12

Cardiovascular Disease Prediction Using Deep Neural Network for Older People

Nagarjuna Telagam*, B.Venkata Kranti and Nikhil Chandra Devarasetti

Dept of Electrical Electronics and Communication Engineering, GITAM University, Bengaluru, Karnataka, India

Abstract

Cardiovascular disease is the primary reason for people's demise rate, and the rate is increasing globally every year, especially in older adults. Nearly 20 million people are dying with heart-related problems every year across the globe. It was estimated that the numbers would increase by 75 million by the end of 2040. The doctors or medical professionals in the medical field can only predict the disease with an accuracy of 67%. Doctors' unrealistic current situation needs a supporting machine learning model for accurate results. This book chapter presents the study of machine learning and deep learning algorithms with detailed and analytical comparisons, which may help new and inexperienced medical professionals or researchers in the medical field. Extracting medical data is becoming more and more necessary for predicting and treating high death rates due to heart attacks. Every day, the hospitals produce tons of terabytes of data, and the hospitals will use clinical tests for decision-making about heart data. The decision tree algorithm predicts heart disease with 86.72% and 75.40% accuracy on the train and test sets, respectively. We saw the difference between training and testing accuracy, which makes decision tree models less reliable. Hence, the Random Forest algorithm with the best hyperparameters is chosen to obtain 90.16% and 87.6% accuracy for the testing and training data set. The training is done using a four-fold cross-validation scheme, ensuring our test set accuracy is better than training set accuracy. The proposed machine learning model has an accurate algorithm that

**Corresponding author*: nagarjuna473@gmail.com
Nagarjuna Telagam: ORCID: 0000-0002-6184-6283
B.Venkata Kranti: ORCID: 0000-0003-3177-6250
Nikhil Chandra devarasetti: ORCID: 0000-0001-7770-5306

Rishabha Malviya, George Ghinea, Rajesh Kumar Dhanaraj, Balamurugan Balusamy and Sonali Sundram (eds.) Deep Learning for Targeted Treatments: Transformation in Healthcare, (369–406)

works with rich healthcare data, a high-dimensional data handling system, and an intelligent framework that uses different data sources to predict heart disease. This book chapter used an ensemble-based deep learning model with optimal feature selection to improve accuracy. The sigmoid function is used for true class or false class identification, and it is also used to calculate the loss or error in the last year in the neural network. The Adam optimizer is used to adjust the three-layer neural network architecture weights. The proposed model shows an accuracy of 97% higher than conventional methods.

***Keywords*:** Decision tree, random forest, confusion matrix, machine learning. artificial intelligence, neural networks, electrocardiogram

12.1 Introduction

The complex problem researchers face in any field is to predict the outcome with some certainty. It is essential to predict any disease and prevent it from affecting people in the medical area. Many high revenue-generating countries Worldwide have been spending billions of dollars on heart disease treatment per day. Therefore, a system is needed to detect cardiovascular disease early for the treatment of patients effectively, preventing a possible heart stroke. Heart diseases can be identified by wearing sensors throughout the human body and extracting the data by conducting medical tests. The physicians or doctors will examine the data generated by the sensors and diagnose the patients accurately. The main problem in wearable sensors is that signal artefacts may corrupt the data, i.e., data has more missing values, and more noise is present in the sensor device. These two problems impact the system degradation performance, and these sensor devices will generate inaccurate results. Electronic Medical Records (EMR) and sensor data analysis are challenging tasks for monitoring cardiac patients. The extraction of features from the data set is vital in predicting disease. Intelligent systems to predict heart disease automatically fuse the information from sensors and EMR.

The hybrid model is the combination of two models. The first model is the feature selection weighting approach to recognize the exact feature weight. The second model utilizes weights as input for machine learning method classifiers to predict disease exactly. Most heart data have invalid, repeated, or missing data, which means redundant and irrelevant features. This kind of data will create confusion among the doctors to define the target class. The time-consuming process is very high in handling this kind of data. The existing methods have heart disease diagnosis depending on the weighting of features methods. Due to the weight allocation, the uncertain combination

of operations may increase mean square error values and decrease the effectiveness of the predictive model. The theoretical support will not be present if the mean square error values are present. The specific weight needs to be applied to the features concerning classification models to avoid such situations. The latest technologies in computer science have brought many opportunities for medical researchers, and computer science is used as an instrument for medical professionals to predict diseases. Medical science with artificial intelligence has gained tremendous momentum since the last decade. Machine learning is such a tool used in different domains, and it works on datasets. The best part of machine learning modelling is that reprogramming is possible at any time if the results are not effective. It has much strength and has enormous opportunities for medical professionals. Heart disease prediction has significant challenges because of different parameters utilized for target values. The numerous data types for various conditions to predict heart disease have many approaches, such as Naïve Bayes, K-means nearest neighboring (KNN), Decision tree, Random Forest, and Neural networks, to indicate any disease with high accuracy. Each algorithm has its speciality for feature classification or accuracy. The neural network models have great potential to reduce the error in heart disease prediction.

The UCI dataset [1] is most predominantly used to evaluate heart attack, and it includes the data of 303 people. It has two classes: one class is for people with No Heart Disease, and the other is for heart disease. The authors use the binary cuckoo optimization algorithm to construct the feature selection and support vector machine. The accuracy achieved is 84.44%, sensitivity 86.49%, and specificity 81.49%. The following article [2] is a cascade correlation neural network [CNN] used to predict the disease. This work accuracy depends on 270 data samples, in which 150 samples are taken for training, and the rest are used to stimulate network architecture. CNN architecture depends on 13 input neurons and one output neuron, and obtained accuracy is 78% for training and 85% for testing with less time complexity. The deep belief network [3] for heart disease prediction for likelihood percentage is designed in MATLAB software, and the CNN method provides an accuracy of 82%.

The dataset's illustration for heart disease is usually raw and inconsistent. The pre-processing of the high-dimensional dataset is compulsory for any researcher to reduce it to a common dataset. The extraction of variables from the available dataset is needed to train the algorithm with less time complexity. Many research articles used time as a parameter and compared other parameters such as accuracy and efficiency. The performance of Sequential Minimal Optimization (SMO) classifiers has more efficiency than Multilayer Perceptron (MLP) classifiers [4]. The machine learning

methods have advantages and disadvantages, and feature optimization provides high Classification in decision tree efficiency [5]. Early detection of heart disease has feature utilization and includes research for medical professionals to perfect heart disease. This article [6] has collected raw data from electrocardiogram devices and used the dataset for training purposes in a neural network model to classify the patterns in the dataset. This article also shows that 95% is achieved with the support vector machine method and neural network for training data classification. It is also used to achieve better results. If the data are multidimensional and nonlinear, this neural network method provides high efficiency relative to other methods. The robust algorithm-based machine learning helps reduce unwanted noise in any dataset, and the redundant data can be eliminated. The deep learning algorithm has a high chance of increasing efficiency and has high accuracy for heart disease detection. The multilayer perceptron [8] with a backpropagation algorithm predicts cardiovascular disease accurately. The authors used artificial neural networks and obtained 88.46% and 80.17% for training and validation sets. Overall, the accuracy is increased by 1.1% for the training set and 0.82% for the validation dataset. The weighted associative classifier (WAC) [9] predicts heart disease with a Guided user interface (GUI) interface connected to the patient records. More than 13 attributes and 303 features are used for testing and training.

The target values are chosen in binary format if it is 1, the patient has heart disease. Similarly, if it is 0, the patient does not have heart disease. The classifier used here is Classification based on multiple association rules (CMAR), Classification based on Associations (CBA), and Classification based on predictive association rules (CPAR), and the authors obtained an accuracy of 81.51%. The article [10] supports the innovative decision system for heart disease prediction. Mitral stenosis and ventricular septal defects are three heart diseases predicted using an artificial neural network decision system. The sound of the heart is gathered for devices such as a stethoscope and microphone. The microphone device is placed between the stethoscope and PCI card to strengthen the sound signals. This ANN-based system identifies the three types of heart disease accurately. The principal component analysis (PCA) [5] and regression technique are mainly used for feature selection, and the prediction accuracy is around 92% is achieved for the heart disease dataset. The feed-forward neural network achieves 95% efficiency. The feed-forward back propagation neural network is used to predict the disease, and it consists of 13 input layers, 20 hidden layers, and one output layer. This network achieves 88% accuracy with 20% of the testing dataset, 20% of the validation dataset, and 60% for training purposes.

The article uses feed-forward multilayer perceptron [12] and supports vector machines to obtain 85% and 87.5% accuracy, out of which 270 samples are used. The division ratio used here is 60 to 40. Samples used for training is 162 and for testing is 108. Target values are usually the same (0,1). In [13], the authors used an entropy ensemble of neural networks with recursive feature elimination [EENNRFE] algorithm for feature extraction. The data set used by the authors is Cleveland from the UCI database. The correlation coefficient is found among the feature classification and, with the help of artificial neural networks, will measure the mean of an ensemble in neural networks. The accuracy of 85.66% is obtained for training data.

Deep learning is used to design a prediction model in medical science. It becomes challenging to achieve higher precision due to the unavailability of values in specific data set fields. The missing values are identified using a systematic methodology. The Cleveland dataset for heart disease is used to test the different models in this book chapter. The authors used the imputation methods for classifying the dataset and identified the MICE imputation method for missing value distribution [27]. Researchers worldwide were working on machine learning algorithms to reduce the death rate. The researchers or doctors entirely depend on the accuracy. The authors used hybrid methods, i.e., combining more than two methods to detect disease with increased accuracy with the importance of feature selection. The two methods are radial basis function and genetic algorithm, and the accuracy increased to 94.2% with nine different characteristics for attribute reduction [28]. The dataset contains some irrelevant features, and the authors used the isolation forest method for improved results. The confusion matrix is used to analyze 14 main attributes in the UCI dataset. The authors achieved 94.2% accuracy using the deep learning approach [29]. Nearly 18 million people are affected by heart disease every year, and overall, 32% of the population suffers from death worldwide. So to identify the disease at an early stage is mandatory for doctors. The authors used the UCI heart disease dataset to determine kernel values. The kernel has a low correlation coefficient and a less mean absolute error, concluding that the proposed model is comparatively good [30]. The authors designed a health care application system that depends on an optimal artificial neural network to diagnose heart disease. The proposed method uses two types of processes, i.e. the first process is distance-based misclassified instance removal, and the next one is learning-based optimization. They designed using the Apache spark method for training and testing data [31]. The dataset has many clinical parameters and has proposed a hybrid decision system. With the help of Python language, the

hybrid decision system was successfully implemented in the simulation environment, and an accuracy of 87% was achieved by the authors [32]. The authors used the hybrid method, a combination of random forest and decision tree, and achieved an accuracy of 89% [33]. The artificial intelligence methods are high-speed growing techniques for predicting heart disease. Data mining techniques efficiently provide disease identification based on patient data such as age, chest pain, and blood pressure [34]. Heart diseases can be reduced by early diagnosis methods with the help of using machine learning models. The authors used the Cleveland dataset to pre-process data and new hybrid classifiers for the training process. The sensitivity and precision FI score is calculated along with pessimistic prediction and false rate. The authors achieved high accuracy of around 99.05% using RFBM and relief feature selection methods [35]. Type 2 diabetes in older adults will likely result in heart diseases. The authors proposed a risk prediction model which uses a machine-learning algorithm to identify heart disease in type 2 diabetes patients. The classifiers' accuracy is in the range of 79% to 88% [36]. The four different types of machine learning models are compared with accuracy and macro-FI as performance metrics. The dataset was collected from 17661 patients. The authors concluded that features such as age and oxygen saturation play a significant role in heart disease prediction [37]. The machine learning models detect the receiver operator curve based on cardiovascular disease. The detected CVD less various levels of intervals for different machine learning algorithms [38]. The authors used ten-fold cross-validation for twenty-four variables based model developing in the training set [39]. The deep neural network is used for the risk prediction model with 834 patients with type two diabetes conditions [40]. This proposed research utilizes 14 features for predicting heart disease with various comparative analyses of machine learning algorithms [41]. The patient has 15 years of suffering from heart disease, and his mortality risk is identified using four machine learning models developed and compared to conventional methods. The accuracy is increased to 5.2% with Australian cohorts [42]. The classifiers are based on many machine learning models from the fifth Korea National health survey dataset. The authors developed models based on age groups for giving the dose of influenza [43]. The proposed model effectively detects the abnormal left ventricular geometry at the early stages based on different features such as age, body mass index and hypertension [44]. The authors conducted the study based on Qatar Biobank's most extensive collection of biomedical measurements and various factors of heart disease patients. The association of the proposed risk factors and comorbidities must be investigated in clinical setup better to understand their role in

CVD [45]. This Review explains the clinical prediction models are analyzed to develop the machine learning counterparts [46]. The best average prediction accuracy is used as the performance metric for various machine learning algorithms for various studies. This proposed research work also validates the area under the receiver operating characteristic curve [47]. The research develops a heart disease prediction based on the ensemble method of multilayer dynamic systems in every layer [48]. The review article has designed a system to integrate the data from omics data to approach the full potential [49]. The artificial intelligence-based heart disease method is developed using a support vector machine with kernel techniques [50]. The cardiovascular disease prediction has made researchers contribute work on many machine learning algorithms. In the medical field, machine learning produces valuable patterns and provide huge helpful information for the researchers, and this study analyzes the documents to provide the disease dataset [51–56].

12.2 Proposed System Model

12.1.1 Decision Tree Algorithm

This algorithm will predict the output based on a tree-type structure. Each branch/set is further divided into subsets to reach the decision finally. It works on nested structures, i.e., each node divides the data into either left or right direction. We imported the required libraries to analyze data, to build the predictive model using different techniques like linear regression, logistic regression, etc. Some of the python libraries used are NumPy, pandas, matplotlib, and seaborn are used for Exploratory Data Analytics [EDA] and Data visualization.

For pre-processing the data, the model building we used is SCIKIT-learn which has StandardScalar, MinMaxScalar, and Robust scalar to rescale the numerical values. MinMaxScalar is used to rescale the values between 0 and 1. StandardScalar is used to bring the data points so that the whole data is around mean 0.

With the help of the panda's library, the excel sheet data are imported into python and stored in a data frame called "df" and checks whether there are any duplicates in the data set and removes drop_duplicates syntax considering all the features into account. Typically, any dataset needs to be split into training and testing sets before modeling. In this proposed model, a four-fold cross-validation scheme is used to divide the training set into four folds and use three folds to train and one fold to validate. 80%

of the data set is used for training, and the remaining 20% is used for testing. Training also involves a validation set.

Basically, in the UCI repository, the data set has 1024 × 14 rows and columns, i.e., 1024 patients with 14 features are considered in our proposed system, with duplicates. Later the data is cleaned, i.e., we have removed the duplicates, the matrix size is converted to 302×14. Initially, the data is divided into two data frames so that one data frame should have all the variables except the target variable. The other should have only the target variable, i.e., the target variable is popped out and assigned to a data frame Y with all the existing rows of the original data frame and has a matrix size of 302×1 (302 patients with one target value). Table 12.2 shows the data variables, and their values are represented in df format. Q2

Similarly, the remaining variables are assigned to data frame X with all the existing rows with a matrix size of 302×13, i.e., 302 patients with 13 features except the target variable. Table 12.2 shows the features imported in df format. Now it is time to split the data into training and testing sets. Imported train_test_split from sklearn.model_selection library helps split X and Y data frames into train and test sets. As mentioned earlier, the data was split in the 80:20 ratio, i.e., train_size=80% of rows (20% given to validation) and test_size is 20% of rows. After splitting the data, X_train and X_test sizes are (241x13) and (61x13), respectively.

After setting up the data frames, instantiate DecisionTreeClassifier as df format by importing it from sklearn.tree library and fit the data. We instantiated the decision tree classifier with only one hyperparameter, the max_depth = 3, and kept all the other parameters as default. Graphviz software and imported pydotplus library to visualize the tree graph libraries and packages have been imported installed to observe a decision tree graph. We were using export_graphviz, which is imported from sklearn.tree library. We defined the class_names as No Disease and Disease. So, with the help of pydotplus and Graphviz libraries, We obtained a Decision tree graph by considering only max_depth(3) as a predefined parameter, as shown in Figure 12.1.

From sklearn. The metrics import confusion matrix gives the information about the correctly predicted and wrongly predicted values compared to the actual values.

12.1.1.1 Confusion Matrix

This Confusion matrix technique will summarize the performance of any machine learning algorithm. Therefore, other parameters like sensitivity, specificity, precision, recall also play an essential role in understanding the effectiveness of an algorithm.

Table 12.1 UCI dataset for predicting heart disease.

Data type	Variable name	Input attributes	Specification/range	Predictable attribute
Numerical	AGE	Age	Above 50 [50-79]	Target Values Value 1=< 50% [No heart disease] Value 0 => 50% [patient suffering from heart disease]
Nominal	SEX	Sex	1-Male 2-Female	
Nominal	CP	Chest pain	1. Angina 2. typical type angina 3. non-angina pain 4. asymptomatic	
Numeric	TRESTBPS	Resting blood pressure	(mm Hg)	
Numeric	CHOL	Serum cholesterol	(mg/dl)	
Numeric	FBS	Fasting blood sugar	>120 mg/dl 1. >120 mg/dl	

(*Continued*)

Table 12.1 UCI Dataset for predicting heart disease. (*Continued*)

Data type	Variable name	Input attributes	Specification/Range	Predictable attribute
Nominal	RESTECG	Resting electrocardiographic results	0. Normal 1. Suffering from ST-TWave abnormality 2. Definite leftventricular hypertrophy	Target Values Value 1=< 50% [No heart disease] Value 0 => 50% [patient suffering from heart disease]
Nominal	THALCH	Heart rate maximum	Beats Per Second	
Nominal	EXANG	Exercise-induced angina	1.Yes 0. No	
Numeric	OLDPEAK	Old peak	ST depression induced by exercise relative to rest	
Nominal	SLOPE	The slope of thepeak exercise ST segment	1. Unsloping 2. Flat 3. Down sloping	
Numeric	CA	No of major vesselscolored by fluoroscopy	Values rangingfrom 0-3	
Nominal	THAL	Thal	3. Normal, 6. Fixed defect, 7. Reversible defect	

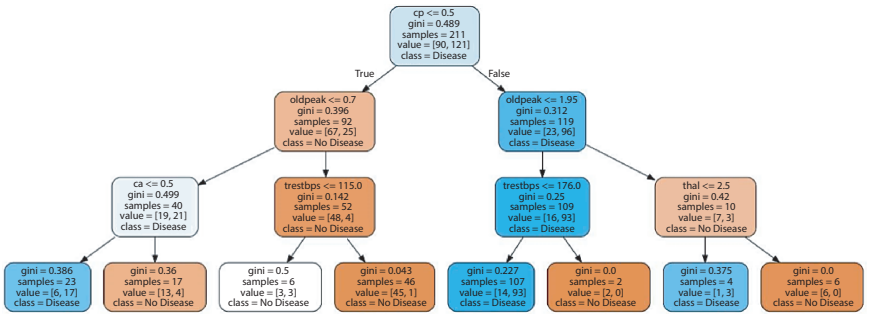


Figure 12.1 Decision tree graph for maximum depth of 3.

The confusion matrix output for a sample decision tree is shown below

$$\textit{Training set Confusion Matrix} = \begin{bmatrix} 79 & 27 \\ 10 & 125 \end{bmatrix} \tag{12.1}$$

Q3

$$\textit{Testing set Confusion Matrix} = \begin{bmatrix} 24 & 8 \\ 5 & 24 \end{bmatrix} \tag{12.2}$$

Accuracy can be calculated as the ratio of correctly predicted labels to the total number of labels, the training accuracy is 84.64%, and the testing accuracy is 78.68% which is less efficient. So, to improve the algorithm's effectiveness, hyperparameter tuning is done and with GridSearchCV, imported from sklearn.model_selection, best hyperparameters were obtained. The entropy criterion is better than the Gini index, maximum depth of tree to be five and the minimum sample per leaf to be 10. Then instantiated the decision tree on best hyperparameters and fit the training and test sets model. The confusion matrix is being obtained after fitting the model. The confusion matrix calculation as shown in and calculated values for training and testing are shown in equations (12.1), (12.2), (12.3) and (12.4).

$$\textit{Training set Confusion Matrix} = \begin{bmatrix} 90 & 16 \\ 16 & 119 \end{bmatrix} \tag{12.3}$$

$$\textit{Testing set Confusion Matrix} = \begin{bmatrix} 24 & 8 \\ 7 & 2 \end{bmatrix} \tag{12.4}$$

The training and testing accuracy obtained is 86.72% and 75.40%, respectively, which is again a poor test accuracy and ineffective in predicting heart disease correctly. So, the random forests technique could be a better alternative to the decision tree, an ensemble of decision trees or an ensemble of different predictive algorithms.

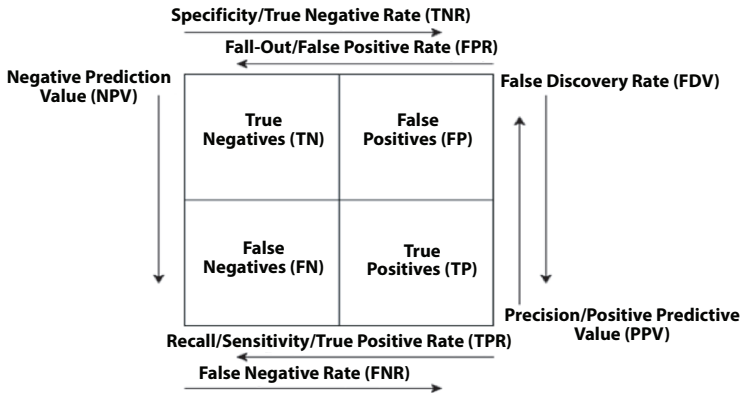
12.2 Random Forest Algorithm

The Random Forest Classifier is being imported from sklearn.ensemble. Using GridSearchCV, imported from sklearn.model_selection, best hyperparameters were obtained. The best hyperparameters obtained are maximum depth = 10, maximum features =4, min_sample_leaf = 5, number of estimators=10. Later, Instantiated grid_search and fit the model considering the best hyperparameters. We obtained a confusion matrix for the random forest model.

The confusion matrix values for training and testing are shown in equations [12.5] and [12.6].

$$Training\ set\ Confusion\ Matrix = \begin{bmatrix} 93 & 13 \\ 17 & 118 \end{bmatrix} \tag{12.5}$$

$$Testing\ set\ Confusion\ Matrix = \begin{bmatrix} 27 & 5 \\ 1 & 28 \end{bmatrix} \tag{12.6}$$



Q4 **Figure 12.2** Confusion matrix.

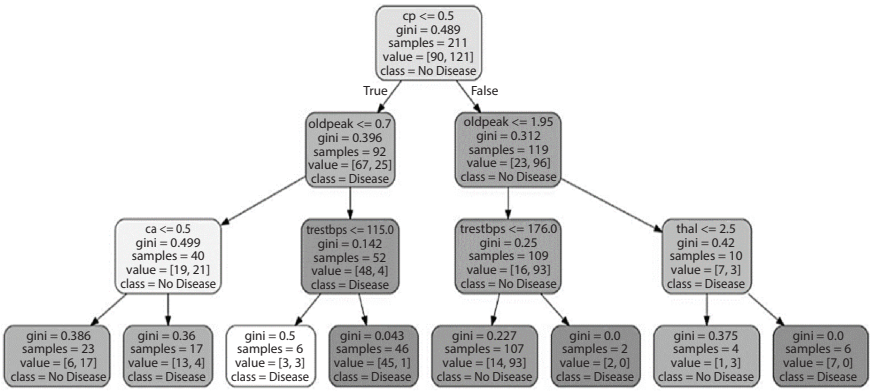


Figure 12.3 Decision tree with best hyperparameters.

From that, the training accuracy comes out to be 87.67%, and the test set accuracy is about 90.16%, the best accuracy any machine learning model can get.

Here, False Negatives are more dangerous than False Positives. Since we know, the model predicting not having a disease when people have the disease is much more complicated than the opposite case. Therefore, the recall percentage must be as close to 100% as possible. In this case, recall is around 96.55%, which is an encouraging factor. So, we say our model predicts that having heart disease with a 96.6% certainty helps them prevent death. The Random Forest algorithm, as shown in Figure 12.4 hence proved extremely reliable to predict heart disease before the attack. Furthermore, this helps people to take necessary actions to prevent fatal heart stroke.

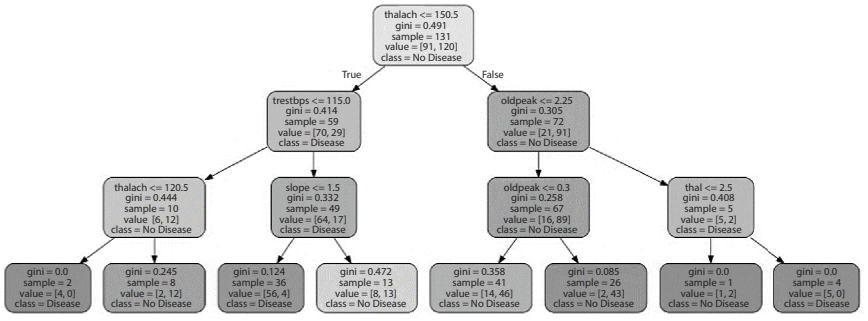


Figure 12.4 Random forest with best hyperparameters.

Table 12.2 Features imported into "df" data form.

	Age	Sex	cp	trestbps	Chol	Fbs	Restecg	Thalch	Exang	Oldpeak	Slope	Ca	thal	Target
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0

12.3 Variable Importance for Random Forests

From the tree diagram for random forests, it is seen on the top that cp(chest pain) is the most important feature of all, followed by thalach, ca, thal (thalassemia), exang, sex and so on. Table 12.1 shows the list of variables in the dataset. Chest pain (Cp) has four different values representing 0, 1, 2, 3, and a specific meaning. From the random forests tree diagram, it is clear that if the value is 0, most people will not have heart disease. If cp <= 0.5 is true, the value is 0. Else the condition is false, and then the value can be either 1 or 2 or 3. Let us consider whether the condition is false, as most people have heart disease when the condition is false. Now if cp <= 0.5 is false, we reach one more condition, thal <= 2.5, that means thalassemia. Thal feature too has four different values (0, 1, 2, 3). From the random forests, we can say that if the people with a thal value greater than 2.5 and cp>= 0.5, it can be confirmed that they have heart disease. Now, if thal <= 2.5, there might be no heart disease. After that, we reach a condition, thalach <= 118.5, which means the maximum heart rate achieved in beats

Table 12.3 Variable representing feature importance.

Variable number	Variable name	Importance
2	CP	0.238423
7	THALCH	0.137478
11	CA	0.135503
12	THAL	0.120802
8	EXANG	0.092325
1	SEX	0.060381
9	OLDPEAK	0.049703
3	TRESTBPS	0.046792
0	AGE	0.042766
10	SLOPE	0.031194
4	CHOL	0.029229
6	RESTECG	0.015404
5	FBS	0.00000

per minute(bpm). If that condition is true, then it can be concluded that the person has heart disease, and if not, the probability of a person having heart disease is very less. This way, we understand the random forest ensemble simply by looking at the graph. Table 12.3 shows the variable importance for the random forest algorithm.

12.4 The Proposed Method Using a Deep Learning Model

Since our model will work only for older people, we converted the data with more senior people by sorting the age above 50. We have seen 230 patients from the UCI repository. After Ensembling, the training data is 80%, and data are 20% for testing. The duplicates are removed. The total data from the UCI repository has 1025 entries or patients. We have sorted the patient's age above 50. We got 732 entries. Even after removing the duplicates, 216 entries are used. The number of features we have used to predict heart disease is 13. From there, we generated 60 features, i.e., by combining the elements by multiplication or division or adding, 60 features are used to predict heart disease with 216 entries. Using machine learning techniques like Pearson correlation, linear support vector classifier (SVC), Lasso, select K best with chi-z, random forest method, and variance threshold algorithms. From the 60 features, only 26 are selected as final features. So, therefore, the matrix size is converted to 216 × 26, i.e., 216 patients and 26 features.

In this proposed model, 4096 neurons are used in the first layer, and each neuron has a certain number of inputs with a bias value. Each signal is

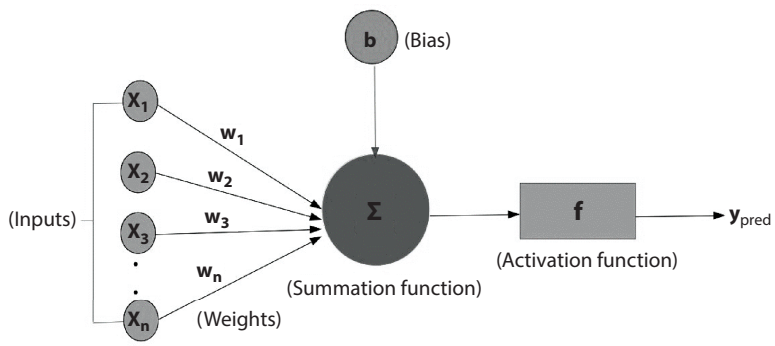


Figure 12.5 Deep learning model.

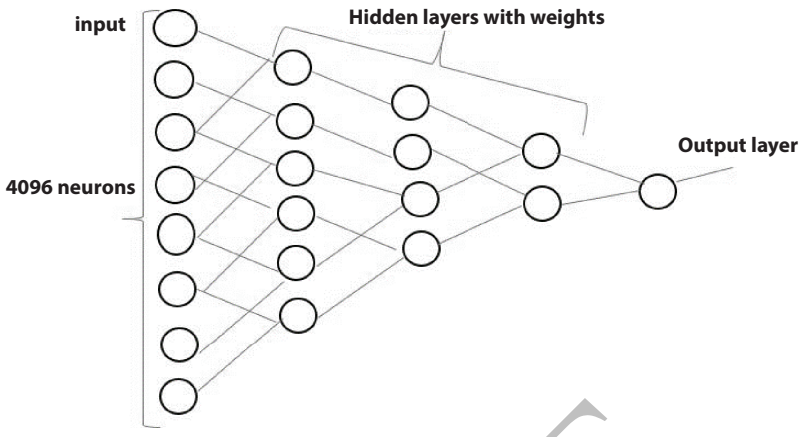


Figure 12.6 Applied deep neural network model with 4096 neurons on the first layer.

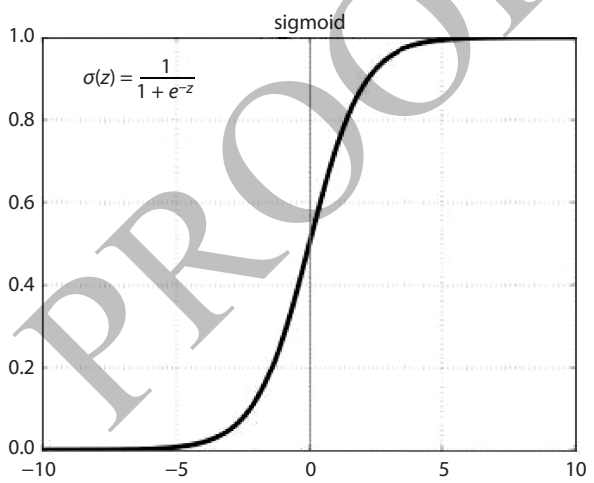


Figure 12.7 Sigmoid activation function range between 0 and 1.

multiplied by a weighted value. Here 4096 neurons are used, so it has 4096 neurons, and values are adjusted during training time. As shown in Figure 12.6, the hidden layers connect one neuron in the input layer to another neuron in the second layer. Each connection always has weight values. The training algorithm updates the weight value to reduce the errors. The activation used in Figure 12.5 has nonlinearity to neural networks. This activation function always squashes the values in a smaller range with a sigmoid function and has values between 0 and 1, as shown in Figure 12.7.

12.4.1 Prevention of Overfitting

We have two methods to prevent the overfitting problem in deep learning.

Q6 **12.4.2 Batch Normalization**

12.4.3 Dropout Technique

Here the number of classes is only two, i.e., the target values are 0 or 1. To calculate the losses in the last layer, we have used the sigmoid function.

Since the binary cross-entropy and backpropagation method is used to adjust the weights, the data's training is 94%. Using the confusion matrix, all ML models are used to compare the accuracy, nearly 20 models are tested, and accuracy is obtained by 94%.

12.5 Results and Discussions

Table 12.4 shows that the top 20 models for heart disease prediction are compared.

Each method has its unique feature.

12.5.1 Linear Regression

Linear regression performs supervised learning for the given data. It is based on a statistical method primarily used for predictive analysis. This algorithm makes continuous predictions and has a decent relationship between dependent variables or independent variables. The dependent variable is kept on changing for the independent variable; hence, it is called linear regression. Generally, the plot between the independent and dependent variables is a sloped straight line, which shows that the direct relationship between the variables is linear. The linear relationship can be positive or negative. The mathematical equation for the linear regression is shown in below equation

$$Y = a+bX1+cX2+dX3\ldots. + \mathcal{E} \tag{12.4}$$

X1, X2, X3 are explanatory variables, Y is the dependent variable, and $\mathcal{E}$ is the random error. If the data is wrongly categorized, errors will occur in the algorithm's accuracy. A linear regression algorithm can easily categorize

Q7 **Table 12.4** 20 models comparison for accuracy training and testing.

Model	Acc_train	Acc_test	Acc_diff
Deep Neural Network model	**97**	**94**	**3**
XGB Classifier	98.84	86.05	12.79
Ridge CV	83.72	86.05	-2.33
Random Forest Classifier	83.14	86.05	-2.91
Gaussian Process Classification	99.42	83.72	15.7
Support Vector Machines	89.53	83.72	5.81
Linear SVC	84.88	83.72	1.16
Stochastic Gradient Decent	84.3	83.72	0.58
Gradient Boosting Classifier	100	81.4	18.6
LGBM Classifier	96.51	81.4	15.11
AdaBoost Classifier	87.79	81.4	6.39
KNN	87.79	81.4	6.39
Extra Trees Classifier	84.3	81.4	2.9
Logistic Regression	82.56	81.4	1.16
Naive Bayes	79.65	81.4	-1.75
Bagging Classifier	97.67	79.07	18.6
Voting Classifier	84.3	79.07	5.23
Decision Tree Classifier	88.95	74.42	14.53
Linear Regression	59.3	48.84	10.46

all the linearly separable datasets. The linear regression is again classified into two types, i.e., multiple and straightforward linear regression. A single independent variable is used to predict the single dependent variable, which comes under simple linear regression.

Similarly, more than one independent variable predicts the single dependent variable's value, which comes under multiple linear regression. Here the linear regression method indicates heart disease, an accuracy of 59.3% for

training data, 48.84% for testing data. The difference between the accuracy is very high, and it is not suitable for predicting heart data. Hence, the decision tree classifier improves accuracy, which is the better linear regression version.

12.5.2 Decision Tree Classifier

The decision tree algorithm has the human thinking ability to decide easily, and the essential logic behind this decision depends on the tree-like structure. The algorithm of the decision tree classifier has five steps, and, for the prediction of the class of a given dataset, the algorithm starts from the root node of any tree. This method compares the different values of root features or attributes with the exact dataset, comparing the branches and jumps to the next node. The first step is to begin the tree with the root node, which contains the dataset completely. The second step depends on attribute selection measures for finding the best attribute in the dataset. The third step is to divide the complete dataset into subsets with the best attributes. The fourth step is to generate the decision tree nodes. The final step is to generate different decision trees with subsets formed in the third step and continue this process until the classification of the nodes cannot be done, i.e., the leaf node is generated. Here, the decision tree classifier method predicts heart disease, 88.95% for training data and 74.42% for testing data. The difference between the accuracy is very high, and it is not suitable for predicting heart data. Hence, the voting classifier improves accuracy, which is a better version of the decision tree classifier.

The decision tree algorithm has the human thinking ability to decide easily, and the essential logic behind this decision depends on the tree-like structure. The algorithm of the decision tree classifier has five steps, and, for the prediction of the class of a given dataset, the algorithm starts from the root node of any tree. This method compares the different values of root features or attributes with the exact dataset, comparing the branches and jumps to the next node. The first step is to begin the tree with the root node, which contains the dataset completely. The second step depends on attribute selection measures for finding the best attribute in the dataset. The third step is to divide the complete dataset into subsets with the best attributes. The fourth step is to generate the decision tree nodes. The final step is to generate different decision trees with subsets formed in the third step and continue this process until the classification of the nodes cannot be done, i.e., the leaf node is generated. Here, the decision tree classifier method predicts heart disease, 88.95% for training data and 74.42% for testing data. The difference between the accuracy is very high, and it is not suitable for predicting heart data. Hence, the voting classifier improves accuracy, which is a better version of the decision tree classifier.

12.5.3 Voting Classifier

The output of this model will be predicted based on the highest chance of the chosen class appearing at the output. This technique predicts the output class with the largest majority voting by merging many different elements of each classifier supplied into the voting classifier. The fundamental idea behind this model is to develop a single model that trains all of the individual dedicated models to find accuracy and forecast output based on the majority of votes for each output class. This model has two types of voting, the first one is hard voting, and the second one is soft voting. The prediction depends on the highest probability of the class, with the majority of votes coming under the hard voting category. In soft voting, the prognosis of the output depends on the average probability of the given class. In this book chapter, the majority voting technique based ensemble approach is used, which combines different types of multiple classifiers to increase the accuracy. Primarily based on the UCI dataset, the Cleveland coronary heart dataset includes 14 attributes, out of which 8 are specific attributes, and six are numerical attributes, and 303 patient statistics are discovered. Based on the voting classifier output prediction, an ensemble of vulnerable classifiers with sturdy classifiers using the most people voting technique improved the accuracy of vulnerable classifiers to a positive extent. The bagging classifier is chosen to improve accuracy—an accuracy of 84.30% for training data and 79.42% for testing data. The difference between the accuracy is minimal, and it is not suitable for the best prediction of heart data. Hence, the bagging classifier improves accuracy, which is the better version of the voting classifier.

12.5.4 Bagging Classifier

This model is an ensemble-based meta estimator that fits base classifiers on the original subset's random subsets and then aggregates their predictions to form a final prediction. This meta estimator reduces the variance of the decision tree with randomization. This classifier eases the variance by voting, and the drawback is that it increases bias, which reduces variance. It combines the predictions from many decision trees. Another name for the bagging ensemble algorithm is bootstrap aggregation. It works on the bootstrap sample, a dataset sample with replacement, i.e., the sample taken from the dataset is replaced, which allows the sample to be selected again and even multiple times in the new sample. It provides an objective approach to estimating the statistical quantities in the dataset and creating an ensemble-based decision tree model. The bagging algorithm is used to identify the

warning signs of heart disease in patients. This method generates multiple versions of any predictor by making bootstrap replicates of learning sets and using these as new learning sets. This book chapter conducted tests on actual and simulated datasets available in the machine learning repository. It used different classification and regression trees with specific subsets of data. The bagging classifier has shown substantial accuracy gains, 97.07% accuracy for training data, and 79.07% accuracy for testing data. The difference between this accuracy is very high; hence, it is not suitable for best heart disease prediction, so we have used the Naïve Bayes method to improve accuracy, which is a better version than the bagging classifier.

12.5.5 Naïve Bayes

One of the supervised learning algorithms depends on the Bayes theorem for solving classification problems for any dataset. Many researchers use this model for text classification, which includes high dimensional training dataset. This model can make quick predictions. This algorithm is comprised of two words, i.e. Naïve and Bayes. Naïve comes because of the assumption that the occurrence of a particular feature is independent of other features. Bayes comes because this model depends on the principles of the Bayes theorem. The main advantages of this model are that it performs well in multiclass predictions. The main disadvantage is that it assumes all features are independent or unrelated to each other so that it cannot learn the relationship between characteristics. In this book chapter, the Naïve Bayes classifiers predict heart disease with the fastest probabilistic classifier, especially for the training phase. The feature selection is a process of removing the irrelevant features from the dataset. This process involves three classes, i.e. filter, wrapper and embedded method. Here the medical dataset is based on types present or absent. This model could classify 79.65% of input instances correctly, and the remaining are incorrect instances. Based on the Naïve Bayes algorithm, diabetic patients with high cholesterol values are in the age group of 45-55, the bodyweight of humans is 60-71, and blood pressure value is 148-230. Furthermore, 79.65% of the accuracy for training data and 81.4% of the testing data are obtained. This accuracy is negligible; hence, it is not suitable for best heart disease prediction, so we have used the logistic regression method to improve accuracy, a better version than the Naïve Bayes classifier.

12.5.6 Logistic Regression

It is also a supervised learning algorithm; it also uses the probability of the target variable to predict the outcome. The dependent variable is always

binary. i.e., either 1 or 0. Logistic regression has binary target variables. It can be divided into several categories: binomial, multinomial, ordinal, etc. the logistic regression method can apply to both datasets, i.e., continuous and discrete datasets. It will classify the observations using different data types to find efficient variables. In this book chapter, the logistic regression model is applied to predict heart disease, based on the logistic results if the probability is greater than 0.05, which shows a less statistically significant relationship with the outcome of the dataset. The logistic regression method predicts heart disease, 82.46% for training data, 81.2% for testing data. The difference between the accuracy is shallow, and it is not suitable for predicting heart data. Hence, the extra trees classifier is used to improve accuracy.

12.5.7 Extra Trees Classifier

This method is also an ensemble technique formed by combining multiple decor-related decision trees in the forest to output its classification result. The only difference between random forest and this extra tree classifier is that each decision tree in the different trees' forest is built from the original training sample. Another name of this algorithm is highly randomized trees. It is mainly related to the decision trees such as bootstrap and decision tree algorithms. Here the extra tree classifier method predicts heart disease, 84.3% for training data, 81.2% for testing data. The difference between the accuracy is significantly less, and it is not suitable for predicting heart data. Hence, the KNN improves accuracy, which is the better version of the extra tree classifier.

12.5.8 K-Nearest Neighbor [KNN] Algorithm

The improvement in artificial intelligence or deep learning methods plays a significant role in identifying how people getting heart disease in the earliest stages is possible with the possible dataset. Here, the KNN algorithm is a supervised learning method and initially assumes the similarity between the new data and already available data and puts the new data into the most similar category to the general data categories. It is also called the non-parametric algorithm, i.e. this algorithm does not consider any assumption data. Sometimes it is also called a lazy learner algorithm, i.e., it will not learn from the training dataset immediately; instead, it acts only on the dataset at the time of Classification. The KNN algorithm mainly helps the identification of a category or any class of a particular dataset. It will work by selecting number K of the neighbors and Euclidean distance of number K. Then, take the K nearest neighbors per the already calculated Euclidean

distance. This algorithm is preferred only if all the features are continuous. It is also chosen based on its high speed of convergence.

The dataset contains a huge number of features that can be used to predict cardiac disease. The majority of the dataset has noisy characteristics. Any classifier's performance will be lowered when the dataset contains more and more noisy features. In this book chapter, the irrelevant dataset is being removed by cleansing. The features, such as weight, height, and ap_lo, are removed in this scenario the KNN provides the accuracy of 61%. This algorithm provides higher accuracy for higher values of K, the improvement in accuracy was observed from 61% to 65% for values of K from 1 to 5. Apart from the value of K, some other features also play a significant role in the excellent accuracy. The main drawback of the KNN algorithm is with large datasets and more considerable dimensional data. The distance calculation between each data point is complicated for large datasets. The KNN machine learning method predicts heart disease, 87.31% for training data, 81.2% for testing data. The difference between the accuracy is significantly less, and it is not suitable for predicting heart data. Hence, the Adaboost classifier improves accuracy, which is the better version of the extra tree classifier.

12.5.9 Adaboost Classifier

It is also known as an adaptive boosting technique. It is mainly used as an ensemble method in machine learning. The weights of the branches are assigned to every instance. It is also used to reduce the bias and variance in supervised learning. Initially, the number of decision trees is made in the mandatory training period of data boosting. So it was developed for binary Classification and can be used to increase performance. The weak models are added one by one, with the weighted training data used to train them. The process of adding soft models continues until the training dataset can no longer be improved. The weak classifiers' prediction or detection of heart disease is made possible via a weighted average. Boosting is an ensemble language for building a reliable classifier. It creates a model using training data and then creates a second model to rectify the first model's inaccuracy. The AdaBoost classifier is a very short decision tree that uses only a single split, i.e. decision stump. Here the Adaboost classifier method predicts heart disease, 87.31% for training data, 81.2% for testing data. The difference between the accuracy is significantly less, and it is not suitable for predicting heart data. Hence, the LGBM classifier improves accuracy, which is the better version of the Adaboost classifier.

12.5.10 Light Gradient Boost Classifier

This algorithm depends on a decision tree algorithm and can predict the accuracy of heart disease with high accuracy. It is also an open-source library, and it extends the boosting gradient algorithm and provides high, more significant gradients. It can give a high predictive performance for complex data. The ensembles are created using decision tree models, and these trees are added one by one and fit the corrected prediction errors. The gradient boosting method speeds up learning and reduces computational complexity. The LGBM classifier method predicts heart disease, 96.31% for training data, 81.2% for testing data. The difference between the accuracy is very high, and it is not suitable for predicting heart data. Hence, the Gradient Boosting classifier improves accuracy, which is the better version of the Light Gradient Boost classifier.

12.5.11 Gradient Boosting Classifier

This classifier depends on three elements, i.e. loss function, weak learner and additive model. The loss function solves different problems, and the issues may be differentiable or squared errors. The process determines the differentiable loss for the framework. The weak learner is decision trees, and the regression trees are built greedily, with split points depending on purity scores such as Gini. The Additive model is the final component, in which all trees are added one at a time, with no changes to previous trees. It is used to reduce the loss after the trees have been added. The parameters are also minimized, such as coefficients in weights or regression equations in neural networks. The weights are updated automatically to reduce the error. The parameters of the tree depend on functional gradient descent. This classifier depends on four enhancements: tree constraints, random sampling, shrinkage, and penalized learning. The LGBM classifier method predicts heart disease, 100 % for training data, 81.2% for testing data. The difference between the accuracy is very high, and it is not suitable for predicting heart data. Hence, the Stochastic Gradient Descent Algorithm improves accuracy, the better version of the Gradient Boost classifier.

12.5.12 Stochastic Gradient Descent Algorithm

This algorithm is mainly used to find the values of coefficients of a function which reduce the cost function. It is primarily used to search for an optimization algorithm. It is also a supervised learning-based machine learning

algorithm and the best way to find the target function which maps input data and output variables. Many algorithms have many representations with different coefficients, and the optimization process is entirely different for machine learning algorithms. The cost function plays a significant role in evaluating the coefficients in any model by prediction calculation and training dataset. It even compares the actual output values predictions and calculates average error. The cost is calculated for each coefficient and can be updated for every dataset with each iteration in the gradient descent algorithm. Each iteration is called a batch, so the other name is called batch gradient descent. The Stochastic Gradient Descent algorithm predicts heart disease, 84.31% for training data and 83.2% for testing data. The difference between the accuracy is significantly less, and it is not suitable for predicting heart data. Hence, the linear support vector classifier improves accuracy, the better version of the stochastic gradient descent algorithm.

12.5.13 Linear Support Vector Classifier

This classifier is like SVC, but the parameter kernel is linear. Sometimes the kernel can be non-linear also. The linear SVC uses one vs the rest classifier wrapper. It performs Classification, and it performs well with a large number of samples. It uses penalty normalization and loss function apart from the support vector classifier algorithm. The Linear support vector classifier method predicts heart disease, 84.31% for training data, 83.2% for testing data. The difference between the accuracy is negligible, and it is not suitable for predicting heart data. Hence, the support vector machine improves accuracy, which is the better version of the linear support vector classifier.

12.5.14 Support Vector Machines

The backing vector machine is a managed learning framework utilized for arrangement and relapse issues. Numerous individuals incredibly prefer backing vector machines as it produces eminent rightness with less calculation power. It is generally used in order issues. We have three sorts of learning: directed, unaided and support learning. A help vector machine is a specific classifier officially characterized by separating the hyperplane. This is the most popular supervised learning algorithm used for regression problems. The main aim is to create the decision boundary of n-dimensional space into new classes. Usually, the decision boundary is called a hyperplane. The support vector machine mainly helps in choosing the vectors to create the hyperplane. The data points that are always close

to the hyperplane and affect the hyperplane's position are termed a support vector. These hyperplanes are decision boundaries and can be helped in the Classification of data points. The support vector machine is classified into two types, i.e., linear and non-linear. The linear support vector machine mainly separates the data into linear wise, called a linear SVM classifier.

Similarly, the nonlinear SVM is primarily used for nonlinearly separated data. The Support vector machine predicts heart disease, 89.31% for training data and 83.2% for testing data. The difference between the accuracy is significantly less, and it is not suitable for predicting heart data. Hence, the Gaussian process classification algorithm improves accuracy, which is the better version of the support vector machine.

12.5.15 Gaussian Process Classification

Gaussian processes are mostly generalized to the probability distribution. It is used for complicated non-parametric machine learning algorithms. This Classification is the type of kernel model and can predict the probabilities of the class. The Gaussian process classifier method predicts heart disease, 99.31% for training data, 83.2% for testing data. The difference between the accuracy is very high, and it is not suitable for predicting heart data. Hence, the random forest classifier improves accuracy, the better version of the Gaussian process classification.

12.5.16 Random Forest Classifier

Random Forest is a well-known artificial intelligence calculation that is used in the implemented learning approach. In machine learning, it is commonly used for both classification and regression problems. It is based on the concept of gathering realization, which is a cycle of combining multiple classifiers to deal with an unforeseen situation and work on the model's exhibition. Random Forest is a classifier that uses the normal to operate on the prescient precision and contains numerous choice trees on different subsets of that dataset, as the name suggests. Instead of depending on one choice tree, the arbitrary backwoods take the expectation from each tree and because of the more significant part votes of forecasts, it predicts the final output. The more prominent number of trees in the timberland prompts higher exactness and forestalls the issue of overfitting. Since the random forest consolidates numerous trees to foresee the class of the dataset, it is conceivable that some decision trees may anticipate the correct output while others may not. In any case, together, every one of the trees foresees a suitable yield. In this manner, beneath are two suspicions for a

superior Random Forest classifier. The Random Forest classifier method predicts heart disease, 83.31% for training data, 86.2% for testing data. The difference between the accuracy is significantly less, and it is not suitable for predicting heart data.

12.6 Evaluation Metrics

Different performance metrics were utilized to determine the Machine Learning model's efficiency, shown in Table 12.4. Table 12.4 shows three parameters and their formulas accuracy, recall. and RMSE. Figure 12.8 shows the comparison of 20 different machine learning models.

Feature importance is a technique that refers the score to certain input features and tells the user about predicting a target variable. Depending on the algorithm, many types of features are essential concerning variable parameters such as correlation scores, decision trees, and essential scores, which helps in improving the efficiency of the heart disease model. The feature importance score will provide high insight into the heart dataset and the most relevant feature to the target. The feature importance also helps predict the specific model and improves the model's accuracy. Figure 12.9 shows the feature importance of the variables used to predict heart disease.

Figure 12.10 shows the recall criteria for twenty machine learning algorithms, plotted for training and testing data. It is also called sensitivity. This performance parameter measured the patients with cardiovascular disease and diagnosed the deep learning model as having heart stroke. The patients with true positives and false negatives are actual positives, considering cardiovascular disease is true positives. The proposed deep neural network method and linear regression model have high recall value regarding other algorithms. The Gaussian process classification, gradient

Table 12.4 Performance metrics used for efficiency calculation in machine learning models.

Name	Description
Accuracy	$(TP + TN)/$ $(TP + TN) + (FP + FN)$
Recall (re)	$(TP)/(\ TP + FN)$
RMSE	$\Sigma.\ \lvert(xi - xi\hat{}\,)^2\rvert$

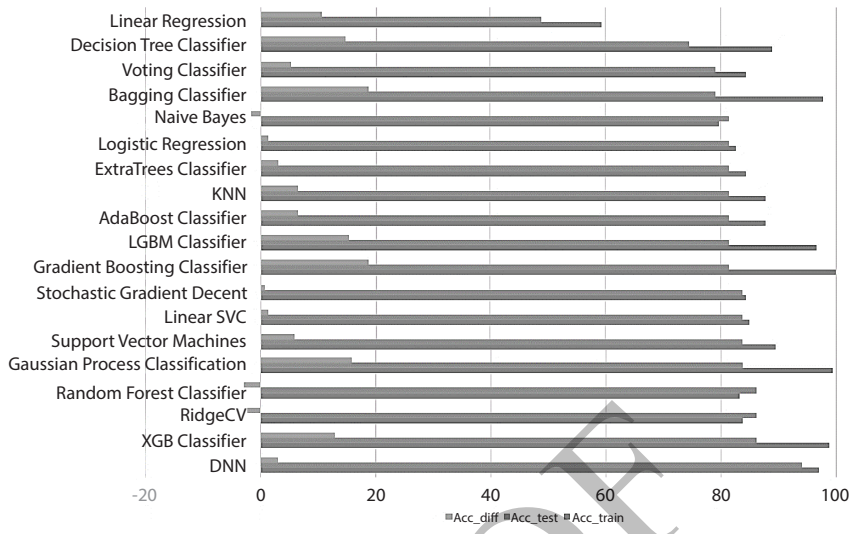


Figure 12.8 Comparison of 20 different models for accuracy of testing, training data.

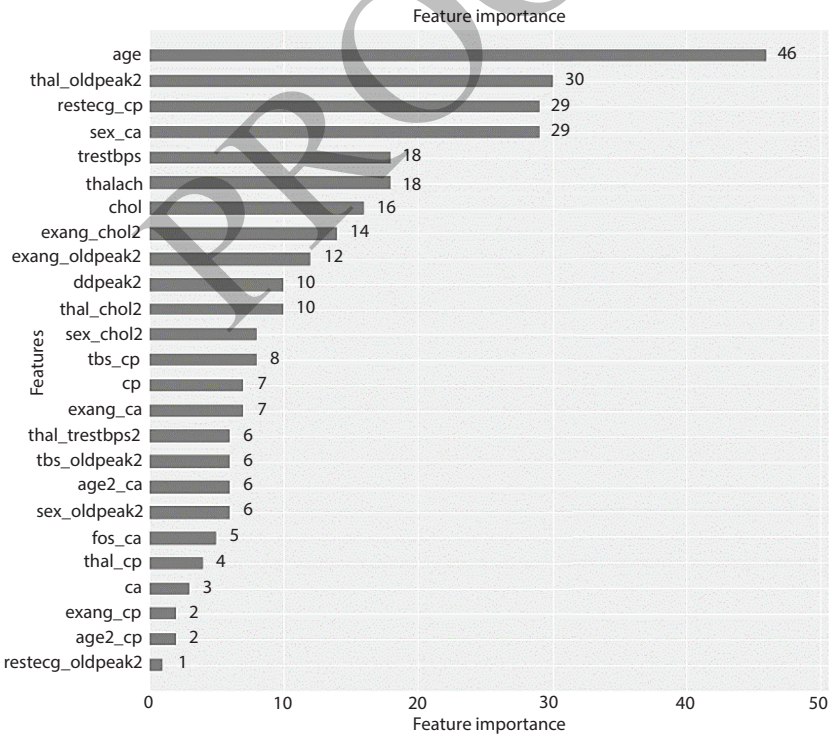


Figure 12.9 Feature importance of different variables.

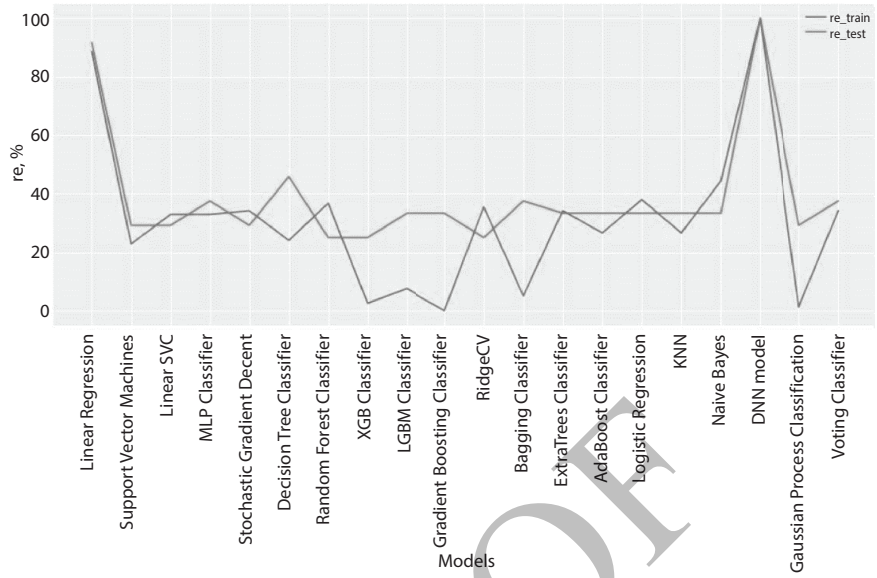


Figure 12.10 Recall criteria for 20 models for train and test datasets.

boosting classifier and bagging classifier has significantly less recall value for the training dataset.

Similarly, the deep learning neural network model has a very high recall value for testing datasets concerning other algorithms. The recall value is almost flat for 17 models, and it does not impact performance because the data is not predicted accurately by 17 models. The most negligible value is noticed in the decision tree classifier machine learning model.

Figure 12.11 shows the RMSE criteria for 20 machine learning algorithms, plotted for training and testing data. The thumb rule represents it. The RMSE normal range is around 0.2 and 0.5, i.e., if the model produces a value between 0.2 and 0.5, that model will predict the data accurately. The proposed deep neural network method and linear regression model have high RMSE values with other algorithms. The Gaussian process classification, gradient boosting classifier and bagging classifier has significantly less RMSE value for the training dataset. Similarly, the deep learning neural network model has a very high recall value for testing datasets concerning other algorithms. The RMSE value is almost 0.4 for the other 15 models, which means data are not predicted accurately. The Ridge CV classifier machine learning model notices the most negligible value.

Figure 12.9 shows the feature importance of variables used for heart disease prediction. Figure 12.10 shows the plot of recall for 20 different models

Q8 **Table 12.5** Detailed comparison of existing methods using heart disease dataset.

References	Year	Classification used	Evaluation metrics
Khemphila *et al.* [14]	2011	Multilayer perceptron with backpropagation machine learning	Accuracy: Training: 89.56%, Testing: 80.99%
Soni *et al.* [15]	2011	Weighted associative classifier	Accuracy: 81%
Ghwanmeh *et al.* [16]	2013	Artificial neural network	Accuracy: 92%
Santhanam *et al.* [17]	2013	Feed forward neural network	Accuracy: 95.2%
Ajam *et al.* [18]	2015	Feed forward back propagation neural network	Accuracy: 88%
Olaniyi *et al.* [19]	2015	Feed forward multilayer perceptron	Accuracy: Training: 85%, Testing: 87%
Roostaee *et al.* [20]	2016	Support vector machine	Accuracy: 84.4%
Karthikeyan *et al.* [21]	2017	Deep belief network	Accuracy: 90%
Silvia priscila *et al.* [22]	2017	Support vector machine with recursive feature elimination	Accuracy: 85.6%

(*Continued*)

Table 12.5 Detailed comparison of existing methods using heart disease dataset. (*Continued*)

References	Year	Classification used	Evaluation metrics
Kishore *et al.* [23]	2018	Adaptive neuro-fuzzy	Accuracy: 94.1%
Latha *et al.* [24]	2019	Ensemble classifiers	Accuracy: 85.4%
Ahmed *et al.* [25]	2019	Decision tree	Accuracy:92.8%
Tuli *et al.* [26]	2019	Deep learning	Accuracy: 89%
Proposed method	**2021**	**Deep Neural network architecture**	**Accuracy: Training: 97%, Testing: 94%**

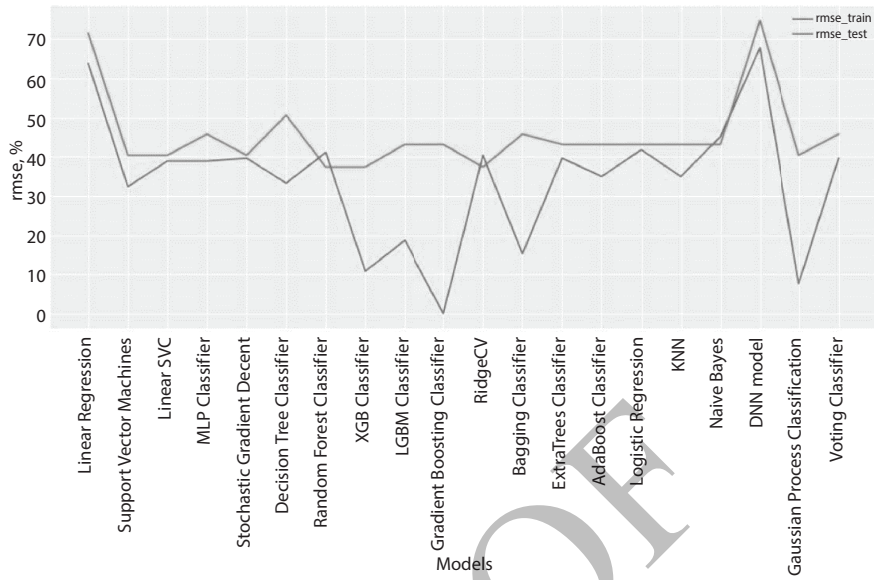


Figure 12.11 RMSE criteria for 20 models for train and test datasets.

for train and test datasets. Deep learning neural network shows the highest re value with maximum accuracy. Similarly, Figure 12.11 shows the plot of RMSE values for different machine learning methods and deep learning neural networks. Our paper investigates the various works on machine learning and deep learning techniques to predict the different types of cardiovascular diseases. The feature selection can improve the accuracy, and vital tasks, i.e., the relevant and irrelevant, can be identified for the proposed system. The neural network is a training method and can predict the relationship between input and output values. It used the backpropagation algorithm as support to predict the disease accurately.

12.7 Conclusion

According to the world health organization, nearly 32% of global deaths are heart disease. The crucial health problem in society is a heart attack, and this article has used machine learning methods and deep learning. This deep learning method predicts heart disease prediction and compares it with other strategies for different datasets. An efficient and accurate forecast of heart disease is needed. Various ML methods and neural

network-based architecture help find the feature selection and precise prediction. This book chapter uses the ensemble-based framework deep learning model and feature selection to improve accuracy. We have obtained an accuracy of 94% by using three layers of neural networks. This book chapter explains the different machine learning models and deep learning neural network models to predict heart disease with the UCI Cleveland dataset. Each machine learning model is explained briefly, and its outcome is compared with other models. We have obtained high accuracy of about 97% by using the deep learning neural network model, the performance parameters used. In this simulation environment are recall, RMSE, and confusion matrix to obtain the accuracy calculation.

References

Q9 1. Carroll, W. and Miller, G.E., STATISTICAL BRIEF# 409: Heart Disease among Elderly Americans: Estimates for the US Civilian Noninstitutionalized Population, 2010, 2013.

2. Kirubha, V. and Priya, S.M., Survey on data mining algorithms in disease prediction. *Int. J. Comput. Trend. Technol.*, *38*, 3, 124–128, 2016.

3. Chandra, P. and Deekshatulu, B.L., Prediction of risk score for heart disease using associative Classification and hybrid feature subset selection, in: *2012 12th International Conference on Intelligent Systems Design and Applications (ISDA)*, IEEE, pp. 628–634, 2012, November.

4. Sultana, M., Haider, A., Uddin, M.S., Analysis of data mining techniques for heart disease prediction, in: *2016 3rd international conference on electrical engineering and information communication technology (ICEEICT)*, IEEE, pp. 1– 5, 2016, September.

Q10 5. Deekshatulu, B.L. and Chandra, P., Classification of heart disease using k-nearest neighbor and genetic algorithm. *Proc. Technol.*, 85–94, 2013.

Q11 6. Kumra, S., Saxena, R., Mehta, S., An extensive review on swarm robotics. 140–145, 2009.

Q12 7. Lakshmi, T.M., Martin, A., Begum, R.M., Venkatesan, V.P., An analysis on performance of decision tree algorithms using student's qualitative data. *Int. J. Mod. Educ. Comput. Sci.*, 5, 5, 18–27, 2013.

8. Sharma, P. and Bhartiya, A.P.R., Implementation of decision tree algorithm to analysis the performance. *Int. J. Adv. Res. Comput. Commun. Eng.*, 1, 10, 861–864, 2012.

Q13 9. Srivastava, D.K. and Bhambhu, L., Data classification using support vector machine. *J. Theor. Appl. Inf. Technol.*, 2009.

10. Bhatia, N. and Author, C., Survey of nearest neighbor techniques. *IJCSIS) Int. J. Comput. Sci. Inf. Secur.*, 8, 2, 302–305, 2010.

11. Schmidhuber, J., Deep learning in neural networks: An overview, 2015.

12. Hochreiter, S. and Urgen Schmidhuber, J., Long short-term memory. *Neural Comput.*, 9, 8, 1735–1780, 1997.
Q14 13. Palaniappan, S. and Awang, R., Intelligent heart disease prediction system using data mining techniques. *2008 IEEE/ACS Int. Conf. Comput. Syst. Appl.*, 108–115, 2008.
14. Khemphila, A. and Boonjing, V., Heart disease classification using neural network and feature selection, in: *2011 21st International Conference on Systems Engineering*, IEEE, pp. 406–409, 2011, August.
15. Soni, J., Ansari, U., Sharma, D., Soni, S., Predictive data mining for medical diagnosis: An overview of heart disease prediction. *Int. J. Comput. Appl.*, *17*, 8, 43–48, 2011.
16. Ghwanmeh, S., Mohammad, A., Al-Ibrahim, A., Innovative artificial neural networks-based decision support system for heart diseases diagnosis, 2013.
17. Santhanam, T. and Ephzibah, E.P., Heart disease classification using PCA and feed forward neural networks, in: *Mining Intelligence and Knowledge Exploration*, pp. 90–99, Springer, Cham, 2013.
18. Ajam, N., Heart diseases diagnoses using artificial neural network. *Int. J. Complex Syst.*, *5*, 4, 2015.
19. Olaniyi, E.O., Oyedotun, O.K., Adnan, K., Heart diseases diagnosis using neural networks arbitration. *Int. J. Intell. Syst. Appl.*, *7*, 12, 72, 2015.
20. Roostaee, S. and Ghaffary, H.R., Diagnosis of heart disease based on meta heuristic algorithms and clustering methods. *J. Electr. Comput. Eng. Innov. (JECEI)*, *4*, 2, 105–110, 2016.
21. Karthikeyan, . and Kanimozhi, V., Deep learning approach for prediction of heart disease using data mining classification algorithm deep belief approach. *Int. J. Adv. Res. Sci., Eng. Technol.*, 4, 3194–3201, 2017.
22. Priscila, S.S. and Hemalatha, M., Improving the performance of entropy ensembles of neural networks (EENNS) on Classification of heart disease prediction. *Int. J. Pure Appl. Math.*, *117*, 7, 371–386, 2017.
23. Nandhu Kishore, A.H. and Jayanthi, V.E., Neuro-fuzzy based medical decision support system for coronary artery disease diagnosis and risk level prediction. *J. Comput. Theor. Nanosci.*, 15, 1027–1037, 2018.
24. Latha, C.B.C. and Jeeva, S.C., Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Inform. Med. Unlocked*, 16, 100203, 2019.
25. Ahmed, H., Younis, E.M.G., Hendawi, A., Ali, A.A., Heart disease identification from patients' social posts, machine learning solution on Spark. *Futur. Gener. Comput. Syst.*, 2019.
26. Tuli, S., Basumatary, N., Gill, S.S., Kahani, M., Arya, R.C., Wander, G.S., Buyya, R., HealthFog: An ensemble deep learning based smart healthcare system for automatic diagnosis of heart diseases in integrated IoT and fog computing environments, Futur. *Generation Comput. Syst.*, 2019.

27. Rani, P., Kumar, R., Jain, A., Multistage model for accurate prediction of missing values using imputation methods in heart disease dataset, in: *Innovative Data Communication Technologies and Application*, pp. 637–653, Springer, Singapore, 2021.
28. Doppala, B.P., Bhattacharyya, D., Chakkravarthy, M., Kim, T.-H., A hybrid machine learning approach to identify coronary diseases using feature selection mechanism on heart disease dataset. *Distrib. Parallel Database*, 20211-20.
29. Bharti, R., Khamparia, A., Shabaz, M., Dhiman, G., Pande, S., Singh, P., Prediction of heart disease using a combination of machine learning and deep learning. *Comput. Intell. Neurosci.*, 2021.
30. Jebakumar, A.Z. and Ravanan, R., A novel machine learning approaches for heart disease dataset. *Elementary Educ. Online*, 20, 5, 7391–7400, 2021.
31. Selvi, R. and Muthulakshmi, I., An optimal artificial neural network based big data application for heart disease diagnosis and classification model. *J. Ambient Intell. Humaniz. Comput.*, 12, 6, 6129–6139, 2021.
32. Rani, P., Kumar, R., Sid Ahmed, N.M., Jain, A., A decision support system for heart disease prediction based upon machine learning. *J. Reliab. Intell. Environ.*, 1–13, 2021.
33. Kavitha, M., Gnaneswar, G., Dinesh, R., Rohith Sai, Y., Sai Suraj, R., Heart disease prediction using hybrid machine learning model, in: *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, IEEE, pp. 1329–1333, 2021.
34. Pavithra, M., Sindhana, A.M., Subajanaki, T., Mahalakshmi, S., Effective heart disease prediction systems using data mining techniques. *Ann. Romanian Soc Cell Biol.*, 6566–6571, 2021.
35. Ghosh, P., Azam, S., Jonkman, M., Karim, A., Javed Mehedi Shamrat, F.M., Ignatious, E., Shultana, S., Beeravolu, A.R., De Boer, F., Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques. *IEEE Access*, 9, 19304–19326, 2021.
36. Hossain, M.E., Uddin, S., Khan, A., Network analytics and machine learning for predictive risk modelling of cardiovascular disease in patients with type 2 diabetes. *Expert Syst. Appl.*, 164, 113918, 2021.
37. Jiang, H., Mao, H., Lu, H., Lin, P., Garry, W., Lu, H., Yang, G., Chen, X., Machine learning-based models to support decision-making in emergency department triage for patients with suspected cardiovascular disease. *Int. J. Med. Inform.*, 145, 104326, 2021.
38. Pollard, J.D., Haq, K.T., Lutz, K.J., Rogovoy, N.M., Paternostro, K.A., Soliman, E.Z., Maher, J., Lima, J.A.C., Musani, S.K., Tereshchenko, L.G., Electrocardiogram machine learning for detection of cardiovascular disease in African Americans: the Jackson Heart Study. *Eur. Heart J. Digit. Health*, 2, 1, 137–151, 2021.

39. Jiang, Y., Zhang, X., Ma, R., Wang, X., Liu, J., Keerman, M., Yan, Y. *et al.*, Cardiovascular disease prediction by machine learning algorithms based on cytokines in Kazakhs of China. *Clin. Epidemiol.*, 13, 417, 2021.
40. Chu, H., Chen, L., Yang, X., Qiu, X., Qiao, Z., Song, X., Zhao, E. *et al.*, Roles of anxiety and depression in predicting cardiovascular disease among patients with type 2 diabetes mellitus: A machine learning approach. *Front. Psychol.*, 12, 2021.
41. Rubini, P.E., Subasini, C.A., Vanitha Katharine, A., Kumaresan, V., Kumar, S.G., Nithya, T.M., A cardiovascular disease prediction using machine learning algorithms. *Ann. Romanian Soc. Cell Biol.*, 904–912, 2021.
42. Sajeev, S., Champion, S., Beleigoli, A., Chew, D., Reed, R.L., Magliano, D.J., Shaw, J.E. *et al.*, Predicting Australian adults at high risk of cardiovascular disease mortality using standard risk factors and machine learning. *Int. J. Environ. Res. Public Health*, 18, 6, 3187, 2021.
43. Kim, M., Kim, Y.J., Park, S.J., Kim, K.G., Oh, P.C., Kim, Y.S., Kim, E.Y., Machine learning models to identify low adherence to influenza vaccination among Korean adults with cardiovascular disease. *BMC Cardiovasc. Disord.*, 21, 1, 1–8, 2021.
44. Angelaki, E., Maria, E., Marketou, G.D., Barmparis, A.P., Vardas, P.E., Parthenakis, F., Tsironis, G.P., Detection of abnormal left ventricular geometry in patients without cardiovascular disease through machine learning: An ECG-based approach. *J. Clin. Hypertens.*, 235, 935–945, 2021.
45. Al-Absi, H.R.H., Refaee, M.A., Rehman, A.U., Islam, M.T., Belhaouari, S.B., Alam, T., Risk factors and comorbidities associated to cardiovascular disease in Qatar: A machine learning based case-control study. *IEEE Access*, 9, 29929–29941, 2021.
46. Allan, S., Olaiya, R., Burhan, R., Reviewing the use and quality of machine learning in developing clinical prediction models for cardiovascular disease. *Postgrad. Med. J.*, 2021.
47. Kim, J.O.R., Jeong, Y.-S., Kim, J.H., Lee, J.-W., Kim, H.-S., Machine learning-based cardiovascular disease prediction model: A cohort study on the Korean national health insurance service health screening database. *Diagnostics*, 11, 6, 943, 2021.
48. Uddin, M.N. and Halder, R.K., An ensemble method based multilayer dynamic system to predict cardiovascular disease using machine learning approach. *Inform. Med.Unlocked*, 100584, 2021.
49. Joshi, A., Rienks, M., Theofilatos, K., Mayr, M., Systems biology in cardiovascular disease: A multiomics approach. *Nat. Rev. Cardiol.*, 18, 5, 313–330, 2021.
50. Sun, W., Zhang, P., Wang, Z., Li, D., Prediction of cardiovascular diseases based on machine learning. *ASP Trans. Internet Things*, 1, 1, 30–35, 2021.
51. Garg, H., Machine Learning techniques for cardiovascular disease, in: *IOP Conference Series: Materials Science and Engineering*, vol. 1116, IOP Publishing, p. 012140, 2021.

Q15 52. Telagam, N. and Kandasamy, N., Review of the medical Internet of Things-based RFID security protocols, in: *Nanoelectronic Devices for Hardware and Software Security*, pp. 163–178, CRC Press, 2021.
53. Nagarjuna, T., Overview of THz applications, in: *Advanced Indium Arsenide-Based HEMT Architectures for Terahertz Applications*, vol. 45, 2021.
54. Telagam, N., Ajitha, D., Kandasamy, N., Review on hardware attacks and security challenges in IoT edge nodes, in: *Security of Internet of Things Nodes: Challenges, Attacks, and Countermeasures*, p. 211, 2021.
55. Dioline, S., Arunkumar, M., Dinesh, V., Nagarjuna, T., Karuppanan, S., Radiology: Clinical trials implemented by composite test-beds via MVDR beamformer system. *Materials Today: Proceedings*, 2021.
56. Gantala, A., Telagam, N., Kumar, G.V., Anjaneyulu, P., Prasad, R.M., Content-based image retrieval using genetic algorithm retrieval effectiveness in terms of precision and recall. *J. Adv. Res. Dyn. Control Syst.*, 9, 18, 2020–2028, 2017.

Answer all Queries (Q) in the margin of the text. When requested, provide missing intext reference citation as well as intext reference for figure/table. Please annotate the PDF and read the whole chapter again carefully. Careful proofing is key to optimal output.

Author Queries

Q1 Chapter title do not fit within text area. Please provide short running title with 45 characters (including spaces) per Scrivener guideline.

Q2 Please note that citation of Table 12.2 was mentioned first before Table 12.1. Citation should be in sequence. Please advise how to proceed.

Q3 Please check if the Equations 12.1 to 12.6 was captured correctly.

Q4 Please provide missing call-out/citation for Figures 12.2 and 12.3 in the main text. Note that figure citations should be cited sequentially.

Q5 Please note that citation of Figure 12.6 was mentioned first before Figure 12.5. Citation should be in sequence. Please advise how to proceed.

Q6 Please provide data for section 12.4.2 or kindly advise how to proceed.

Q7 Please note that there are two Table 12.4 should the second Table 12.4 be renumber also the corresponding citation will be renumbered.

Q8 Please provide missing call-out/citation for Table 12.5 in the main text. Note that figure citations should be cited sequentially.

Q9 Please provide journal title, volume and page number of references 1, 11, 16.

Q10 Please provide volume number of references 5, 32, 34, 41, 48.

Q11 Please provide journal title and volume number of reference 6.

Q12 Please provide missing call-out/citation for references 7 and 11 in the main text.

Q13 Please provide volume and page number of references 9, 25, 26, 28, 29, 46.

Q14 Please provide page number of references 13, 18, 40

Q15 Please provide publisher location of reference 52.