EAI Endorsed Transactions
on Pervasive Health and Technology
Research Article EAI.EU

Automated Cardiovascular Disease Prediction Models: A Comparative Analysis

T. H. Choudhury[1] and B. Choudhury[1,*]

[1]Department of Computer Science and Engineering, Assam Down Town University, Guwahati, India

Abstract

INTRODUCTION: Cardiovascular disease (CVD) is one of the primary causes of the increased mortality rate universally. Therefore, automated methods for early prediction of CVD are of utmost importance to prevent the disease.
OBJECTIVES: In this study, we have pointed out the major advantages, drawbacks, and the scope of enhancing the prediction accuracy of the existing automated cardiovascular disease prediction methods. In addition to that, we have analyzed various combinations of attributes that can help in prediction at the earliest.
METHODS: We have exploited various machine learning models to analyse their performances in predicting the CVD at the earliest.
RESULTS: For a publicly available database, the Artificial Neural Network attained the highest accuracy of 88.5% and recall of 90%.
CONCLUSION: We justified the notion that it will be beneficial to identify potential physiological and behavioural attributes to predict CVD accurately as early as possible.

Keywords: Cardiovascular disease, mortality, prediction, machine learning, heart-attack, attributes.

Received on 01 December 2021, accepted on 16 February 2023, published on 29 May 2023

doi: 10.4108/eetpht.8.3402

Introduction

Cardiovascular disease (CVD) is a collection of diseases affecting the heart involving the blood circulatory system, including coronary heart disease, congenital heart disease, rheumatic heart disease, angina pectoris, deep vein thrombosis, peripheral arterial disease, pulmonary embolism, myocardial infarction, heart failure, arrhythmia, and cerebrovascular disease [1]. Every year, approximately 17 million people decease due to CVDs, particularly heart attacks and strokes. These perilous events are caused mainly by a blockage due to fatty deposits that prevent blood from flowing to the heart or brain. Other than the main underlying pathological processes of atherosclerosis, other risk factors for heart attack and strokes include unhealthy diet, physical inactivity, obesity, harmful use of tobacco or/and alcohol, being a male, growing age, genetic disposition, diabetes, hypertension, hyperlipidemia, and psychological factors [2]–[4].

CVD is a prime public health concern in both developing and developed nations. According to World Health Organization (WHO), in 2016, CVD accounted for 31% of the total deaths worldwide [5] and 85% of them are caused due to heart attack and stroke. The heart-related CVDs have become the prime reason of death in India as well. According to the Global Burden of Disease Report of 2016, approximately 1.7 million Indians were killed by heart disease during that time. Heart-related diseases greatly affect the economy as the spending on health care increases significantly, but the productivity of the individual decreases. According to WHO, heart related or cardiovascular diseases cost India a lost up to $237 billion from 2005-2015. It is anticipated that the mortality count due to CVD would increase up to 24.2 million by 2030 [6]. Thus, feasible and accurate prediction of cardiovascular diseases is very essential.

The CVD risk prediction equations derived from epidemiologic cohort studies, e.g., the Framingham equation, have proved to be useful traditional tools in primary prevention of CVD at the clinical level. Traditionally, a medical

*Corresponding author. Email:bismi.choudhury@gmail.com

practitioner/ specialist checks or counts all the vital risk factors of CVD in the patients and predict the overall risk of CVD. This is a time consuming and tedious process. The Machine Learning (ML) models can effectively enhance the CVD prediction accuracy compared to the traditional CVD detection methods. A ML model can help the specialist to analyse all the risk factors and predict risk of CVD more accurately in less time.

In this paper, we have focused on reviewing various automated methods for CVD prediction. We have discussed the problems and the gaps with the existing models and the future scope. In addition to that, we have engaged in analysing various machine learning predictors for accurate CVD prediction. By considering various combinations of physiological features, we performed a comparative analysis of CVD prediction capability of machine learning algorithms, such as k-Nearest Neighbor (KNN), Logistic Regression, Random Forest, and Artificial Neural Network (ANN), in a publicly available UCI database.

Therefore, in the following section, we have provided the state-of-the-art methods available for predicting CVD, then in Section 3, we have provided a comparative analysis of various machine learning models for CVD prediction including k-Nearest Neighbor (KNN), Logistic Regression, Random Forest, and Artificial Neural Network (ANN). In Section 4, experimental results of these models are evaluated on UCI database with different combination of features followed by a discussion and conclusion.

2. State-of-the Art Methods

In this paper, we have provided a vast literature review on CVD prediction or risk assessment techniques. In the literature, we can see that CVD risk is presented mainly in terms of absolute, relative, lifetime, and recurrent risk. Most of the CVD prediction methods are based on physiological factors such as age, gender, blood sugar level, and cholesterol level, diabetes and other behavioural factors such as smoking. The majority of the methods are statistical methods to find the correlation among various such physiological factors and those models/frameworks are population-based and diverse in nature. Again the majority of the models used Cox proportional hazards regression, logistic regression, or accelerated failure time analysis other than traditional statistical methods [7]. In [8], the authors investigated physiological factors such as three types of blood pressure (Stress Blood Pressure, Ambulatory Blood Pressure, and Resting blood pressure), cholesterol level, three types of Electrocardiograph (resting, ambulatory, and stress ECG), Arterial Stiffness, Ankle-Brachial Index, Pulse Wave Analysis, and Blood Glucose for predicting a high risk of CVD. The CVD prediction models are mainly statistical methods to find the correlation among various physiological factors that can predict CVD risk at the earliest or accurately [8]. Some of the well-known techniques or tools are listed in Table 1. The European SCORE project used Framingham equations to calculate CVD risk based on age, gender, Total Cholesterol (TC), TC/HDL, systolic blood pressure, and smoking [9]. From the table, we can observe that the available tools are diverse in terms of predicting features, databases, and sample sizes; and therefore, direct comparison of these tools are not very practical.

Due to the technological advancement and machine learning algorithms, in recent years, researchers have applied many different data mining techniques to improve the accuracy of CVD prediction. Data mining techniques have showed promising performance in diagnosing heart-related diseases with good prediction accuracy in less time and hence, minimize the incidence of a heart attack. Some of the popular machines learning models are Naive Bayes, Decision Tree, neural network, Genetic algorithm, Support Vector Machine (SVM), and classification via clustering. The combination of data mining and soft computing techniques can be useful to unravel hidden relationships among various physiological features and diagnose diseases more efficiently. Some of the existing works on cardiovascular disease detection methods are as follows:

Kumar et al. in [10] did a comparative study on various machine learning algorithms including logistic regression, nearest neighbor, support vector machine, and multilayer perceptron in approximating the severity level of cardiovascular disease. The accuracy of these models was compared considering different physiological parameters. They combined features from four databases from UCI repository including Cleveland, Hungary, Switzerland, and the VA Long Beach. Total 899 instances were integrated and considered 11 features for prediction including age, sex, chest pain types, resting blood pressure, years of smoking, fasting blood sugar, diabetes history, family history of coronary artery disease, resting ECG, maximum heart rate achieved, and degree of severity. The dataset was split into 80% training data and 20% testing data. In their experiment, Logistic regression obtained an accuracy of 88.88%, K-nearest neighbours obtained an accuracy of 84.44%. Support vector machine achieved an accuracy of 88.88%, and ANN obtained an accuracy of 88.89%. In [11], the authors developed a data science framework for prediction of heart disease by utilizing various classification algorithms. They analysed the influence and distribution of various attributes for disease prediction with visualizations. This experimental work on Cleveland cardiovascular medical database focused on analysing performance of different algorithms such as SVM, Logistic regression, Random Forest, Naïve Bayes, and XG-Boost. With all 14 features into consideration, SVM and Logistic Regression achieved the highest accuracy of 79%, Naïve Bayes obtained accuracy of 78%, XG-Boost obtained accuracy of 76%, and Random Forest obtained accuracy of 75%. Shan Xu et al. in [12] proposed a practical CVD risk prediction system based on data mining techniques to support medical service. For patients' data collection and analysis, their system comprised of four stages, such as data interface, data preparation, feature selection, and classification.

In [1], presented a CVD prediction model for a 3-year risk assessment of CVD in eastern China. In 2014, they used an electronic health record system for regular follow-ups of 29930 subjects with a high risk of CVD. On that particular population, they found nearly 30 risk factors of CVD using logistic regression analysis. Some of the indicators were abnormal fasting blood glucose, lipoprotein, cholesterol, low-density, obesity, old age, male gender, family income, smoking, and drinking. They applied several algorithms to build a prediction model including multivariate regression model, Naïve Bayes,

classification, and regression tree (CART), Ada Boost, Bagging, and Random Forest. From the performance comparison, they concluded that the Random Forest performed better than all tested methods with an AUC of 0.787.

Nakai et al. in [13], constructed a CVD risk prediction model based on an urban Japanese cohort study. They used multivariable cox proportional hazards models to develop the prediction model for coronary heart disease and stroke. They considered gender, age, diastolic blood pressure, systolic blood pressure, high-density, low-density and non-high-density lipoprotein cholesterol, diabetes mellitus, urinary protein, and smoking as risk factors and paired them with ECG data and without ECG data. Then, the performance of their model was compared against the Framingham CVD risk score (FRS) model. They evaluated the discriminatory ability of the proposed and FRS model using concordance statistics (C-statistics). The C-statistics for models with ECG data were 0.782 (95% CI, 0.766–0.799) and without ECG data were 0.781 (95% CI, 0.765–0.797). The C-statistic for the FRS model was 0.768 (95% CI, 0.750–0.785).

Rezaee et al. in [14] developed risk prediction models for multiple cardiovascular diseases including stroke, coronary artery disease, abdominal aortic aneurysm, and deep vein thrombosis. The same prediction model can also predict the risk of type 2 diabetes mellitus. They considered various conventional risk factors, including gender, age, systolic and diastolic blood pressure, body mass index (BMI), current and past smoking history, physical activity, and family history along with 22 blood biomarkers, including three blood count tests and 19 biochemical markers for the prediction model. They collected data from 254,220 UK Biobank participants and utilized Cox proportional hazards regression to develop the prediction model. The C-index achieved for coronary artery disease is 0.794, stroke is 0.778 deep vein thrombosis is 0.743 and the abdominal aortic aneurysm is 0.893. The C-index for predicting type 2 diabetes mellitus is 0.909.

Alty et al. in [15] used a patient's volume pulse to predict the risk of CVD. They analysed the reading of the aortal pulse wave velocity using a digital volume pulse and used SVM to predict the high or low arterial stiffness. They analysed various combinations of features from the digital volume pulse waveform and selected three efficient features, viz. Crest-time, Peak-to-peak time, and maximum slope. Using the Gaussian Radial Basis kernel function, SVM successfully predicted a high and low risk of CVD with an accuracy of 85%, sensitivity of 93%, and specificity of 78% for 134 (101 records were used for training and 33 records were used for testing) records-collected from the south-east London hypertension clinic.

Tuan et al. in [16] applied statistical and geostatistical linear prediction models to obtain efficient features of the mass spectrometry data of blood samples for early prediction of CVD. The computational theories are used to analyse the proteins for the predicting the major risk of adverse cardiac events. After collecting the blood samples, the protein level of myeloperoxidase (MPO) and other known cardiovascular biomarkers were measured. They used linear prediction coding for feature selection. With statistical distortion measure and geostatistical distortion measure, the prediction model achieved an average accuracy of 83% and 97% respectively.

Zhu et al. in [17] combined data mining technology to design and develop a readmission risk assessment system for patients with cardiovascular disease. Their model was 90.6% accurate for automatically predicting the risk level and risk factors of CVD. The dataset was collected from a hospital in Beijing consisting of a total of 10228 instances and 1074 instances of them were the readmission cases within 30 days. K- Means clustering was used to cluster the attributes including age, BMI, SBP, complication, operation scale, and wound grade. A linear regression model was used to analyse the risk factors from the clustered attributes. Finally, an Artificial Neural Network with two hidden layers was used for the readmission risk prediction. Out of 9897 instances, 50% of data were used for training and 50% were used for testing.

Park et al. in [18] proposed a frequency-aware based Attention-based LSTM to specifically put weight on the important medical features for predicting CVD. They developed a specialized prediction mechanism to obtain the

correlation between input features and the prediction target. The main goal was to predict the value of ejection fraction for the next visit and predicted the risk using the features indirectly related to CVD. In their research, they used a dataset of CVD related records of nine years, collected from a hospital in Seoul. They extracted 40 features from the records of 4551 patients including cholesterol, glucose, and albumin. For every visit of patients in six months, they checked the frequency of the selected features. Then, the features were interpolated using K-NN interpolation. Considering the ejection fraction as a prediction target, the model obtained root means square error (RMSE) of 3.65 and a mean absolute error (MAE) of 2.49 in predicting the risk of CVD.

Peng et al. in [19] provided a survey on CVD prediction using an artificial neural network. Their survey focused on mainly two datasets, viz., MIT-BIH and UCI and the use of Electrocardiogram (ECG) features. In [20], [21] used ANN to predicted CVD with 96.2% accuracy using the Pan-Tompkins feature extraction method and 100% accuracy using the LM algorithm respectively. In [22], Multi-Layered Feed Forward Network and Probabilistic Neural Networks are used to discriminate six types of ECG signals from the normal signal. In [23] used the BPNN to classify normal, obstruction, tachycardia, and bradycardia. Raj et al. in [24] used Wavelet Packet Decomposition for feature extraction from ECG signals and ANN for classifying normal and abnormal signals. In [25]–[27] used the combination of ANN and LM algorithm for CVD prediction to achieve accuracy over 90%. By experimenting on the UCI database, [28] used BPNN and genetic algorithm obtaining 94% accuracy, [29] used tiered multivariate analysis and ANN obtaining 86% accuracy, [30] used Bayesian ANN attaining 93% accuracy, [31] used enhanced Random forest-achieving 99.6% accuracy, and [32] used Deep Neural Network attaining 96% accuracy. Table 2 summarizes some of the existing machine learning methods for predicting the risk of CVD.

From the vast literature review, it can be noticed that the there are lots of research going on for CVD risk prediction and majority of the methods are relying on similar type of physiological risk factors. It is time for identifying new biomarkers for CVD risk assessment and prediction. However, it

is important to validate the recent cardiovascular risk factors before they are implemented in standard clinical care [8]. The non-clinical features, such as geographical location, early life and childhood, social relationship, and lifestyle can also be possible risk factors. Therefore, the non-clinical features should-be explored more to filter out the risk assessment process. The automated CVD risk prediction methods have the great potential to prevent the fatal condition and provide treatment to the patients as early as possible.

Table 1: Various Exiting Tools for CVD Risk Assessment

Tools	Predictors	Dataset	Sample Size	Performance
SCORE (high/low risk) [33]	Gender, Age, Diabetes, total cholesterol, Systolic blood pressure, Smoking	12 pooled European studies	117,098 men and 88,080 women	AUROC = 0.80
QRISK 2 (men/women) [34]	Age, Gender, Diabetes mellitus, Atrial fibrillation Height / Weight, Chronic kidney disease, Smoking Systolic blood pressure and on treatment for it, Rheumatoid arthritis	QRESEARCH database	2.29 million	AUROC = 0.79
Reynolds (men/women) [35], [36]	Age, Gender, Diabetes mellitus, Smoking, Family history of CVD, Systolic blood pressure, Total cholesterol / High-density lipoprotein CRP	Women's Health study & Physician's health study II	10,724 men and 24,558 women	AUROC = 0.71
PROCAM (coronary/cerebral) [37]	Age, Gender, Smoking, DM Height / Weight, Systolic blood pressure, Total cholesterol / High-density lipoprotein, FHx	Prospective study	18,460 men and 8515 women	AUROC = 0.82
INTERHEART modifiable risk score (men/women) [38]	Age, Gen, DM, Smoking, Systolic blood pressure, poB:A1 ratio, obesity, dietary, physical and psychosocial factors	INTERHEART case-control study	19470 individuals	AUROC = 0.71

However, sample size and population greatly affect the performance of the automated predictors in real life scenario. Therefore, in this era of large datasets, instead of focussing on new methods, research should be focused at validating the existing models. The existing models should be tried on the local settings and population to compare the relative performances of the prediction models and can be improved the prediction using better predictors, novel attributes, and enhanced algorithms.

3. Comparative Analysis of Machine Learning Models

In this paper, we exploited various machine learning algorithms to analyse the data and predict the high and low risk of CVD based on physiological attributes. Fig. 1 shows the block diagram of the proposed methodology for risk of CVD prediction.

The proposed methodology includes the following steps:

- **Exploratory Data Analysis:** The dataset will be collected from a publicly available heart disease dataset. Here, we have used UCI database for CVD.
- **Pre-processing Data:** The pre-processing phase involves multiple steps including data integration, data cleaning, filling of missing values, removing redundant data, which lead to fault prediction. The data cleaning can be performed by handling the corrupted and missing values. All the missing values of the dataset will be filled in. All the existing categorical data will be converted into numeric values using One Hot Encoding.
- **Feature Analysis and Selection:** The database might contain various factors or predictors that will be more beneficial to predict CVD effectively and efficiently. Therefore, we need to carefully analyze that to have a proper insight into predictors.

- **Algorithms for Classifier:** A classifier needs to be trained to correctly predict the risk of CVD by analysing the features. A machine learning model can predict in terms of probability. The CVD dataset contains categorical data, which can't be implemented directly. Therefore, to deal with categorical data, One-hot encoding has been adopted to convert data into an appropriate representation. Instead of providing class values in the target column, in one-hot encoding, the target column is extended to multiple columns, one for each class.

The various machine learning algorithms exploited for predicting CVD are explained below:

Table 2: Existing CVD Prediction Methodologies in Literature

Author	Method Used	Database Used	Remarks
Kumar et al. (2020) [10]	MLP, k-NN, logistic regression and SVM	UCI Repository	Max Accuracy= 88.89% (ANN)
Yang et al. (2020) [1]	CART, Naïve Bayes, Random forest	Database from China	AUC= 0.787
Prakash et al. (2020) [11]	Random Forest, SVM, Logistic regression, Naïve Bayes, and XG-Boost	Cleveland Heart-Disease Database	Max Accuracy= 79% (SVM, Logistic regression)
Xu et al. (2017) [12]	Random forest	Cleveland Heart-Disease Database, dataset of PKU People's Hospital	Accuracy = 91.6%, 97%
Amma et al. (2012) [28]	ANN, Genetic Algorithm	UCI Repository	Accuracy = 94%
Wiharto et al. (2017) [29]	Multivariate analysis and ANN	UCI Repository	Accuracy = 86%
Sihem et al. (2018) [31]	Enhanced Random forest	UCI Repository	Accuracy = 99.6%
Darmawahyuni et al. (2019) [32]	Deep Neural Network	UCI Repository	Accuracy = 96%

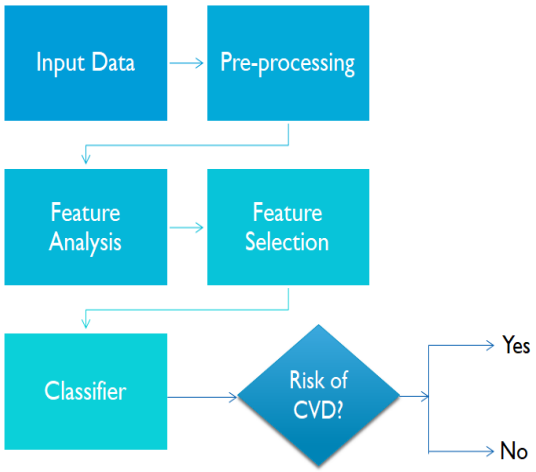

Figure 1. Process Flow of the Proposed Methodology

3.1. Logistic Regression:

It is a supervised learning algorithm, which accepts – labelled data and try to predict the correct class for the new input data. It uses a logistic function or sigmoid function that can be defined by the following equation:

$$logistic\ (\eta) = \frac{1}{1+\exp\ (-\eta)} \quad (1)$$

The logistic function is shown in Fig 2.

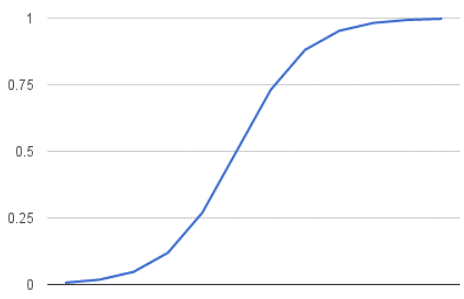

Figure 2. Logistic Regression Function

3.2. K-Nearest Neighbors:

K-nearest neighbor is a popular supervised classification algorithm that classifies a sample to a target class based on the class label of its K neighboring samples. The algorithm randomly selects K samples in the dataset and measures the distance between these samples and the given input sample using any distance metric, for e.g., Euclidean, Manhattan. The given sample is classified to a target class based on the majority of the neighboring samples' class as shown in Fig. 3.

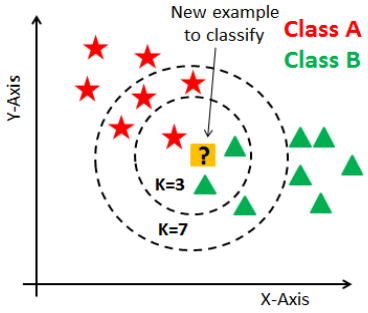

Figure 3. k-Nearest Neighbor Algorithm

3.3. Random Forest:

This supervised learning algorithm can be used for both classification and regression. It is an ensemble of multiple decision trees that classifies samples based on the decision of all the tree. By using a majority voting schemes, the given input sample is classified to a particular class as shown in Fig 4.

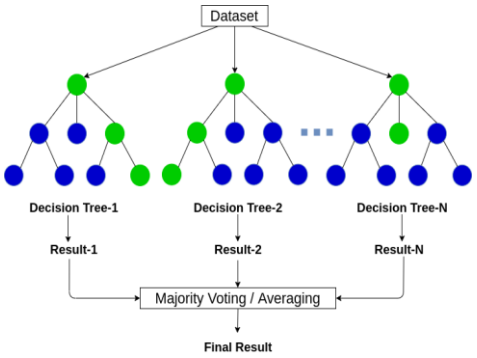

Figure 4. Random Forest Algorithm

3.4. Artificial Neural Network:

Artificial Neural Network (ANN) mimics the functioning of human brain. ANN composed of multiple nodes, analogous to neurons in human brain, arranged in a sequence of layers. The input features are processed and analysed in each layer of ANN. The activation function helps in the learning process and the output of each layer is passed on to the next layer using the following equation:

$$y = \sum_{i=1}^{n} w_i x_i + b \quad (2)$$

Where, y denotes the output of a neuron, w is the weight associated with each input feature x and b is the bias. The weights associated with every neuron specify the meaningful features. The overall prediction is estimated by updating the weights in every epoch during the training with the help of Gradient Descent Algorithm. Fig. 5 shows the architecture of ANN.

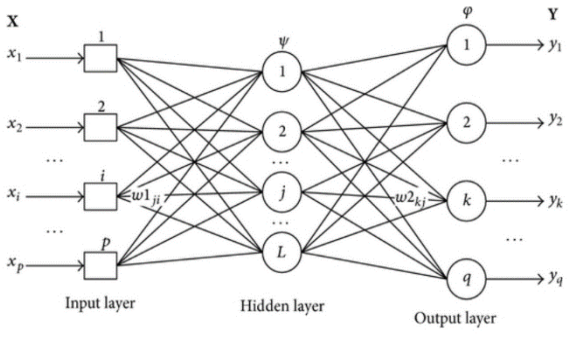

Figure 5. Architecture of ANN

4. Experimental Analysis:

In this section, we have thoroughly discussed the experimental evaluation of the various prediction models for CVD prediction.

4.1. Database Used:

The dataset is collected from the publicly available database at the UCI machine learning repository. The collected dataset has 14 features and a total of 303 readings. The attributes are as follows:

- Age
- Gender
- Serum cholesterol in mg/dl
- Fasting blood sugar > 120 mg/dl
- Resting blood pressure
- Resting electrocardiographic results (values 0,1,2)
- Maximum heart rate achieved
- Chest pain type (4 values)
- Exercise-induced angina
- Oldpeak = ST depression induced by exercise relative to rest
- The slope of the peak exercise ST segment
- Number of major vessels (0-3) colored by fluoroscopy
- Thal (thalassemia): 3 = normal; 6 = fixed defect; 7 = reversible defect

The dataset consists of 165 positive cases (Class=1) of heart diseases and 138 negative cases (Class=0).

4.2. Experimental results:

We have conducted experiments on the collected database with different combinations of features for better prediction of possible heart disease. Based on the count and types of predictors the performances of various machine learning models are trained to predict CVD accurately. We have explored four machine learning algorithms, viz., Logistic Regression, K-Nearest Neighbor Algorithm, Random Forest, and Artificial Neural Network; and used them as prediction models. We have analyzed the performance of these models in terms of the accuracy of prediction. In addition to that, the Confusion matrix is generated to analyze the true positive, true negative, precision etc.

The data pre-processing has been done before training the prediction models. Data cleaning is an important step in case of missing values and data sample equalization. However, there is no missing value in the dataset and all attributes has equal no or records that is 303. The co-relation of various features or data points (i.e., age, trestbps, cp, thalach,chol,) concerning the label of the data points (targets) is shown in Fig. 6. From the data visualization, it can be observed that the attributes are not directly separable, i.e., we cannot directly classify the severity of heart disease based on one or two attributes.

We have exploited four algorithms for CVD prediction, viz. k-NN, Logistic Regression, Random Forest, and Artificial Neural Network. The performances of these classifiers are evaluated in terms of four parameters: Accuracy, Precision, Recall and F1 Value. Accuracy measures the total number of correct predictions and can be calculated by following equation:

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + Flase\ Positive + True\ Negative + False\ Negative} \quad (3)$$

The precision depicts how accurate the model is out of the total predicted positive and it can be computed using following equation:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (4)$$

Recall or Sensitivity depicts how correctly the model can identify the true positive and it can be computed using following equation:

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (5)$$

F1 score is the harmonic mean of precision and recall and can be calculated using following equation:

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (6)$$

Data Analysis I:

To have a better insight into the features of predicting CVD, we have analyzed the data in different combinations. We have implemented Logistic Regression, K-NN, Random Forest, and Artificial Neural Network (ANN) for performance evaluation. These models are trained with 13 features of 303 records for binary classification. The dataset is randomly divided into an 8:2 ratio where 80% data are used for training and 20% data are used for testing. The 80/20 ratio is an empirically set standard training and testing split as this split provides the most accurate valid accuracy estimate without underestimating the approximation error [39]. With 20 nearest neighbor (k = 20), k-NN has achieved highest accuracy of 83.61%, precision 70.58%, recall 75% and F1 score 72%.

An ANN is designed with 4 layers including input layer, 2 hidden layers, and 1 output layer. The ReLU activation is used in the hidden layers and Logistic/Sigmoid activation function is used in the output layer with a single neuron. The Adam optimizer is used to optimize the cross-entropy loss during the backpropagation. While conducting the experimentation, it is found that more than 2 hidden layers increases the parameters and complexity of the model. On the other hand, a single hidden layer is not good enough for learning the necessary features for accurate prediction of CVD. The Sigmoid activation function is used because in this application, we are predicting the probability of the CVD, and Sigmoid function provides probability in between 0 to 1. Moreover, it is efficient and provides smooth gradients. An Adam optimizer is chosen because with minimum tuning of hyper-parameters and less memory requirements, it provides the best optimization for the stochastic gradient decent algorithm. With a learning rate of 0.001, the ANN is trained with a batch size 10 and 100 epoch. The ANN achieved an accuracy of

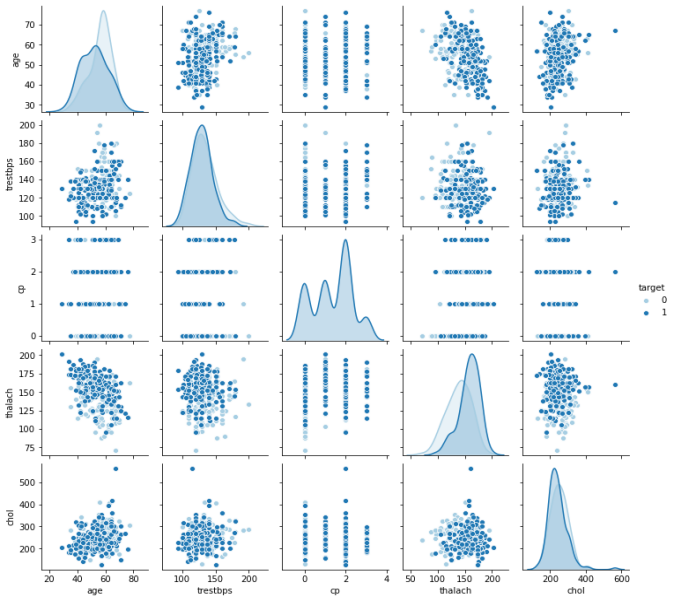

Figure 6. Correlation of Various Data Points.

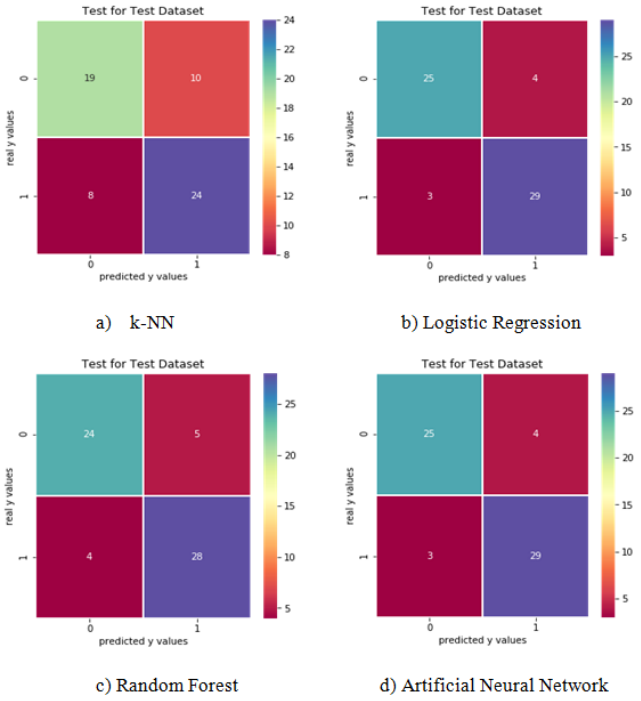

Figure 7. Confusion matrices using 13 features.

-88.52%, precision 87.87%, recall 90%, and F1 score 89%. A Random Forest is created with 200 decision trees and maximum features 0.5 for splitting at each node where 3 data points are allowed at leaf nodes. For the logistic regression setting, random state is 42 and a Limited-memory Broyden–Fletcher–Goldfarb–Shanno solver is used for minimizing the cost function as the dataset is not very large. The CVD prediction performance of all four classifiers considering all the13 features is shown in Table 3. On testing set, with 13 features Logistic Regression has achieved an accuracy of 88.52%, with 20 nearest neighbor (k=20), k-NN has achieved 83.61%, the Random Forest has achieved 85.24%, and Artificial Neural Network achieved 88.52%. The confusion matrices are shown in Fig. 7.

Table 3. Performance Comparison using 13 Features

Models	Accuracy (%)	Precision (%)	Recall (%)	F1 Score
Logistic Regression	88.52	87.87	90	0.89
K Nearest Neighbor	83.61	70.58	75.00	0.72
Random Forest	85.24	84.84	87.50	0.86
Artificial Neural Network	88.52	87.87	90	0.89

Data Analysis II:

Here, we analysed the CVD prediction performance of the four classifiers by considering a combination of 10 features (age, gender, resting blood pressure, fasting blood sugar > 120 mg/dl, chest pain type (4 values), serum cholesterol in mg/dl, resting electrocardiographic results (values 0,1,2), maximum heart rate achieved, exercise-induced angina, oldpeak (ST depression induced by exercise relative to rest). The main goal of this analysis is to see the performance of the selected classifiers with lesser features. If we can predict risk of CVD with lesser features, it is more advantageous for the fact that we do not need to rely on evaluation of all clinical features for accurate prediction of CVD risk. If the performance of predictions with a fewer number of features is equivalent or better than that of a greater number of features, it will be time efficient. Therefore, we skipped three features: the slope of the peak exercise, number of major vessels colored by fluoroscopy, and thal. The Pearson correlation of the features are analysed (Fig 6). It is

found that the features are not directly separable. We cannot directly classify the severity of heart disease based on one or two attributes. Moreover, these three features are not included in the most significant risk factors of CVD [40], [41]. Therefore, these three features are dropped. In this experiment also the dataset is randomly divided into an 8:2 ratio where 80% data are used for training and 20% data are used for validation or testing. With same settings of parameters and hyperparameters, on testing set, Logistic Regression attained 83.60% accuracy, 86.67% precision, 81.25% recall, and 84% F1 score; with k = 20, k-NN achieved 80.36% accuracy, 70.58% precision, 75% recall, and F1 score/harmonic mean 72%; Random Forest obtained an accuracy of 85.24%, precision of 92.59%, recall of 78.12%, and F1 score 84%; and ANN achieved an accuracy of 86.88%, precision and recall of 87%, and F1 score 87%. Table 4 shows the comparative analysis of four classifiers. Fig 8 shows the confusion matrices for all 4 classifiers.

Table 4. Performance Comparison using 10 Features

Models	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Logistic Regression	83.60	86.67	81.25	84
K Nearest Neighbor	80.36	70.58	75.00	72
Random Forest	85.24	92.59	78.12	84
Artificial Neural Network	86.88	87.00	87.00	87

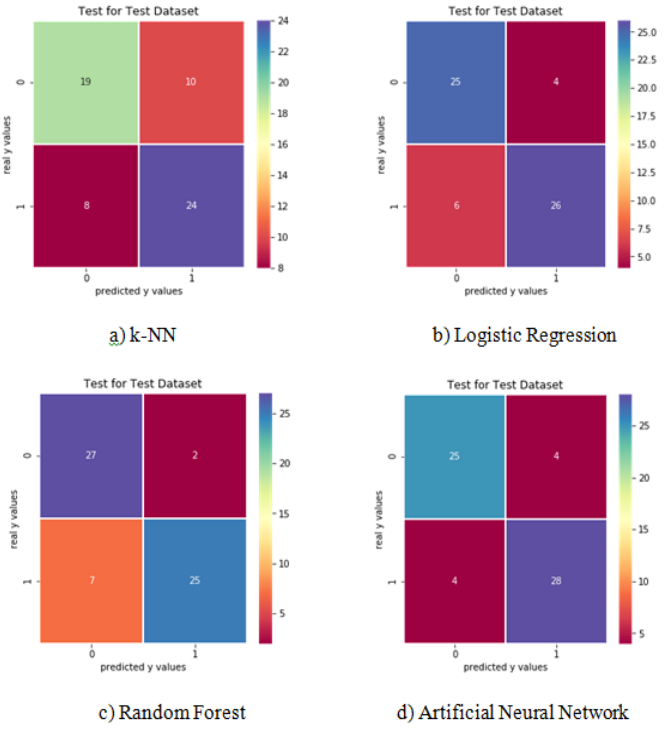

Figure 8. Confusion Metrics using 10 features.

5. Discussion

From the experimental results it can be observed that when all 13 features are considered, Logistic Regression and ANN are attaining the highest accuracy of 88.52%. For both LR and ANN, the CVD prediction performance is same in terms of all evaluation parameters (accuracy, precision, recall, and F1 score). For all four classifiers, recall values are higher than precision, which is good in this case. We are concerned about correctly identifying the heart patients those actually have heart disease so that none of the heart patients miss out the treatment, and it will be ensured by the high recall/ or sensitivity value. On the other hand, in the combination of 10 features, ANN again achieved the highest accuracy of 86.88% among 4 classifiers. However, this time, Random Forest attained highest precision of 92.5% and ANN achieved highest recall of 87%. In case of Logistic Regression and Random Forest, precision is higher than recall and in case of k-NN and ANN recall is higher precision. Precision provides the measure of patients that are correctly identified as having a heart disease out of all the patients actually having it. High precision is good when cost of false positive is high, i.e., if we want to avoid giving treatment to a patient who actually doesn't have a heart disease, but our model predicted as having it. Therefore, with high precision, Random Forest and Logistic Regression can avoid such situations. But, for our case, we would like to completely avoid any situations where a patient has heart disease, but the model identifies the patient as true negative. Therefore, high recall value is preferable in CVD prediction as the cost associated with false negative case is high and we also want to identify as many patients with CVD risk as possible. Again, for 10 combinations

of features, ANN has maintained a balanced precision and recall value which is an ideal condition. We can consider a situation when the patients who were incorrectly classified as having heart disease and the doctor inform that these could be indicative of some other disease. Then, in such situation we both high recall and high precision is important. Here, ANN achieved the highest F1 score, which provides the trade-off between precision and recall.

We can conclude that ANN is giving the best performance in CVD prediction with recall 90% for 13 features; Random Forest is giving the best performance in prediction with precision 92.59% for 10 features; and ANN is giving the best performance in terms of both accuracy and F1 score for both 13 and 10 combinations of features. Therefore, here, we can conclude that ANN is the best classification model overall for CVD prediction.

6. Conclusion

In this paper, we provided an extensive review on various CVD prediction methods. In addition to that we provided a comparison-based analysis of different machine learning algorithms for predicting cardiovascular diseases. We have exploited mainly four models, viz. k-NN, Logistic Regression, Random Forest, and Artificial Neural Network. We have also combined different sets of features for CVD prediction. We have conducted all the experiments on a publicly available database for heart disease prediction. For a different combination of attributes, the different predictive model has performed differently. For a combination of 13 attributes and 10 attributes, Artificial Neural Network with 5 hidden layers has attained the highest accuracy of 89% and 87% respectively among all four classifiers. ANN has also obtained highest recall of 90% with 13 features. There is scope of performance improvement with the help of better feature extraction method and by enhancing these machine learning algorithms. The main contribution of this study is that from multiple experimentations on CVD risk prediction one can select the best prediction model. An attempt on combination of different features provided a possibility of predicting CVD with lesser number of features to avoid overhead. The difference between the performance for CVD prediction using 13 and 10 features is minimal. With perfect combination of features/attributes, it is possible to predict CVD with high performance even with lesser number of features, which will be time efficient. In future, we will try to improve the accuracy of CVD prediction with better sets of parameters and hyper-parameters. We can extend our work with deep learning models to gain better performance. However, it will require a large dataset.

References

[1] L. Yang *et al.*, (2020) Study of cardiovascular disease prediction model based on random forest in eastern China, *Sci. Rep.*, 10(1), pp. 1–8, 2020.

[2] S. Kanjilal *et al.*, (2008) Application of cardiovascular disease risk prediction models and the relevance of novel biomarkers to risk stratification in Asian Indians, *Vasc. Health Risk Manag.*, 4(1), pp. 199–211.

[3] *Published by the World Health Organization in collaboration with the World Heart Federation and the World Stroke Organization*. .

[4] Y. Ruan *et al.*, (2018) Cardiovascular disease (CVD) and associated risk factors among older adults in six low-and middle-income countries: results from SAGE Wave 1, *BMC Public Health*, 18(1), pp. 778.

[5] Cardiovascular diseases (CVDs). [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). [Accessed: 20-Jan-2021].

[6] "HO | The Atlas of Heart Disease and Stroke, *WHO*, 2010.

[7] J. A. A. G. Damen *et al.*, (2016) Prediction models for cardiovascular disease risk in the general population: Systematic review, *BMJ (Online)*, 353(i2416). BMJ Publishing Group, 16-May.

[8] W. H. Lin, H. Zhang, and Y. T. Zhang, (2013) Investigation on cardiovascular risk prediction using physiological parameters, *Computational and Mathematical Methods in Medicine*, vol. 2013. Taylor and Francis Ltd.

[9] F. P. Cappuccio, P. Oakeshott, P. Strazzullo, and S. M. Kerry, (2002) Application of Framingham risk estimates to ethnic minorities in United Kingdom and implications for primary prevention of heart disease in general practice: Cross sectional population based study, *Br. Med. J.*, 325(7375), pp. 1271–1274.

[10] A. Kumar, R. Gyawali, and S. Agarwal, (2020) Cardiovascular disease prediction using machine learning tools, in *Advances in Intelligent Systems and Computing*, 1085, pp. 441–451.

[11] C. S. Prakash, M. Madhu Bala, and A. Rudra, (2020) Data Science Framework - Heart Disease Predictions, Variant Models and Visualizations, in *2020 International Conference on Computer Science, Engineering and Applications, ICCSEA 2020*, pp. 1–4.

[12] S. Xu, Z. Zhang, D. Wang, J. Hu, X. Duan, and T. Zhu, (2017) Cardiovascular risk prediction method based on CFS subset evaluation and random forest classification framework, in *2017 IEEE 2nd International Conference on Big Data Analysis, ICBDA 2017*, pp. 228–232.

[13] M. Nakai *et al.*, (2020) Development of a cardiovascular disease risk prediction model using the suita study, a population-based prospective cohort study in Japan, *J. Atheroscler. Thromb.*, 27(11), pp. 1160–1175.

[14] M. Rezaee, I. Putrenko, A. Takeh, A. Ganna, and E. Ingelsson, (2020) Development and validation of risk prediction models for multiple cardiovascular diseases and Type 2 diabetes, *PLoS One*, 15(7), p. e0235758, Jul.

[15] S. R. Alty, S. C. Millasseau, P. J. Chowienczyk, and A. Jakobsson, (2006) Cardiovascular disease prediction using support vector machines, pp. 376–379.

[16] T. D. Pham *et al.*, (2008) Computational prediction models for early detection of risk of cardiovascular events using mass spectrometry data, *IEEE Trans. Inf. Technol. Biomed.*, 12(5), pp. 636–643.

[17] C. Y. Zhu, S. Q. Chi, R. Z. Li, D. Y. Tong, Y. Tian, and J. S. Li, (2017) Design and development of a readmission risk assessment system for patients with cardiovascular disease, in *Proceedings - 2016 8th International Conference on Information Technology in Medicine and Education, ITME 2016*, pp. 121–124.

[18] H. D. Park, Y. Han, and J. H. Choi, (2018) Frequency-Aware Attention based LSTM Networks for Cardiovascular Disease, in *9th International Conference on Information and Communication Technology Convergence: ICT Convergence Powered by Smart Intelligence, ICTC 2018*, pp. 1503–1505.

[19] C. C. Peng, Y. C. Lai, C. W. Huang, J. G. Wang, S. H. Wang, and Y. Z. Wang, (2020) Cardiovascular Diseases Prediction Using Artificial Neural Networks: A Survey, in *2nd IEEE Eurasia*

Conference on Biomedical Engineering, Healthcare and Sustainability 2020, ECBIOS 2020, pp. 141–144.
[20] R. Ghongade and A. A. Ghato, (2007) A brief performance evaluation of ECG feature extraction techniques for artificial neural network based classification, in *IEEE Region 10 Annual International Conference, Proceedings/TENCON*, pp. 1-4.
[21] M. H. F. M. Jalil, M. F. Saaid, A. Ahmad, and M. S. A. M. Ali, (2014) Arrhythmia modelling via ECG characteristic frequencies and artificial neural network, in *Proceedings - 2014 IEEE Conference on System, Process and Control, ICSPC 2014*, pp. 121-126.
[22] H. Haseena, P. K. Joseph, and A. T. Mathew, (2009) Artificial neural network based ECG arrhythmia classification, *J. Mech. Med. Biol.*, 9(4), pp. 507-525.
[23] T. Debnath, M. Hasan, and T. Biswas, (2017) Analysis of ECG signal and classification of heart abnormalities using artificial neural network, in *Proceedings of 9th International Conference on Electrical and Computer Engineering, ICECE 2016*, pp. 353-356.
[24] A. A. S. Raj, N. Dheetsith, S. S. Nair, and D. Ghosh, (2015) Auto analysis of ECG signals using artificial neural network," in *2014 International Conference on Science Engineering and Management Research, ICSEMR 2014*, pp. 1-4.
[25] H. Gothwal, S. Kedawat, and R. Kumar, (2011) Cardiac arrhythmias detection in an ECG beat signal using fast fourier transform and artificial neural network, *J. Biomed. Sci. Eng.*, 4, pp. 289-296.
[26] R. Ceylan and Y. Özbay, (2007) Comparison of FCM, PCA and WT techniques for classification ECG arrhythmias using artificial neural network, *Expert Syst. Appl.*, 33 (2), pp. 286-295.
[27] N. K. Dewangan and S. P. Shukla, (2017) ECG Arrhythmia classification using discrete wavelet transform and artificial neural network, in *2016 IEEE International Conference on Recent Trends in Electronics, Information and Communication Technology, RTEICT 2016 - Proceedings*, pp. 1892-1896.
[28] N. G. B. Amma, (2012) Cardiovascular disease prediction system using genetic algorithm and neural network, in *2012 International Conference on Computing, Communication and Applications, ICCCA 2012*, pp. 1-5.
[29] Wiharto, H. Kusnanto, and H. Herianto, (2017) Hybrid system of tiered multivariate analysis and artificial neural network for coronary heart disease diagnosis, *Int. J. Electr. Comput. Eng.*, 7(2), pp. 1023-1031.
[30] D. Gao, M. Madden, D. Chambers, and G. Lyons, (2005) Bayesian ANN classifier for ECG arrhythmia diagnostic system: A comparison study, in *Proceedings of the International Joint Conference on Neural Networks*, 4, pp. 2383-2388.
[31] S. Nita, S. Bitam, and A. Mellouk, (2018) An Enhanced Random Forest for Cardiac Diseases Identification based on ECG signal, in *2018 14th International Wireless Communications and Mobile Computing Conference, IWCMC 2018*, pp. 1339-1344.
[32] A. Darmawahyuni, S. Nurmaini, and F. Firdaus, (2019) Coronary Heart Disease Interpretation Based on Deep Neural Network, *Comput. Eng. Appl. J.*, 8(2), pp. 1-12.
[33] R. M. Conroy *et al.*, (2003) Estimation of ten-year risk of fatal cardiovascular disease in Europe: The SCORE project," *Eur. Heart J.*, 24, pp. 987–1003.
[34] J. Hippisley-Cox *et al.*, (2008) Predicting cardiovascular risk in England and Wales: Prospective derivation and validation of QRISK2, *BMJ*, 336, pp. 1475–1482.
[35] P. M. Ridker, N. P. Paynter, N. Rifai, J. M. Gaziano, and N. R. Cook, (2008) C-reactive protein and parental history improve global cardiovascular risk prediction: The Reynolds risk score for men, *Circulation*, 118, pp. 2243–2251.
[36] P. M. Ridker, J. E. Buring, N. Rifai, and N. R. Cook, (2007) Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: The Reynolds Risk Score, *J. Am. Med. Assoc.*, 297, pp. 611–619.
[37] G. Assmann, P. Cullen, and H. Schulte, (2002) Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the Prospective Cardiovascular Münster (PROCAM) study, *Circulation*, 105, pp. 310–315.
[38] C. McGorrian *et al.*, (2011) Estimating modifiable coronary heart disease risk in multiple regions of the world: The INTERHEART Modifiable Risk Score, *Eur. Heart J.*, 32, pp. 581–589.
[39] A. Gholamy, V. Kreinovich, and O. Kosheleva, (2018) Technical Report on Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation (El Paso: Computer Science Department, The University of Texas at El Paso) 1209.
[40] R. Hajar, (2017) Risk Factors for Coronary Artery Disease: Historical Perspectives, *Heart Views*, 18(3), p. 109.
[41] D. M. T. Tran, N. Lekhak, K. Gutierrez, and S. Moonie, (2021) Risk factors associated with cardiovascular disease among adult Nevadans, *PLoS One*, 16(2), p. e0247105, Feb.