Scientific Research Publishing

Open Journal of Statistics, 2024, 14, 228-242
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Advanced Computing for Cardiovascular Disease Prediction

Santosh Gaire[1,2], Poshan Belbase[2*], Amrit Kafle[1,2], Rajendra Bhandari[1,2]

[1]Vitreous State Laboratory, The Catholic University of America, Washington DC, USA
[2]Department of Physics, The Catholic University of America, Washington DC, USA
Email: *belbase@cua.edu

How to cite this paper: Gaire, S., Belbase, P., Kafle, A. and Bhandari, R. (2024) Advanced Computing for Cardiovascular Disease Prediction. *Open Journal of Statistics*, **14**, 228-242.
https://doi.org/10.4236/ojs.2024.143011

Received: April 18, 2024
Accepted: May 27, 2024
Published: May 30, 2024

Abstract

Developing a predictive model for detecting cardiovascular diseases (CVDs) is crucial due to its high global fatality rate. With the advancements in artificial intelligence, the availability of large-scale data, and increased access to computational capability, it is feasible to create robust models that can detect CVDs with high precision. This study aims to provide a promising method for early diagnosis by employing various machine learning and deep learning techniques, including logistic regression, decision trees, random forest classifier, extreme gradient boosting (XGBoost), and a sequential model from Keras. Our evaluation identifies the random forest classifier as the most effective model, achieving an accuracy of 0.91, surpassing other machine learning and deep learning approaches. Close behind are XGBoost (accuracy: 0.90), decision tree (accuracy: 0.86), and logistic regression (accuracy: 0.70). Additionally, our deep learning sequential model demonstrates promising classification performance, with an accuracy of 0.80 and a loss of 0.425 on the validation set. These findings underscore the potential of machine learning and deep learning methodologies in advancing cardiovascular disease prediction and management strategies.

Keywords

Machine Learning, Deep Learning, Classification, Performance Matrix, Accuracy

1. Introduction

The proper functioning of the cardiovascular system ensures a healthy heart, which is the most essential aspect of the well-being of our body. The problems that arise in the system cause cardiovascular diseases (CVDs). According to National Health Services, UK, cardiovascular disease is a general term for the conditions affecting the heart or blood vessels. Generally, CVDs encompass all types of diseases that affect the heart or blood vessels, including coronary heart disease, cerebrovascular disease, rheumatic heart disease, and other conditions like chest pain, stroke, and heart attack [1]. The effects of behavioral risk factors for CVDs like unhealthy diet, physical inactivity, and consumption of tobacco and alcohol [2], may appear as high blood pressure, blood glucose, raised blood lipids, overweight, and obesity in an individual. [3]. Most patients experience shortness of breath, arm, shoulder, and chest pain along with an overall feeling of weakness. These symptoms increase the risk of stroke, angina, and heart attack due to restricted or clogged blood vessels, which primarily cause the untimely death of patients [4].

The World Health Organization reports that 32 percent of the 17 million deaths globally in 2019 that were related to non-communicable diseases were caused by CVDs [5]. More than 75 percent deaths from the CVDs occur in least-developed countries and it badly affects the mid- and low-income people [6]. More than four out of five CVD deaths are due to heart attacks and strokes, and one-third of these deaths occur prematurely in people under 70 years of age [7]. However, clinical decision-making during diagnosis and treatment is complex, and cardiologists face difficulties in detecting and treating patients in the early stages [8]. Early and accurate detection and diagnosis of CVDs is a must to provide appropriate treatments to the patients, which helps to prevent the premature death of the person [9].

The diagnosis and treatment of CVDs rely on data in several forms, such as patient history, physical examination, laboratory data, and invasive and non-invasive imaging techniques [10]. Invasive imaging techniques such as cardiac catheterization and intravascular ultrasound are associated with more risk factors and require a unique hospital setting [11]. Angiography is more dependable among other non-invasive imaging techniques like X-ray and magnetic resonance imaging (MRI), however, it requires solid technological knowledge and has side effects [12] [13] [14]. Such conventional methods are time-consuming and expensive, making detecting CVDs more complicated than it needs to be. On the other hand, machine learning (ML) can solve this issue by enabling an automatic method to assess examples and draw consistent and accurate conclusions.

A higher degree of accuracy in predicting the risk of CVDs can be achieved by appropriately processing the data mined with various ML algorithms to identify patterns, trends, and relationships between distinct parameters. Classification models, for example, make it possible to design more individualized and efficient treatment plans, improving patient care [15]. This study examines a comparative computational approach, using supervised classification machine learning to forecast cardiovascular diseases. The research framework, outlined in **Figure 1**, progresses from basic to advanced models, including artificial neural networks (ANN). Model selection is based on performance evaluation metrics including sensitivity, precision, F1 score, accuracy, and the area under the Receiver Operating Characteristic (ROC) curve (AUC). Our work is organized into

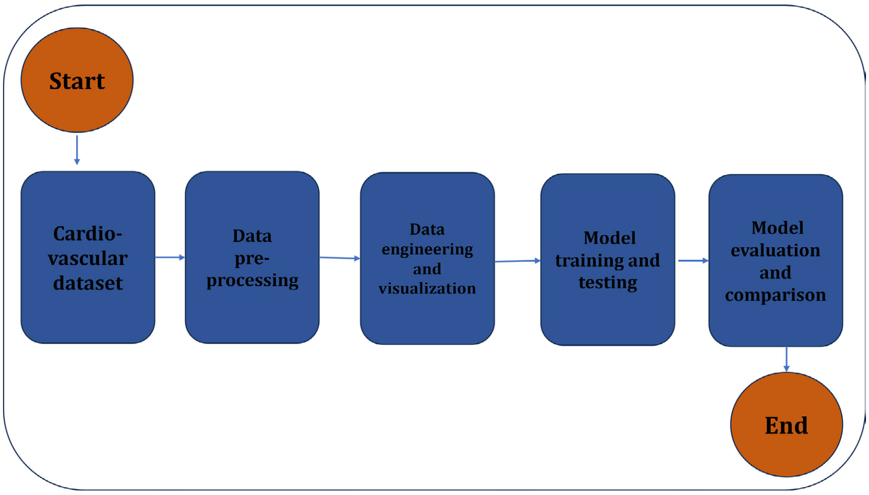
Start
Cardio-vascular dataset
Data pre-processing
Data engineering and visualization
Model training and testing
Model evaluation and comparison
End

Figure 1. Schematic diagram of working procedure.

several sections. Chapter 2 summarizes key findings from relevant literature. Chapter 3 delves into our methodology, encompassing discussions on data structure, data engineering, data visualization, and concise descriptions of the models employed. Furthermore, Chapter 4 presents the results, followed by a detailed discussion that leads to the conclusive findings in Chapter 5.

2. Literature Review

Numerous studies have delved into machine learning methods for forecasting CVDs. The findings from these investigations consistently demonstrate the capability of machine learning to accurately predict CVDs. Here we review a few literature and compare the results that have been obtained using different models.

Degroat *et al.* [16] combined classical statistical methods with advanced Machine Learning (ML) algorithms to improve disease prediction in CVD patients. They proposed four feature selection algorithms: Chi-Square Test, Pearson Correlation, Recursive Feature Elimination (RFE), and Analysis of Variance (ANOVA). Using these methods, they identified 18 transcriptome biomarkers in the CVD population with up to 96% predictive accuracy. The study included 61 CVD patients and 10 healthy individuals as controls.

In a study published in 2020, Drożdż *et al.* [17] employed machine learning, utilizing liver ultrasonography and biochemical analysis in 191 CVD patients, revealing the association between metabolic-associated fatty liver disease (MAFLD) and CVD risk factors. They utilized techniques like principle component analysis (PCA) and logistic regression to construct a predictive model for high CVD risk, focusing on diabetes duration, plaque scores, and hypercholesterolemia. Evaluation via receiver operating characteristic (ROC) curves yielded Area Under the ROC Curve (AUCs) ranging from 0.84 to 0.87. The optimal model, utilizing five variables, accurately detected 85.11% of high-risk and 79.17% of low-risk patients. These findings emphasize ML's utility in identifying MAFLD patients at CVD risk based on readily available patient data.

Ambekar *et al.*'s paper [18] suggested employing unimodel disease risk algorithms based on convolutional neural networks to estimate a patient's risk level based on the heart disease dataset, *i.e.*, high or low. They carry out data imputation and cleaning procedures to turn unstructured data into structured data. Later on, the KNN and naïve bayes algorithm is applied to the input values and heart disease is predicted based on this information. They contrast the outcomes of the KNN and Naïve Bayes algorithms, noting that NB has an accuracy of 82%, which is higher than that of the KNN algorithm. They were able to forecast disease risk with about 65% accuracy because of the organized dataset.

Larroza *et al.* [19] employed machine learning and MRI texture features to differentiate between acute myocardial infarction (AMI) and chronic myocardial infarction (CMI). Analyzing 44 cases (22 AMI, 22 CMI) with cine and late gadolinium enhancement (LGE) MRI, 279 texture features were extracted from infarcted areas on LGE and the entire myocardium on cine. Classification was performed using three prediction models: random forest, SVM with Gaussian kernel, and SVM with polynomial kernel. The study demonstrated that texture analysis when paired with machine learning, may effectively differentiate between AMI and CMI on both LGE and cine MRI. The SVM with a polynomial kernel showed the best performance, achieving AUC of 0.86 ± 0.06 on LGE MRI (72 features) and 0.82 ± 0.06 on cine MRI (75 features).

Oyewola *et al.* [20] compare different algorithms between long short-term memory (LSTM), feedforward neural network (FFNN), cascade forward neural network (CFNN), Elman Neural Network (ELMAN), and ensemble deep learning (EDL) to predict the best model using the Kaggle cardiovascular dataset with 12 attributes and 70,000 patients. The EDL model surpasses other algorithms with an incredible 98.45% accuracy. After more investigation, EDL's 100% classification accuracy for CVD diagnosis was shown to exceed LSTM, FFNN, CFNN, and ELMAN. Overall, the study suggests that EDL model could be a robust tool for the early detection of CVDs.

Alaa *et al.* [21] analyzed 437 characteristics of 423,604 individuals in the UK Biobank who did not have CVDs at baseline. The AutoPrognosis model outperformed established techniques such as the Framingham score area under the receiver operating characteristic curve (AUCROC: 0.724, 95%), Cox proportional hazards models with conventional risk factors (AUCROC: 0.734, 95%), and Cox proportional hazards with all UK Biobank variables (AUCROC: 0.758, 95%) in terms of risk prediction (AUCROC: 0.774, 95%). AutoPrognosis correctly identified 368 more occurrences of CVDs after 5 years than the Framingham score. They also emphasized how the addition of more risk variables outweighed than use of complex models.

3. Methodology

3.1. Data Engineering

The dataset analyzed in this study was sourced from the renowned data science community, Kaggle, specifically identified as the cardiovascular risk prediction dataset within Kaggle's repository. This dataset comes from the Behavioral Risk Factor Surveillance System (BRFSS), which is known as the nation's top system for health-related telephone surveys. It contains a thorough collection of health-related information. Comprising 19 carefully selected variables and spanning 308,854 rows, the dataset encapsulates various aspects of an individual's lifestyle that may contribute to their susceptibility to cardiovascular diseases. Through meticulous curation and analysis, this dataset offers valuable insights into the intricate interplay between lifestyle factors and cardiovascular health, facilitating informed decision-making and intervention strategies in public health initiatives.

The dataset contained 19 variables, with 12 being categorical and the rest being of float or integer type. Some of the categorical columns included lengthy string names such as green vegetable consumption or Fried potato consumption, which could be challenging to handle. Therefore, we opted to convert them into a more precise format using the renaming feature in pandas.

To enhance data clarity, we have restructured some features in our dataset. One significant change is seen in the Body Mass Index (BMI) and Age categories. Initially, BMI was represented as individual values but is now categorized into standard groups. Similarly, Age, initially in-class format, has been transformed into categories for better understanding. The rationale behind this modification, elucidated in detail in **Table 1**, facilitates clearer communication.

3.2. Data Visualization

Our scope encompasses not merely the focus on model performance, but also the illumination of critical insights extracted from our data analysis through visualization techniques. Employing methods such as heat map visualization, we try to understand the important connection between heart disease and key features. In the subsequent analysis, we present various variables with instances of heart disease, elucidating their significance and implications.

Table 1. Categories of age and BMI.

BMI Index		Age category	
BMI Index range	BMI category	Age class	Age category
12.02 - 18.5	Under weight	18 - 24	Young
18.5 - 25	healthy weight	25 - 39	Adult
25 - 30	Over weight	40 - 54	Mid-aged
30 - 35	Obese-I	55 - 64	Senior-Adult
35 - 40	Obese-II	65 - 79	Elderly
40 above	severe Obesity	80+	Elderly

In Figure 2, we have highlighted some important factors closely tied to heart disease using histograms. Figure 2(a), shows how heart disease relates to people's body weights. For example, among those with a healthy weight (BMI: 18.5 - 25), only about 6% have heart disease. But among those categorized as overweight or obese, the rates are higher: around 15% for overweight, 22% for moderately obese, and approximately 25% for severely obese individuals. For the corresponding BMI ranges please refer to Table 1.

In addition, Figure 2(b) elucidates the relationship between smoking and the prevalence of cardiovascular disease. In terms of percentage, heart disease affects approximately 11.6% of smokers while it affects only 5.6% of non-smokers. This clearly indicates that smokers are more likely to get affected by heart disease. Figure 2(c) and Figure 2(d), show that individuals with diabetes and arthritis are more susceptible to heart disease on their own. Based on statistical data, around 20.85% of individuals with diabetes have heart disease, whereas only 6.06% of non-diabetics have the condition. Furthermore, an examination of the arthritis data shows that 14.09% of those with arthritis also had heart disease, compared to 5.43% of those without arthritis.

Figure 3 illustrates the crucial role of regular exercise and checkups in mitigating the risk of heart disease. In Figure 3(a), we observe a comparison of heart disease cases relative to regular exercise habits. The data indicates that 8% of individuals who do not engage in regular exercise are afflicted with heart disease, compared to only 3% among those who exercise regularly. Similarly, in Figure 3(b), we analyze heart disease cases in relation to regular checkup habits. The findings suggest that individuals who undergo regular checkups are at a lower risk of developing heart disease compared to those who schedule checkups every two years, every five years, or never. These insights underscore the significance of maintaining both a regular exercise routine and a consistent checkup schedule in reducing the likelihood of heart disease or facilitating early detection.

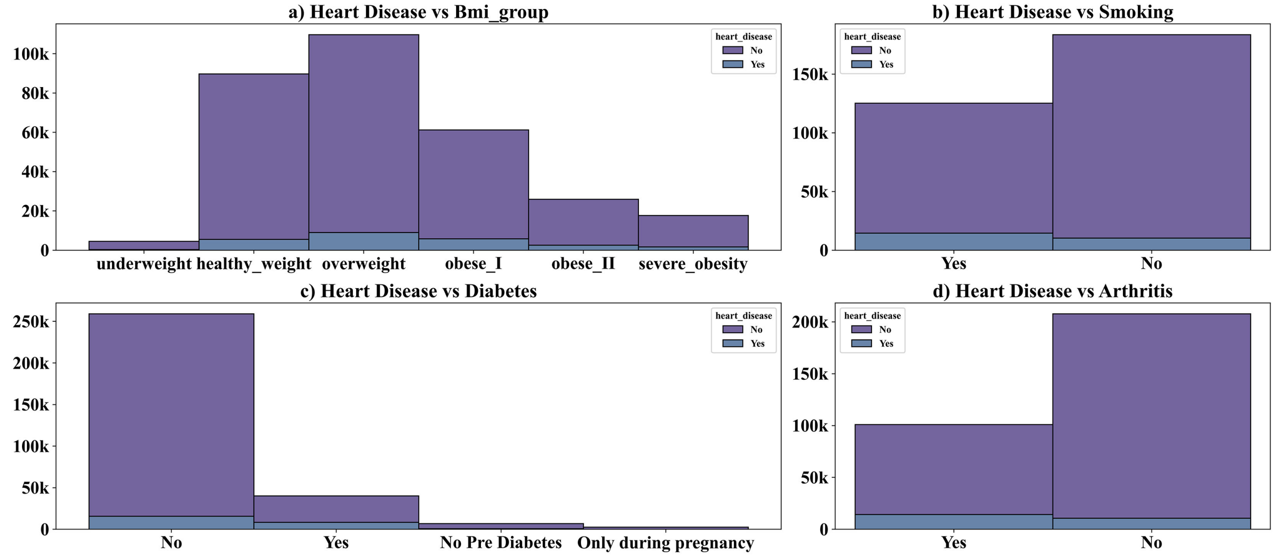

Figure 2. A visual display illustrating how our target variable is related to different key features.

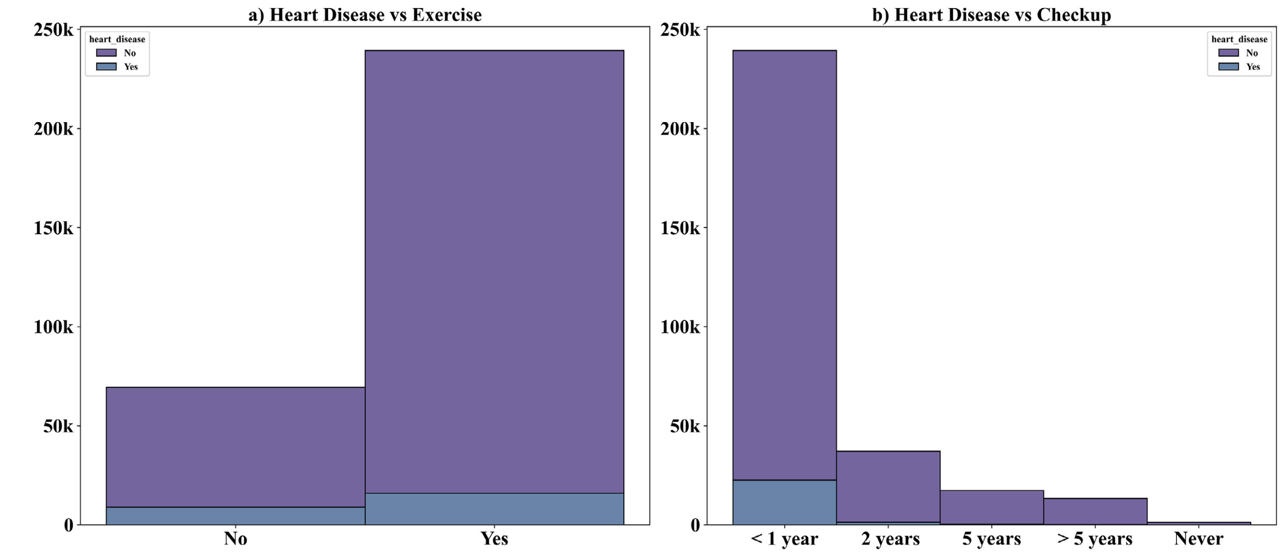

Figure 3. Histograms depicting the inverse correlation between specific features and the target variable.

3.3. Used Models

This study employs four supervised machine learning algorithms and one deep learning algorithm for training, validating, and testing the dataset. Below, we provide a brief overview of each model used.

3.3.1. Logistic Regression

The method of modeling the probability of a discrete result given input variables is known as logistic regression. The most frequent logistic regression models include a binary result, which can accept two values like true/false, yes/no scenarios. Multinomial logistic regression can be used to model events with more than two discrete outcomes. Despite its name, logistic regression is a classification model rather than a regression model. For binary and linear classification problems, logistic regression is a simpler and more efficient method [22]. Furthermore, logistic regression, unlike linear regression, does not require a linear connection between input and output variables [23]. In logistic regression, the goal is to model the probability that a given instance belongs to a particular class (positive or negative). The logistic regression model uses a linear combination of input features, transformed by a sigmoid function. The linear combination is represented as:

$$z = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + \cdots + b_k \cdot x_k \tag{1}$$

where b_0 is the bias term and $b_1, b_2, \cdots, b_k$ are the coefficients associated with the features $x_1, x_2, \cdots, x_k$. The probability of belonging to the positive or negative class is then given by the sigmoid function:

$$P(Y = 1) = \frac{1}{1 + \mathrm{e}^{-z}} \tag{2}$$

The logistic regression model predicts the class based on whether the probability is above a certain threshold (typically 0.5) [22].

3.3.2. Decision Tree

A well-known machine learning technique called a decision tree divides data into multiple groups according to predetermined criteria. Nodes and leaves are the two navigable components of the tree. Decision nodes divide data, whereas leaves represent choices or results. Combining decision trees with other techniques can help solve problems (ensemble learning). Using fundamental decision rules generated from data properties, the objective is to construct a model that predicts the value of a target variable. A piecewise constant approximation can be seen in a decision tree [24] [25]. Assume that samples with quantities for the categorical attribute "*A*" that have "*n*" different potential values make up training dataset "*D*". The parameter and information obtained for attribute "*A*" can be obtained using the following formula:

$$Gain(A,D) = Entropy(D) - \sum_{i=1}^{n} \frac{|D_i|}{|D|} \cdot Entropy(D_i) \tag{3}$$

Here, D_i represents the subset of instances in D where attribute A has the *i*-th value. $|D_i|$ is the number of instances in D_i. $Entropy(D)$ is the entropy of the dataset *D*. $Entropy(D_i)$ is the entropy of the *i*-th subset D_i [26].

3.3.3. Random Forest

Random forest is a popular machine learning technique that combines the output of numerous decision trees to produce a single conclusion. Its ease of use and flexibility, as well as its ability to tackle classification and regression challenges, has boosted its popularity [26] [27]. The random forest algorithm is a bagging method extension that uses both bagging and feature randomness to produce an uncorrelated forest of decision trees. The forecast determination will differ depending on the type of difficulty. Individual decision trees will be averaged for a regression task, and a majority vote—*i.e.* the most common categorical variable—will produce the predicted class for a classification problem. Finally, the odd sample is used for cross-validation, which completes the prediction [23]. In order to minimize overfitting and enhance the model's capacity for generalization, randomization is incorporated into the feature selection process as well as the data selection process. The Random Forest technique is made more effective overall by the Gini impurity, which is used as a criterion for dividing nodes in the decision trees. The Gini impurity for a set with *n* classes is calculated using the formula:

$$\text{Gini} = 1 - \sum_{i=1}^{n} (p_i)^2 \tag{4}$$

where p_i is the probability of an object being classified into the *i*-th class. In the context of decision trees and random forests, gini impurity is used as a criterion to evaluate the purity of a node. The algorithm aims to split nodes in a way that minimizes the gini impurity [26] [27].

3.3.4. Extreme Gradient Boosting

Extreme gradient boosting (XGBoost) is a type of ensemble machine learning method that may be used to solve predictive modeling tasks like classification and regression. It is extremely effective as well as computationally efficient [28]. Using an ensemble approach called "boosting," errors produced by previous models are corrected by adding new models. To minimize the loss when adding new models, it makes use of a gradient descent approach [29]. Consider a dataset (X_i, Y_i) having M features and N records. T predict the best output $\hat{Y}$, we need the best set of function to minimize the overall loss such that,

$$\ell(\phi) = \sum_i L(Y_i, \hat{Y}_i) + \sum_k \Omega(f_k) \tag{5}$$

The loss function, denoted by $L(Y_i, \hat{Y}_i)$, is the difference between the actual output (Y_i) and the projected output $\hat{Y}_i$. Where $\sum_k \Omega(f_k)$ indicates the complexity of the model, this helps prevent the model from being overfit [30].

3.3.5. Sequential Model

A sequential model is a fundamental type of model used to construct neural networks layer by layer, following a sequential order. In this project, a sequential model is composed of several layers interconnected within the widely-used deep learning framework, Keras. Deep learning, a subset of machine learning, utilizes artificial neural networks (ANNs) which are computer algorithms inspired by the biological functioning of the human brain in processing information. Instead of learning through programming, ANNs are trained through experience by looking for patterns and relationships in data [31] [32]. **Figure 4** displays the model's flow. The dense layer, which is the fully connected layer, receives the characteristics, activates using Relu function and passes it forward. Other dense layer gives the output with the help of sigmoid activation function [33].

3.4. Hyper-Parameter Tuning

In this study, we did not explicitly tune or use any specific hyperparameters for machine-learning models except for deep learning. The deep learning sequential model underwent careful tuning for optimal performance. This included setting the learning rate to 0.001 for the Adam optimizer, utilizing a first Dense layer with 256 neurons and a rectified linear unit (ReLU) activation function, and employing a sigmoid activation function in the output layer for classification tasks. Additionally, we implemented an early stopping mechanism with a patience of 3 epochs, monitoring validation loss and restoring the best weights when training stops, ensuring the model's robustness and generalization ability.

3.5. Model Evaluation

We assess the developed machine learning model using performance metrics like precision, recall, F1 score, accuracy, ROC curve and AUC. Additionally, we evaluate the deep learning model for predicting cardiovascular disease based on accuracy and loss curves. The essential criteria for evaluating machine learning models are derived from the components of the confusion matrix. **Table 2** outlines the analytical structure of a confusion matrix.

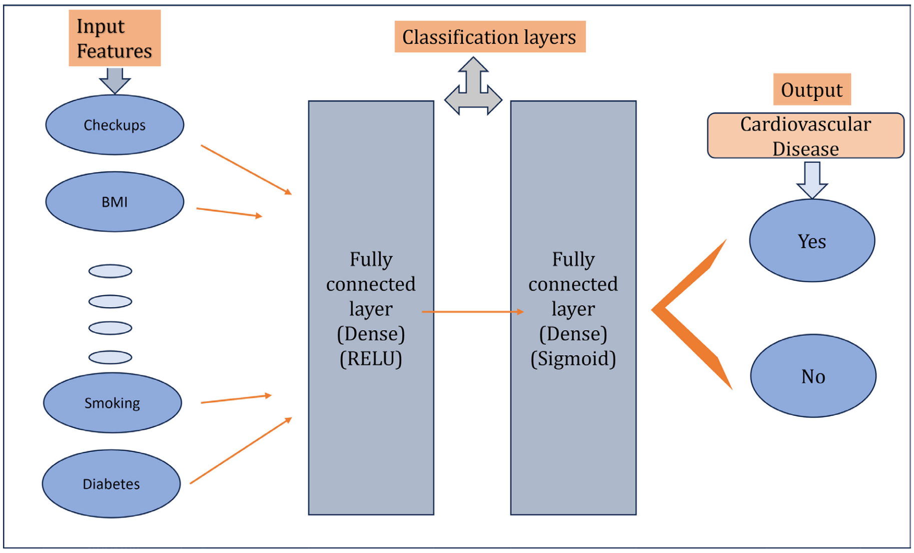

Figure 4. Working of sequential with Keras.

Table 2. Confusion matrix for the evaluation of machine learning models.

Confusion Matrix		
	Actual positive	Actual negative
Predictive positive	True positive (TP)	False positive (FP)
Predictive negative	False negative (FN)	True negative (TN)

The precision, recall (sensitivity) and F1 scores are calculated with the help of confusion matrix using following mathematical expressions:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

$$\text{F1 score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

And, the accuracy is the percentage of cases that are correctly anticipated (forecasted negative for patients without CVD and correctly predicted positive for individuals with CVD) and is mathematically represented as,

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{9}$$

Along with this, model's performance is measured with the help of AUC. Its value is in the range of 0 and 1. If a model's AUC is near to 1, it is considered good. The model is better if the AUC is greater, and vice versa.

As previously stated, the deep learning sequential model is validated using accuracy and loss curve. Out of all the cases, the percentage of correctly categorized instances by the model is known as accuracy. Consequently, the accuracy curve enhances the model's capacity to produce precise predictions by providing insight into how well the model matches the data. The loss curve, on the other hand, measures the inaccuracy or dissimilarity between the model's anticipated output and the true parameter and provides us with information about how the model performs over time. The loss shows how much the actual values deviate from the model's predictions. The model attempts to approximate the true values as closely as possible by minimizing the loss. Thus, the loss curve illustrates how the model's inaccuracy diminishes with learning, signifying a boost in its overall effectiveness.

4. Results and Discussion

The study employed various performance metrics including precision, recall, accuracy, F1 score, ROC curve and AUC to evaluate the effectiveness of four ML classifiers: logistic regression, random forest (RF), decision tree, and XGBoost. The dataset underwent a 70 - 30 split for training and testing, respectively, to identify cardiovascular disease presence.

Results showcased that the RF algorithm achieved the highest cross-validation accuracy of 0.91, with notable precision, recall, and F1 score of 0.90, 0.92, and 0.91, respectively, for predicting negative results (0—absence of cardiovascular disease). Furthermore, for positive results (1—presence of cardiovascular disease), the RF model demonstrated a precision, recall, and F1 score of 0.90, 0.92, and 0.91, respectively (**Table 3**). Similarly, the XGBoost, decision tree, and logistic regression algorithms produced accuracies of 0.90, 0.86, and 0.70, with corresponding precision, recall, and F1 score ranges as follows; XGBoost: (0.89, 0.91), (0.91, 0.89), (0.90, 0.90), Decision tree: (0.88, 0.85), (0.84, 0.89), (0.86, 0.87), Logistic regression: (0.69, 0.70), (0.72, 0.67), (0.70, 0.69). The AUC curve (**Figure 5**) displayed that RF had the highest AUC score of 0.91, followed by XGBoost at 0.89, while decision tree and logistic regression scored 0.86 and 0.70, respectively.

Transitioning to the deep learning sequential model, it demonstrated superior accuracy and loss on both training (accuracy: 0.8113, loss: 0.4149) and validation sets (accuracy: 0.8142, loss: 0.4100) before early stopping at the 15th epoch. The learning curve for this model is depicted in **Figure 6**.

Comparing to Lupague *et al.*'s findings on the same data in 2023 using logistic regression, they achieved accuracies of 79.18% for CVDs classification and 73.46% for healthy individuals, with an AUC value of 0.837. However, our study obtained an AUC score of 0.70, possibly due to different hyperparameters. Nonetheless, RF outperformed both, our other models and earlier studies mentioned.

Table 3. Evaluation matrix.

ML Models	Precision (0, 1)	Recall (0, 1)	F1-Score (0, 1)	Accuracy
Logistic Regression	0.69/0.70	0.72/0.67	0.70/0.69	0.70
Decision Tree	0.88/0.85	0.84/0.89	0.86/0.87	0.86
Random Forest Classifier	0.92/0.90	0.90/0.92	0.91/0.91	0.91
XGBoost	0.89/0.91	0.91/0.89	0.90/0.90	0.90

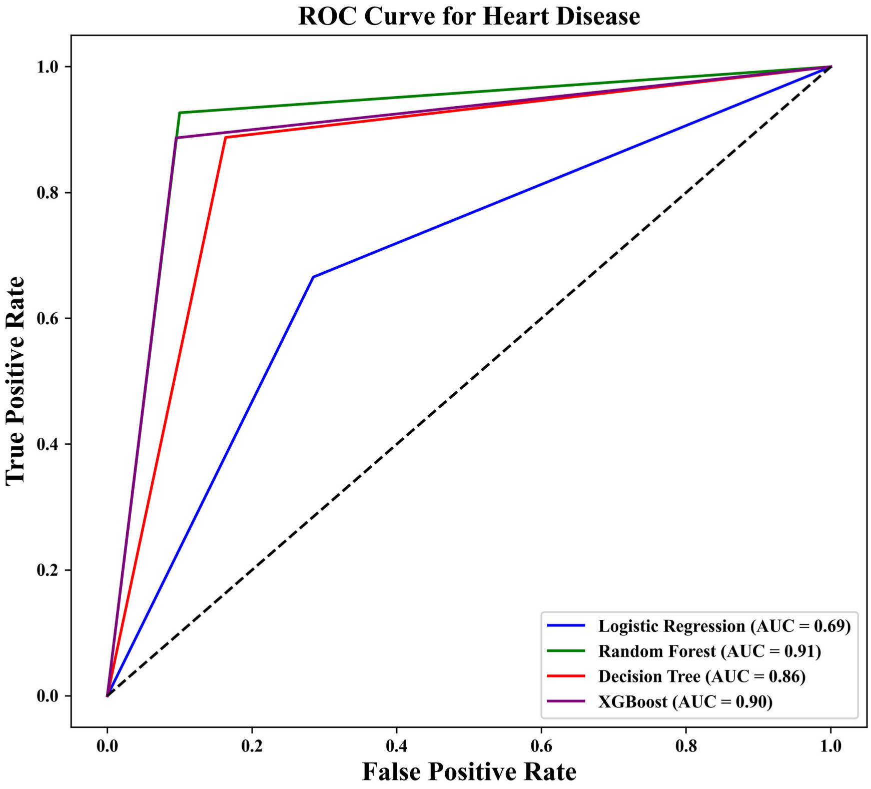

Figure 5. Roc curve for the machine learning models.

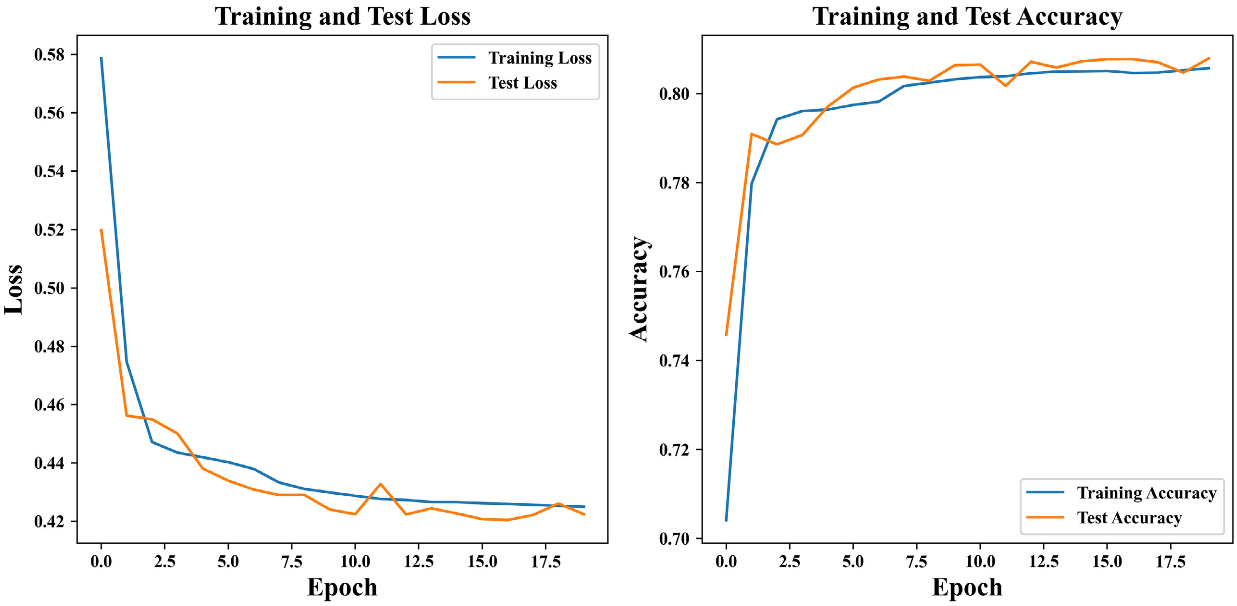

Figure 6. Accuracy and loss curve for deep learning sequential model.

5. Conclusions

In this study, we tackled the urgent global health challenge posed by CVDs, aggravated by behavioral risk factors like poor diet, physical inactivity, and substance use. Early detection of CVDs is critical for effective intervention and averting premature mortality. Leveraging a repertoire of machine learning and deep learning techniques, including logistic regression, decision trees, random forest classifier, XGBoost, and a sequential model, our objective was to devise a robust method for early diagnosis using data from the BRFSS program. Our investigation unveiled the RF classifier as the standout performer, boasting an impressive accuracy of 0.91, surpassing alternative machine learning and deep learning methodologies. Following closely were XGBoost (accuracy: 0.90), decision tree (accuracy: 0.86), and logistic regression (accuracy: 0.70). Furthermore, our deep learning sequential model exhibited promising classification performance, recording an accuracy of 0.80 and a loss of 0.425 on the validation set.

These findings underscore the potency of machine learning and deep learning approaches in bolstering cardiovascular disease prediction and management strategies. By harnessing publicly available datasets and employing advanced computational methodologies, we are positioned to make significant strides in improving public health outcomes and combating the scourge of cardiovascular disease.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Du, X., Su, X., Zhang, W., Yi, S., Zhang, G., Jiang, S., Li, H., Li, S. and Xia, F. (2021) Progress, Opportunities, and Challenges of Troponin Analysis in the Early Diagnosis of Cardiovascular Diseases. *Analytical Chemistry*, **94**, 442-463. https://doi.org/10.1021/acs.analchem.1c04476

[2] Fernández-Ruiz, I. (2019) Artificial Intelligence to Improve the Diagnosis of Cardiovascular Diseases. *Nature Reviews Cardiology*, **16**, Article No. 133. https://doi.org/10.1038/s41569-019-0158-5

[3] Cercato, C. and Fonseca, F.A. (2019) Cardiovascular Risk and Obesity. *Diabetology & Metabolic Syndrome*, **11**, Article No. 74. https://doi.org/10.1186/s13098-019-0468-0

[4] Dalal, S., Goel, P., Onyema, E., Alharbi, A., Mahmoud, A., Algarni, M. and Awal, H. (2023) Application of Machine Learning for Cardiovascular Disease Risk Prediction. *Computational Intelligence and Neuroscience*, **2023**, Article ID: 9418666. https://doi.org/10.1155/2023/9418666

[5] Pal, M., Parija, S., Panda, G., Dhama, K. and Mohapatra, R. (2022) Risk Prediction of Cardiovascular Disease Using Machine Learning Classifiers. *Open Medicine*, **17**, 1100-1113. https://doi.org/10.1515/med-2022-0508

[6] Awuah, R., Afrifa-Anane, E. and Agyemang, C. (2015) Cardiovascular Diseases and Established Risk Factors in Low-and Middle-Income Countries. In: de Graft Aikins, A. and Agyemang, C., Eds., *Chronic Non-Communicable Diseases in Low and Middle-Income Countries*, CABI Digital Library, Wallingford, 1-13. https://doi.org/10.1079/9781780643328.0001

[7] Hasani, W., Muhamad, N., Hanis, T., Maamor, N., Chen, X., Omar, M., Cheng Kueh, Y., Abd Karim, Z., Hassan, M. and Musa, K. (2023) The Global Estimate of Premature Cardiovascular Mortality: A Systematic Review and Meta-Analysis of

Age-Standardized Mortality Rate. *BMC Public Health*, **23**, Article No. 1561. https://doi.org/10.1186/s12889-023-16466-1

[8] Itchhaporia, D. (2022) Artificial Intelligence in Cardiology. *Trends in Cardiovascular Medicine*, **32**, 34-41. https://doi.org/10.1016/j.tcm.2020.11.007

[9] Mathur, P., Srivastava, S., Xu, X. and Mehta, J. (2020) Artificial Intelligence, Machine Learning, and Cardiovascular Disease. *Clinical Medicine Insights: Cardiology*, **14**. https://doi.org/10.1177/1179546820927404

[10] Muhammad, Y., Tahir, M., Hayat, M. and Chong, K. (2020) Early and Accurate Detection and Diagnosis of Heart Disease Using Intelligent Computational Model. *Scientific Reports*, **10**, Article No. 19747. https://doi.org/10.1038/s41598-020-76635-9

[11] Regar, E. (2011) Invasive Imaging Technologies: Can We Reconcile Light and Sound? *Journal of Cardiovascular Medicine*, **12**, 562-570. https://doi.org/10.2459/JCM.0b013e3283492b5a

[12] Groepenhoff, F., Klaassen, R., Valstar, G., Bots, S., Onland-Moret, N., Den Ruijter, H., Leiner, T. and Eikendal, A. (2021) Evaluation of Non-Invasive Imaging Parameters in Coronary Microvascular Disease: A Systematic Review. *BMC Medical Imaging*, **21**, Article No. 5. https://doi.org/10.1186/s12880-020-00535-7

[13] Karatzia, L., Aung, N. and Aksentijevic, D. (2022) Artificial Intelligence in Cardiology: Hope for the Future and Power for the Present. *Frontiers in Cardiovascular Medicine*, **9**, Article 945726. https://doi.org/10.3389/fcvm.2022.945726

[14] Chattu, V. (2021) A Review of Artificial Intelligence, Big Data, and Blockchain Technology Applications in Medicine and Global Health. *Big Data and Cognitive Computing*, **5**, Article 41. https://doi.org/10.3390/bdcc5030041

[15] Ahmed, Z., Mohamed, K., Zeeshan, S. and Dong, X. (2020) Artificial Intelligence with Multi-Functional Machine Learning Platform Development for Better Healthcare and Precision Medicine. *Database*, **2020**, baaa010. https://doi.org/10.1093/database/baaa010

[16] DeGroat, W., Abdelhalim, H., Patel, K., Mendhe, D., Zeeshan, S. and Ahmed, Z. (2024) Discovering Biomarkers Associated and Predicting Cardiovascular Disease with High Accuracy Using a Novel Nexus of Machine Learning Techniques for Precision Medicine. *Scientific Reports*, **14**, Article No. 1. https://doi.org/10.1038/s41598-023-50600-8

[17] Drożdż, K., Nabrdalik, K., Kwiendacz, H., Hendel, M., Olejarz, A., Tomasik, A., Bartman, W., Nalepa, J., Gumprecht, J. and Lip, G. (2022) Risk Factors for Cardiovascular Disease in Patients with Metabolic-Associated Fatty Liver Disease: A Machine Learning Approach. *Cardiovascular Diabetology*, **21**, Article No. 240. https://doi.org/10.1186/s12933-022-01672-9

[18] Ambekar, S. and Phalnikar, R. (2018) Disease Risk Prediction by Using Convolutional Neural Network. 2018 *Fourth International Conference on Computing Communication Control and Automation* (*ICCUBEA*), Pune, 16-18 August 2018, 1-5. https://doi.org/10.1109/ICCUBEA.2018.8697423

[19] Larroza, A., Materka, A., López-Lereu, M.P., Monmeneu, J.V., Bodíand, V. and Moratal, D. (2017) Differentiation between Acute and Chronic Myocardial Infarction by Means of Texture Analysis of Late Gadolinium Enhancement and Cine Cardiac Magnetic Resonance Imaging. *European Journal of Radiology*, **92**, 78-83. https://doi.org/10.1016/j.ejrad.2017.04.024

[20] Oyewola, D., Dada, E. and Misra, S. (2024) Diagnosis of Cardiovascular Diseases by Ensemble Optimization Deep Learning Techniques. *International Journal of*

Healthcare Information Systems and Informatics. **19**, 1-21. https://doi.org/10.4018/IJHISI.334021

[21] Alaa, A., Bolton, T., Di Angelantonio, E., Rudd, J. and Schaar, M. (2019) Cardiovascular Disease Risk Prediction Using Automated Machine Learning: A Prospective Study of 423,604 UK Biobank Participants. *PLOS ONE*, **14**, e0213653. https://doi.org/10.1371/journal.pone.0213653

[22] Musa, A. (2013) Comparative Study on Classification Performance between Support Vector Machine and Logistic Regression. *International Journal of Machine Learning and Cybernetics*, **4**, 13-24. https://doi.org/10.1007/s13042-012-0068-x

[23] Nashif, S., Raihan, M., Islam, M. and Imam, M. (2018) Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System. *World Journal of Engineering and Technology*, **6**, 854-873. https://doi.org/10.4236/wjet.2018.64057

[24] Baghdadi, N., Farghaly Abdelaliem, S., Malki, A., Gad, I., Ewis, A. and Atlam, E. (2023) Advanced Machine Learning Techniques for Cardiovascular Disease Early Detection and Diagnosis. *Journal of Big Data*, **10**, Article No. 144. https://doi.org/10.1186/s40537-023-00817-1

[25] Kecman, V. (2005) Support Vector Machines—An Introduction. In: Wang, L., Ed., *Support Vector Machines*: *Theory and Applications*, Springer, Berlin, 1-47. https://doi.org/10.1007/10984697_1

[26] Uddin, K., Ripa, R., Yeasmin, N., Biswas, N. and Dey, S. (2023) Machine Learning-Based Approach to the Diagnosis of Cardiovascular Vascular Disease Using a Combined Dataset. *Intelligence-Based Medicine*, **7**, Article 100100. https://doi.org/10.1016/j.ibmed.2023.100100

[27] Pal, M. and Parija, S. (2021) Prediction of Heart Diseases Using Random Forest. *Journal of Physics*: *Conference Series*, **1817**, Article 012009. https://doi.org/10.1088/1742-6596/1817/1/012009

[28] Chen, T. and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. *Proceedings of the* 22*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 785-794. https://doi.org/10.1145/2939672.2939785

[29] Ogunleye, A. and Wang, Q. (2019) XGBoost Model for Chronic Kidney Disease Diagnosis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **17**, 2131-2140. https://doi.org/10.1109/TCBB.2019.2911071

[30] Osman, A., Ahmed, A., Chow, M., Huang, Y. and El-Shafie, A. (2021) Extreme Gradient Boosting (Xgboost) Model to Predict the Groundwater Levels in Selangor Malaysia. *Ain Shams Engineering Journal*, **12**, 1545-1556. https://doi.org/10.1016/j.asej.2020.11.011

[31] Agatonovic-Kustrin, S. and Beresford, R. (2000) Basic Concepts of Artificial Neural Network (ANN) Modeling and Its Application in Pharmaceutical Research. *Journal of Pharmaceutical and Biomedical Analysis*, **22**, 717-727. https://doi.org/10.1016/S0731-7085(99)00272-1

[32] Dahal, K., Dahal, J., Banjade, H. and Gaire, S. (2021) Prediction of Wine Quality Using Machine Learning Algorithms. *Open Journal of Statistics*, **11**, 278-289. https://doi.org/10.4236/ojs.2021.112015

[33] Aggarwal, S., Bhatia, M., Madaan, R. and Pandey, H. (2021) Optimized Sequential Model for Plant Recognition in Keras. *IOP Conference Series*: *Materials Science and Engineering*, **1022**, Article 012118. https://doi.org/10.1088/1757-899X/1022/1/012118