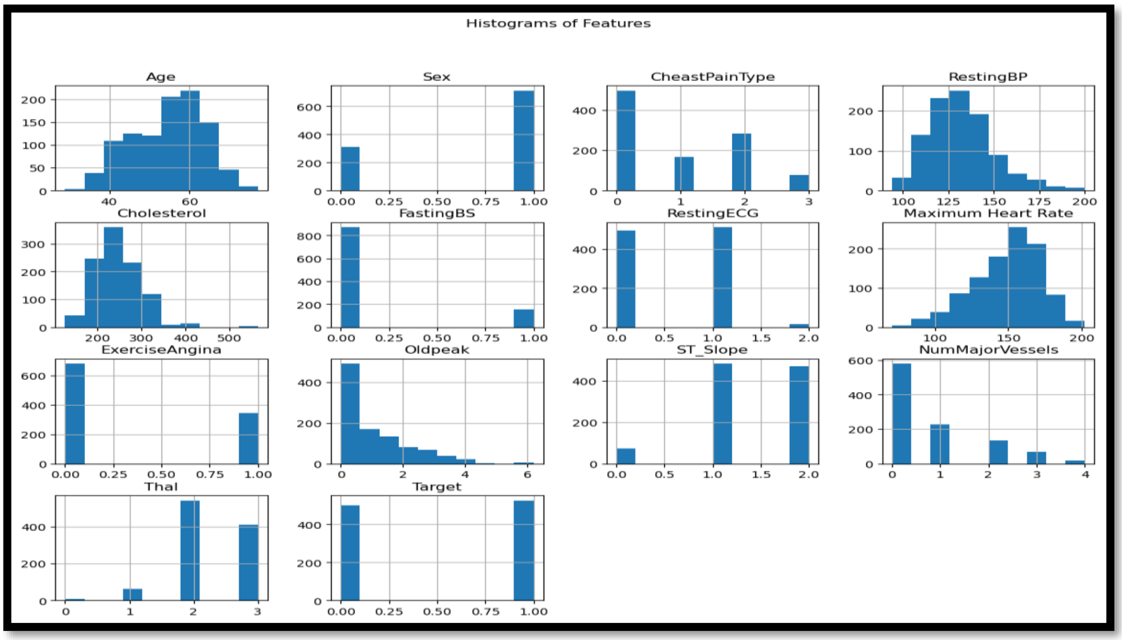


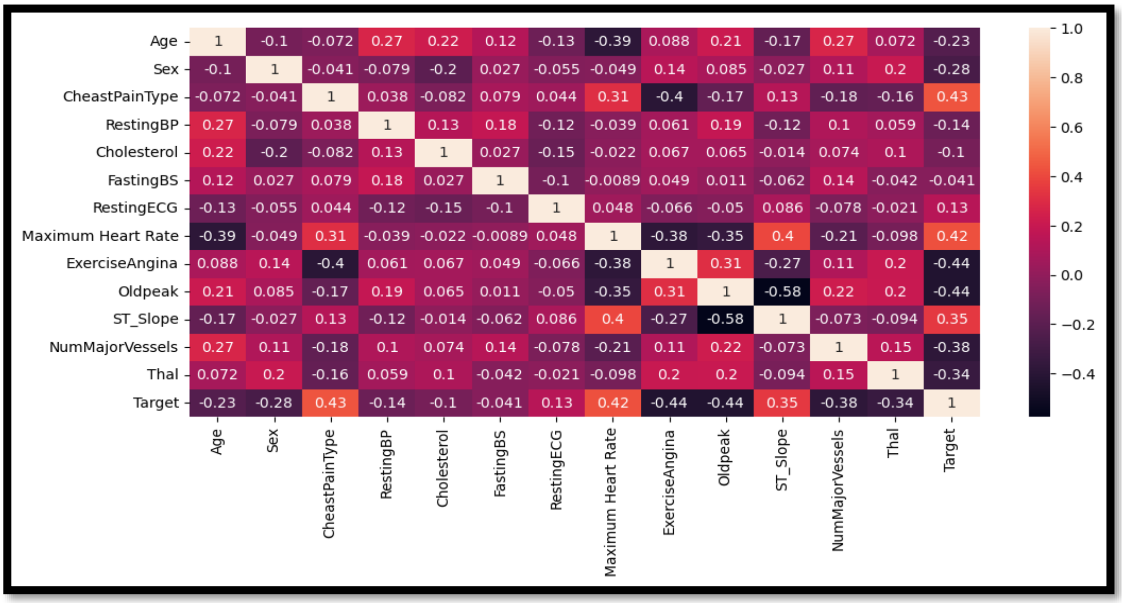
Table1

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1025 entries, 0 to 1024
Data columns (total 14 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Age                  1025 non-null   int64
 1   Sex                  1025 non-null   int64
 2   CheastPainType       1025 non-null   int64
 3   RestingBP            1025 non-null   int64
 4   Cholesterol          1025 non-null   int64
 5   FastingBS            1025 non-null   int64
 6   RestingECG           1025 non-null   int64
 7   Maximum Heart Rate   1025 non-null   int64
 8   ExerciseAngina       1025 non-null   int64
 9   Oldpeak              1025 non-null   float64
 10  ST_Slope             1025 non-null   int64
 11  NumMajorVessels      1025 non-null   int64
 12  Thal                 1025 non-null   int64
 13  Target               1025 non-null   int64
dtypes: float64(1), int64(13)
memory usage: 112.2 KB
```

The histogram displays the distribution of cardiovascular disease cases in relation to the clinical features.



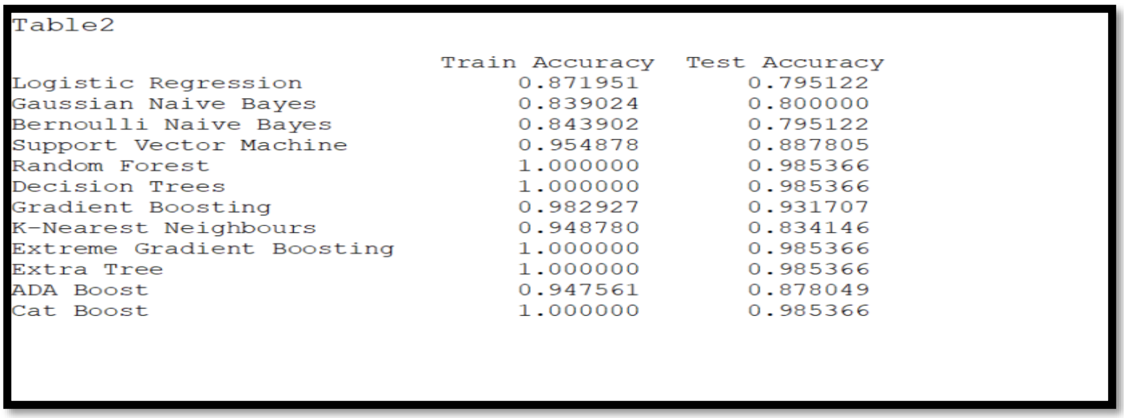
The heatmap has been utilized to explore the relationships among all clinical features:



6. Prediction of Cardiovascular Disease: We analysed the performance of twelve machine learning algorithms and assessed their potential as clinical decision support tools for predicting cardiovascular disease.

6.1. Examine the effectiveness of Machine Learning models:

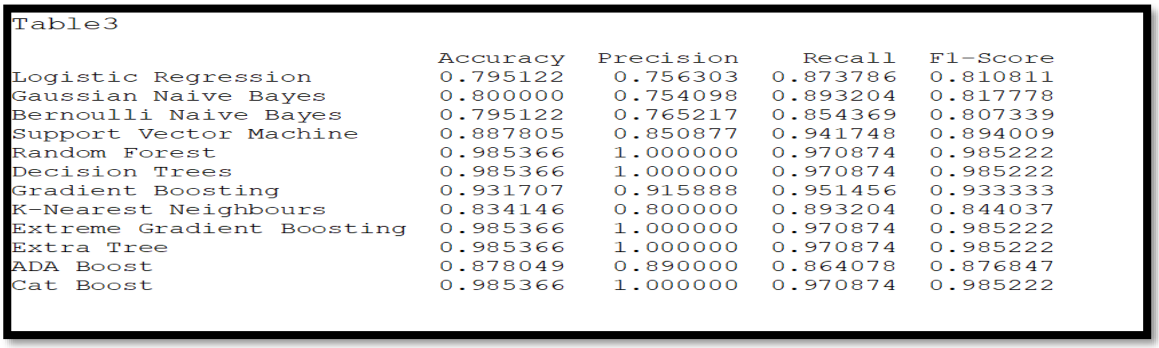
Divide the data into training, validation, and test sets to evaluate the model's performance. Standardize the data to ensure uniformity, which is essential for many ML algorithms. In the process of developing the model, we randomly selected clinical features from 80% of the patients for training purposes. Furthermore, we validated the model's predictive accuracy externally by testing it on an independent sample size, as detailed in Table 2.



```
Table2

                            Train Accuracy  Test Accuracy
Logistic Regression               0.871951       0.795122
Gaussian Naive Bayes              0.839024       0.800000
Bernoulli Naive Bayes             0.843902       0.795122
Support Vector Machine            0.954878       0.887805
Random Forest                     1.000000       0.985366
Decision Trees                    1.000000       0.985366
Gradient Boosting                 0.982927       0.931707
K-Nearest Neighbours              0.948780       0.834146
Extreme Gradient Boosting         1.000000       0.985366
Extra Tree                        1.000000       0.985366
ADA Boost                         0.947561       0.878049
Cat Boost                         1.000000       0.985366
```

6.2. Selection of Model for Predicting Cardiovascular Disease: The Random Forest, Decision Trees, XGBoost, Extra Tree and CatBoost models, emerged as the most effective for cardiovascular disease prediction, achieving an accuracy of 98.53%, precision of 1.00%, sensitivity of 97.08% and an F1 score of 98.52%. Table 3 provides a detailed comparison of accuracy, precision, recall, and F1 scores across all models.

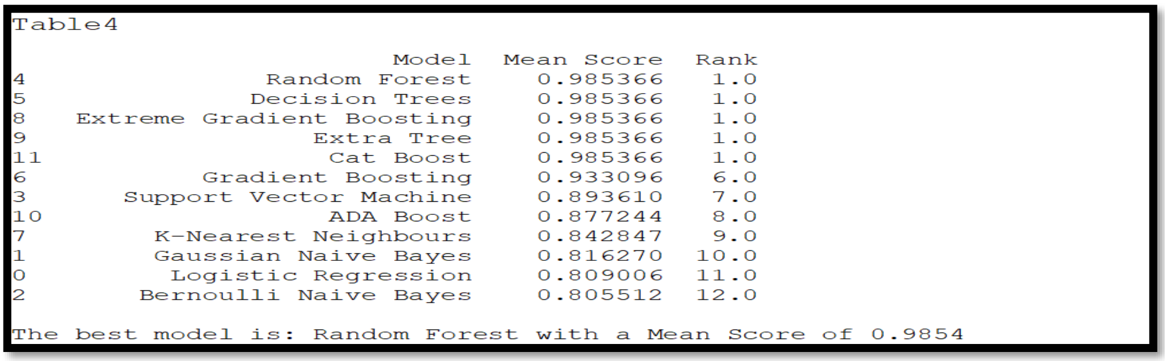


```
Table3

                           Accuracy  Precision    Recall  F1-Score
Logistic Regression        0.795122   0.756303  0.873786  0.810811
Gaussian Naive Bayes       0.800000   0.754098  0.893204  0.817778
Bernoulli Naive Bayes      0.795122   0.765217  0.854369  0.807339
Support Vector Machine     0.887805   0.850877  0.941748  0.894009
Random Forest              0.985366   1.000000  0.970874  0.985222
Decision Trees             0.985366   1.000000  0.970874  0.985222
Gradient Boosting          0.931707   0.915888  0.951456  0.933333
K-Nearest Neighbours       0.834146   0.800000  0.893204  0.844037
Extreme Gradient Boosting  0.985366   1.000000  0.970874  0.985222
Extra Tree                 0.985366   1.000000  0.970874  0.985222
ADA Boost                  0.878049   0.890000  0.864078  0.876847
Cat Boost                  0.985366   1.000000  0.970874  0.985222
```

6.3. Selecting the Optimal Model for Predicting Cardiovascular Disease:

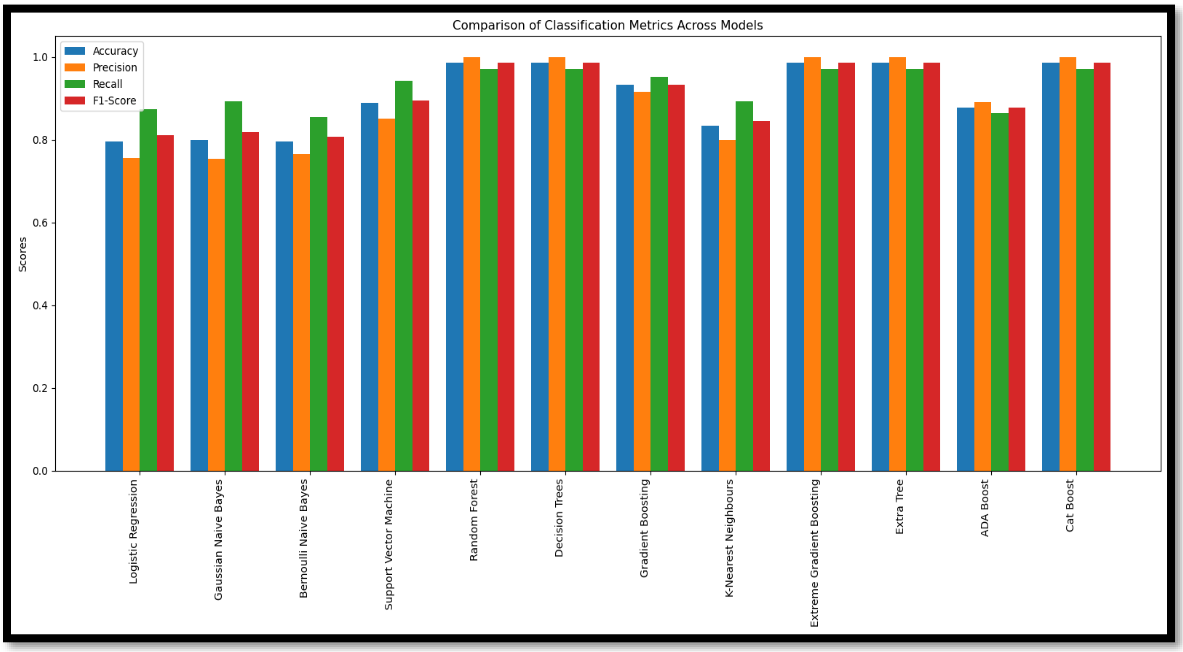
Twelve machine learning models were assessed for their predictive performance, and the highest-ranking model was chosen. Among these, the Random Forest model achieved the top rank. The rankings of all the models are provided below:

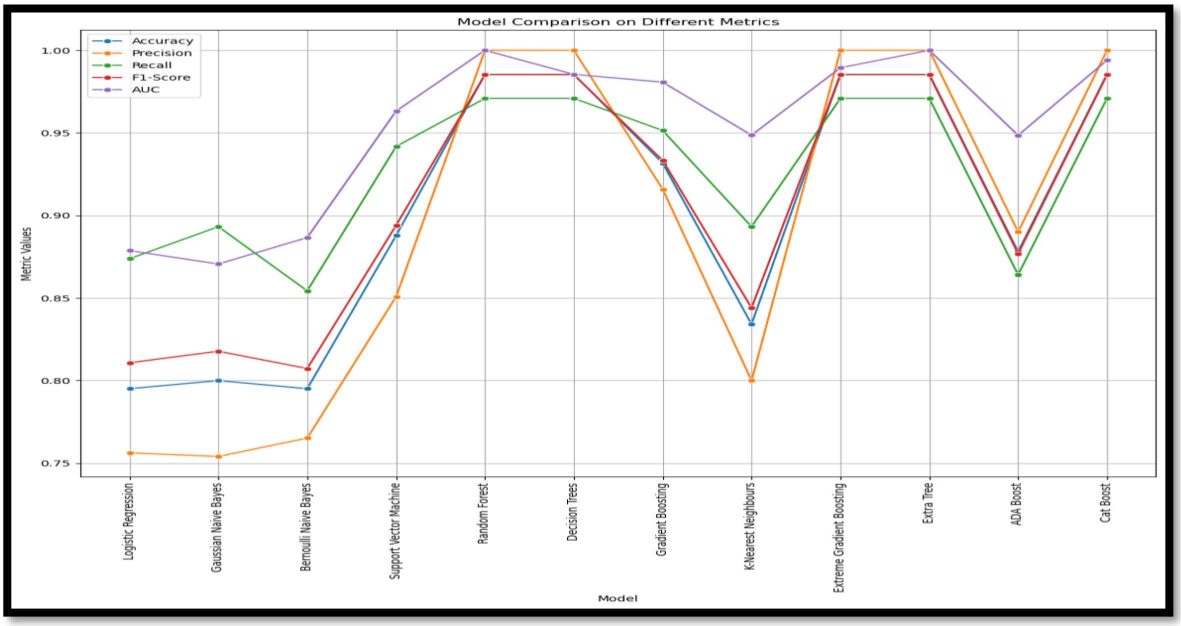


```
Table4
                        Model  Mean Score  Rank
4               Random Forest    0.985366   1.0
5              Decision Trees    0.985366   1.0
8   Extreme Gradient Boosting    0.985366   1.0
9                  Extra Tree    0.985366   1.0
11                  Cat Boost    0.985366   1.0
6           Gradient Boosting    0.933096   6.0
3      Support Vector Machine    0.893610   7.0
10                  ADA Boost    0.877244   8.0
7        K-Nearest Neighbours    0.842847   9.0
1        Gaussian Naive Bayes    0.816270  10.0
0         Logistic Regression    0.809006  11.0
2       Bernoulli Naive Bayes    0.805512  12.0

The best model is: Random Forest with a Mean Score of 0.9854
```

We also used graphs to visualize the data, providing a clearer insight into the models' performance:

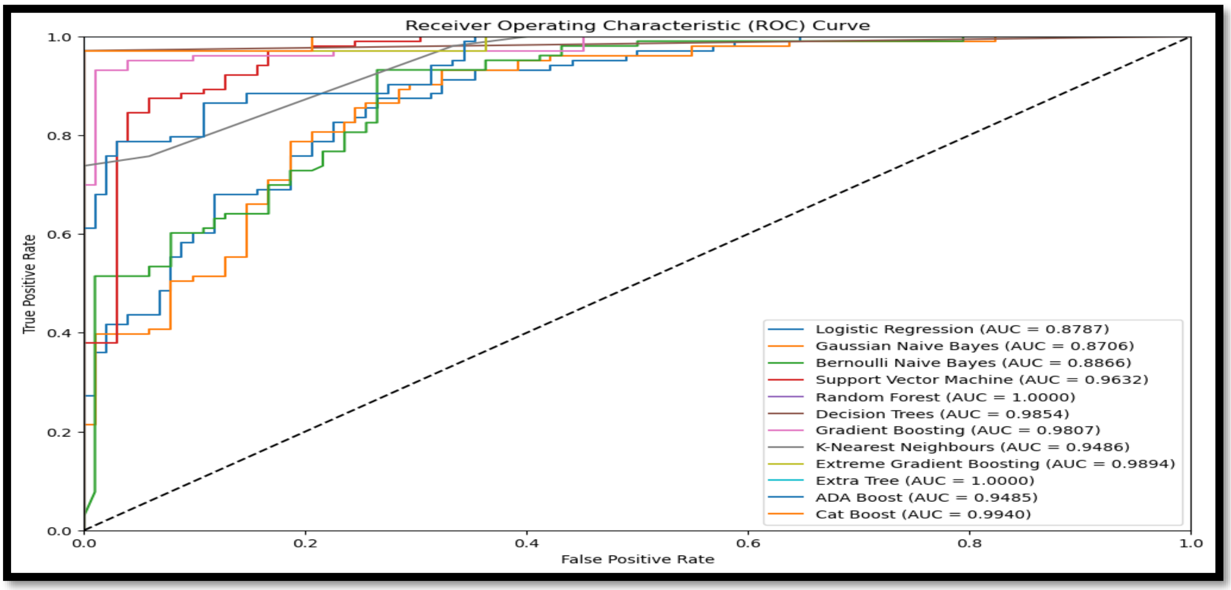




Cardiovascular disease prediction models can utilize Random Forest algorithm, offering fresh perspectives on the development of learning algorithm. Despite incorporating multiple base classifiers, the Random Forest model seldom overfits and effectively minimizes the exponential loss function by constructing a stepwise additive model.

7. Result:

The Random Forest model emerged as the best-performing model in terms of overall metrics, achieving the highest F1 score and cross-validation consistency, followed closely by Extra Tree, Decision Tree, XGBoost and CatBoost models.



This study focused on assessing various models built on different algorithms for cardiovascular disease prediction. The Random Forest model stood out, achieving remarkable performance with an accuracy of 98.53%, precision of 1.00%, sensitivity of 97.08% and an F1 score of 98.52% in predicting cardiovascular disease. This analysis underscores the importance of selecting high-performance machine learning models in predicting cardiovascular disease, with ensemble models proving most effective for this dataset.

8. Conclusion:

The utilization of sophisticated methods in machine learning represents a paradigm shift in the detection and management of heart issues. Through the fusion of computational prowess and medical expertise, researchers and practitioners have unlocked new avenues for early intervention, personalized treatment, and preventive care. The insights gleaned from vast datasets have empowered healthcare professionals to discern intricate patterns, identify subtle indicators, and predict potential cardiac events with unprecedented accuracy. However, amidst the promising advancements, several challenges persist, including data privacy concerns, model interpretability, and equitable access to technology-enabled healthcare solutions.

As we navigate the evolving landscape of machine learning in cardiovascular medicine, collaboration across interdisciplinary domains will be paramount. By fostering synergies between data scientists, clinicians, policymakers, and industry stakeholders, we can harness the full potential of these innovative technologies to combat heart diseases effectively. Ultimately, the integration of sophisticated machine learning methods into clinical practice holds the promise of saving lives, alleviating suffering, and ushering in a new era of cardiac care characterized by precision, efficiency, and compassion.

References:

[1] Tang, N., Do, C., Dinh, T. B., & Dinh, T. B. (2012). Urban traffic monitoring system. In D.-S. Huang, Y. Gan, P. Gupta, & M. M. Gromiha (Eds.), Advances in intelligent computing: Theories, applications, and aspects of artificial intelligence (pp. 573–580). Springer. https://doi.org/10.1007/978-3-642-25944-9_74

[2] Kannan, P. K., & Li, H. "Alice" (2017). Digital marketing: A framework, review, and research agenda. International Journal of Research in Marketing, 34(1), 22–45. https://doi.org/10.1016/j.ijresmar.2016.11.006

[3] Zakria, N., Raza, A., Liaquat, F., & Khawaja, S. G. (2017). Machine learning based analysis of cardiovascular disease prediction. Journal of Medical Systems, 41(12), 207. https://doi.org/10.1007/s10916-017-0842-6

[4] Bhatt, A., Dubey, S. K., Bhatt, A. K., & Joshi, M. (2017). Data mining approach to predict and analyze the cardiovascular disease. In S. C. Satapathy, V. Bhateja, S. K. Udgata, & P. K. Pattnaik (Eds.), Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications (pp. 117–126). Springer. https://doi.org/10.1007/978-981-10-3153-3_12

[5] Shinde, P. P., & Shah, S. (2018). A review of machine learning and deep learning applications. In 2018 Fourth International Conference on Computing, Communication, Control and Automation (ICCUBEA) (pp. 1–6). https://doi.org/10.1109/ICCUBEA.2018.8697857

[6] Zhang, S., Zhou, H., & Zhang, L. (2018). Recent machine learning progress in image analysis and understanding. Advances in Multimedia, 2018, 1–2. https://doi.org/10.1155/2018/1685890

[7] Ait Ali, N., Cherradi, B., El Abbassi, A., Bouattane, O., & Youssfi, M. (2018). GPU fuzzy c-means algorithm implementations: Performance analysis on medical image segmentation. Multimedia Tools and Applications, 77(21), 21221–21243. https://doi.org/10.1007/s11042-017-5589-6

[8] Dwivedi, A. K. (2018). Performance evaluation of different machine learning techniques for prediction of heart disease. Neural Computing and Applications, 29(2), 685–693. https://doi.org/10.1007/s00521-016-2604-1

[9] Amin, M. S., Chiam, Y. K., & Varathan, K. D. (2019). Identification of significant features and data mining techniques in predicting heart disease. Telematics and Informatics, 36, 82–93. https://doi.org/10.1016/j.tele.2018.11.007

[10] Tao, R., Zhang, S., Huang, X., et al. (2019). Magnetocardiograph-based ischemic heart disease detection and localization using machine learning methods. IEEE Transactions on Biomedical Engineering, 66(6), 1658–1667. https://doi.org/10.1109/TBME.2018.2832018

[11] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. IEEE Access, 7, Article ID 81542. https://doi.org/10.1109/ACCESS.2019.2901057

[12] Dornadula, V. N., & Geetha, S. (2019). Credit card fraud detection using machine learning algorithms. Procedia Computer Science, 165, 631–641. https://doi.org/10.1016/j.procs.2020.01.057

[13] Tiwari, U., Jain, M., & Mehfuz, S. (2019). Handwritten character recognition—an analysis. In S. N. Singh, F. Wen, & M. Jain (Eds.), Advances in Systems Optimization and Control (pp. 207–212). Springer. https://doi.org/10.1007/978-981-13-0665-5_18

[14] Ali, L., Rahman, A., Khan, A., Zhou, M., Javeed, A., & Khan, J. A. (2019). An automated diagnostic system for heart disease prediction based on Chi-square statistical model and optimally configured deep neural network. IEEE Access, 7, Article ID 34938. https://doi.org/10.1109/ACCESS.2019.2906157

[15] Garate Escamilla, A. K., Hajjam El Hassani, A., & Andres, E. (2019). A comparison of machine learning techniques to predict the risk of heart failure. In G. A. Tsihrintzis, M. Virvou, E. Sakkopoulos, & L. C. Jain (Eds.), Machine Learning Paradigms and Applications (pp. 9–26). Springer International Publishing. https://doi.org/10.1007/978-3-030-15628-2_2

[16] Fitriyani, N. L., Syafrudin, M., Alfian, G., & Rhee, J. (2020). HDPM: An effective heart disease prediction model for a clinical decision support system. IEEE Access, 8, Article ID 133034. https://doi.org/10.1109/ACCESS.2020.2991293

[17] Wang, J., Liu, C., Li, L., et al. (2020). A stacking-based model for non-invasive detection of coronary heart disease. IEEE Access, 8, Article ID 37124. https://doi.org/10.1109/ACCESS.2020.2979137

[18] Khan, M. A., & Algarni, F. (2020). A healthcare monitoring system for the diagnosis of heart disease in the IoMT cloud environment using MSSO-ANFIS. IEEE Access, 8, Article ID 122259. https://doi.org/10.1109/ACCESS.2020.2991200

[19] Mienye, I. D., Sun, Y., & Wang, Z. (2020). An improved ensemble learning approach for the prediction of heart disease risk. Information Medicine Unlocked, 20, 100402. https://doi.org/10.1016/j.imu.2020.100402

[20] Islam, Md. M., Haque, Md. R., Iqbal, H., Hasan, Md. M., Hasan, M. N., & Kabir, M. N. (2020). Breast cancer prediction: A comparative study using machine learning techniques. SN Computer Science, 1(1), 290. https://doi.org/10.1007/s42979-020-00305-w

[21] Spencer, R., Tabtah, F., Abdelhamid, N., & Tompson, M. (2020). Exploring feature selection and classification methods for predicting heart disease. Digital Health, 6, Article ID 2055207620914777. https://doi.org/10.1177/2055207620914777

[22] Sogancioglu, E., Murphy, K., Calli, E., Scholten, E. T., Schalekamp, S., & Van Ginneken, B. (2020). Cardiomegaly detection on chest radiographs: Segmentation versus classification. IEEE Access, 8, Article ID 94631. https://doi.org/10.1109/ACCESS.2020.2992098

[23] Abdeldjouad, F. Z., Brahami, M., & Matta, N. (2020). A hybrid approach for heart disease diagnosis and prediction using machine learning techniques. In M. Jmaiel, M. Mokhtari, B. Abdulrazak, H. Aloulou, & S. Kallel (Eds.), The Impact of Digital Technologies on Public Health in Developed and Developing Countries (pp. 123–135). Spr. International Publication. https://doi.org/10.1007/978-3-030-12376-3_12

[24] Terrada, O., Cherradi, B., Raihani, A., & Bouattane, O. (2020). Atherosclerosis disease prediction using supervised machine learning techniques. In 2020 1st International Conference on Innovative Research Applications in Science, Engineering, and Technology (IRASET) (pp. 1–5). https://doi.org/10.1109/IRASET48871.2020.9092082

[25] Terrada, O., Cherradi, B., Hamida, S., Raihani, A., Moujahid, H., & Bouattane, O. (2020). Prediction of patients with heart disease using artificial neural network and adaptive boosting techniques. In 2020 3rd International Conference on Advanced Communication Technologies and Networking (CommNet) (pp. 1–6). https://doi.org/10.1109/CommNet49926.2020.9199620

[26] Tougui, I., Jilbab, A., & El Mhamdi, J. (2020). Heart disease classification using data mining tools and machine learning techniques. Health Technology, 10, 1137–1144. https://doi.org/10.1007/s12553-020-00438-1

[27] Alom, Z., et al. (2021). Early stage detection of heart failure using machine learning techniques. In Proceedings of the International Conference on Big Data, IoT, and Machine Learning (pp. 23–25). Cox's Bazar, Bangladesh.

[28] Shameer, K., et al. (2021). Machine learning predictions of cardiovascular disease risk in a multi-ethnic population using electronic health record data. International Journal of Medical Informatics, 146, 104335. https://doi.org/10.1016/j.ijmedinf.2021.104335

[29] Collins, C., Dennehy, D., Conboy, K., & Mikalef, P. (2021). Artificial intelligence in information systems research: A systematic literature review and research agenda. International Journal of Information Management, 60, 102383. https://doi.org/10.1016/j.ijinfomgt.2021.102383

[30] Bachute, M. R., & Subhedar, J. M. (2021). Autonomous driving architectures: Insights of machine learning and deep learning algorithms. Machine Learning and Applications, 6, 100164. https://doi.org/10.1016/j.mlwa.2021.100164

[31] Mienye, I. D., & Sun, Y. (2021). Improved heart disease prediction using particle swarm optimization based stacked sparse autoencoder. Electronics, 10(19), 2347.

[32] Zhenya, Q., & Zhang, Z. (2021). A hybrid cost-sensitive ensemble for heart disease prediction. BMC Medical Informatics and Decision Making, 21(1), 73.

[33] Shorewala, V. (2021). Early detection of coronary heart disease using ensemble techniques. Information Medicine Unlocked, 26, 100655. https://doi.org/10.1016/j.imu.2021.100655

[34] Patro, S. P., Nayak, G. S., & Padhy, N. (2021). Heart disease prediction by using novel optimization algorithm: A supervised learning prospective. Information Medicine Unlocked, 26, 100696. https://doi.org/10.1016/j.imu.2021.100696

[35] Tyagi, A., & Mehra, R. (2021). Intellectual heartbeats classification model for diagnosis of heart disease from ECG signal using hybrid convolutional neural network with Goa. SN Applied Sciences, 3(2), 265.

[36] Gour, S., Panwar, P., Dwivedi, D., & Mali, C. (2022). A machine learning approach for heart attack prediction. In A. K. Nagar, D. S. Jat, G. Marín-Raventós, & D. K. Mishra (Eds.), Intelligent Sustainable Systems (pp. 741–747). Springer Singapore. https://doi.org/10.1007/978-981-16-6309-3_70

[37] Gupta, C., Saha, A., Reddy, N. S., & Acharya, U. D. (2022). Cardiac disease prediction using supervised machine learning techniques. Journal of Physics: Conference Series, 2161, 012013.

[38] Ait Ouallane, A., Bakali, A., Bahnasse, A., Broumi, S., & Talea, M. (2022). Fusion of engineering insights and emerging trends: Intelligent urban traffic management system. Information Fusion, 88, 218–248. https://doi.org/10.1016/j.inffus.2022.07.020

[39] World Health Organization. (2023). Cardiovascular diseases (CVDs). Available online: https://www.afro.who.int/health-topics/cardiovascular-diseases (accessed on 5 May 2023).

[40] Denysyuk, H. V., Pinto, R. J., Silva, P. M., Duarte, R. P., Marinho, F. A., Pimenta, L., Gouveia, A. J., Gonçalves, N. J., Coelho, P. J., Zdravevski, E., Lameski, P., Leithardt, V., Garcia, N. M., Pires, I. M. (2023). Algorithms for automated diagnosis of cardiovascular diseases based on ECG data: A comprehensive systematic review. Heliyon, 9, e13601. https://doi.org/10.1016/j.heliyon.2023.e13601

[41] Basak, A., Schmidt, K. M., & Mengshoel, O. J. (2023). From data to interpretable models: Machine learning for soil moisture forecasting. International Journal of Data Science and Analytics, 15, 9–32. https://doi.org/10.1007/s41060-022-00

[42] Wang, T., Guo, H., Ge, Z., Zhang, Q., & Yang, Z. (2023). An MMSE graph spectral magnitude estimator for speech signals residing on an undirected multiple graph. EURASIP Journal on Audio, Speech, and Music Processing, 7. https://doi.org/10.1186/s13636-023-00272-z

[43] Huang, S., Xu, J. T., & Yang, M. (2023). Review: Predictive approaches to breast cancer risk. Heliyon, 9, e21344. https://doi.org/10.1016/j.heliyon.2023.e21344

[44] Singh, D. P. (2024). An extensive examination of machine learning methods for identifying diabetes. Tuijin Jishu/Journal of Propulsion Technology, 45(2).

[45] Singh, D. P. (2024). An extensive analysis of machine learning models to predict the breast cancer recurrence. Tuijin Jishu/Journal of Propulsion Technology, 45(2).

[46] Singh, D. P. (2024). A notable utilization of machine learning techniques in the healthcare sector for optimizing resources and enhancing operational efficiency. European Journal of Biomedical and Pharmaceutical Sciences, 11(7), 212-224. http://www.ejbps.com

[47] Singh, D. P. (2024). An extensive analysis of machine learning techniques for predicting the onset of lung cancer. Tuijin Jishu/Journal of Propulsion Technology, 45(4).