©IDOSR PUBLICATIONS
International Digital Organization for Scientific Research
IDOSR JOURNAL OF COMPUTER AND APPLIED SCIENCES 6(1):66-73, 2021. ISSN: 2579-0803

Heart Disease Prediction Using Machine Learning and Data Mining Techniques: Application of Framingham Dataset

Walaa Adel Mahmoud[1], Mohamed Aborizka[1] and Fathy Amer[2]

[1]Information System Department, College of Computing & Information Technology, Arab Academy for Science, Technology and Maritime Transport, Cairo, Egypt
[2]Department of Computing & Information Technology, College of Computing, Oct.6 University,Cairo, Egypt
Email:walaaadel720@hotmail.com; m.aborizka@aast.edu.eg; dr_fathi_amer@yahoo.com

ABSTRACT

Heart diseases have an abundant pact of attention in medical research due to its impact on human health where early diagnosis is critical to delay the development of cardiovascular disease, the world's leading cause of death. thus, it is much needed to predict the possibility of occurrence of cardiovascular disease based on their characteristics. This paper studies the different machine learning classification algorithms to predict the cardiovascular disease. The Ten-fold cross-validation resampling is used to validate the prediction model. Aim and the prediction scores of each algorithm are evaluated with performance metrics such as prediction accuracy, confusion matrix, F1-meuser and suggested geometric mean. It was shown that using different classification algorithms for the classification of the HD dataset gives very promising results in term of the classification accuracy for the KNN, SVM, DT, LR and RFalgorithms with accuracy of classification of 83.95, 84.5, 84.82,84.89 and 85.05 % respectively.The RF algorithm predicts 84.8 % (true positive rate) of the deceased cases correctly. Based on the prediction results of various machine learning classification algorithms on the Framinghamdataset, this paper shows that the RF algorithm predicts Framinghampossibilities well for the smaller (4240 records) dataset than other algorithms.

Keywords:Framinghamdataset, Classification, Recall, Precision, Accuracy, F1-meuser, Geometric Mean, Confusion Matrix, Supervised, Unsupervised and Reinforcement learning, Outlier data, Missing values.

INTRODUCTION

Data mining, is known as the extraction of implicit, previously unknown, and potentially useful information from data. It encompasses a set of processes performed automatically, whose task is to discover and extract hidden features from large datasets[1]. Machine Learning (ML) is three type, supervised machine learning, unsupervised machine learning and reinforcement learning [2]. ML is a wide field which depend on concepts from computer science, statistics, cognitive science, engineering, optimization theory and many other disciplines of mathematics and science. An overall work flow of our study has been shown in Fig. 1. Classification is one of the most studied problems in machine learning and data mining. Predicting the outcome of a disease is one of the most interesting and challenging tasks in which to develop data mining applications.There are many algorithms in Machine Learning. Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbors (KNN), Decision Tree (DT), and Logistic Regression (LR) used for data classification. Data Cleaning and Preparation. The process of systematically and accurately cleaning the data to make it ready for analysis is known as data cleaning. Most of times, there will be discrepancies in the gathered data like wrong data formats, missing data values, errors while collecting the data. For training and testing of algorithm, the data need to be split in two parts. The training set contains well known classified output and the model trains on this data to be generalized to other data later [3].

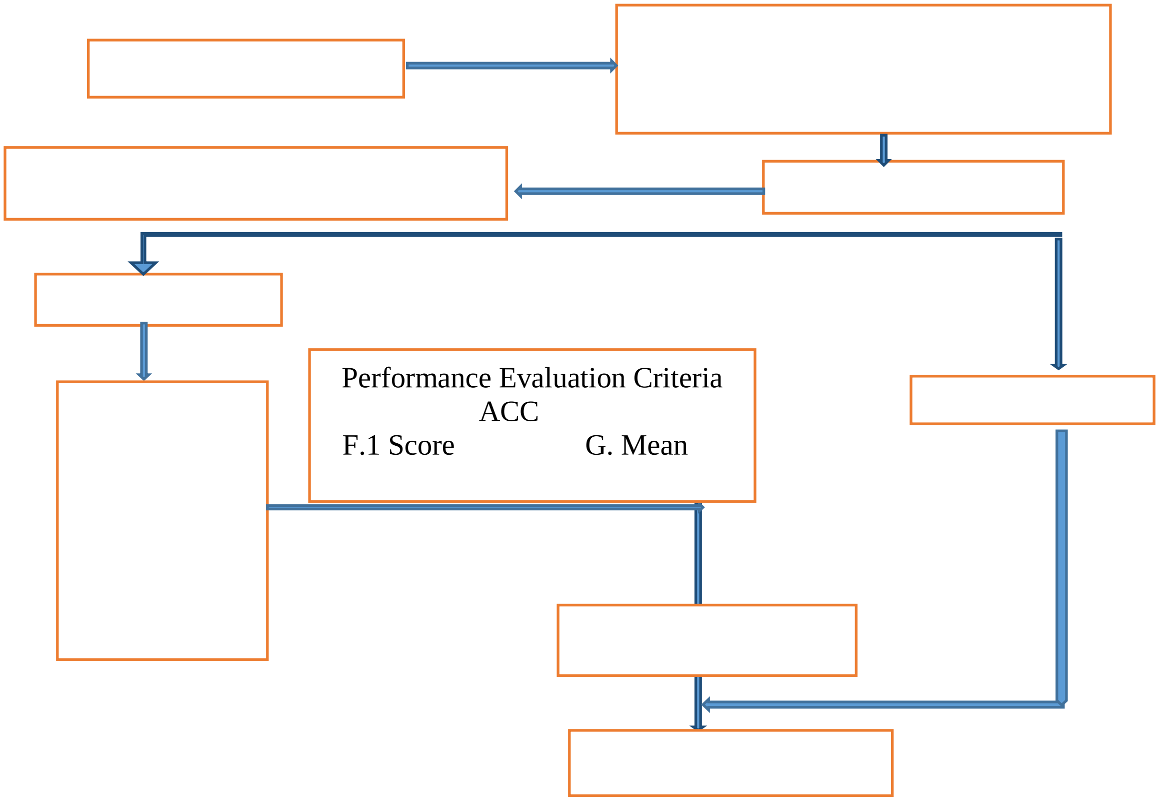

Fig.1: Methodology of the research work

The organized of the paper is organized as follows. In the next section, introduces the different machine learning Algorithms. While in Section 3, introduces the dataset description.In Section 4, we illustrate the performance metrics. A real data set is analyzed in Section 5 and concluding remarks are included in Section 6.

Algorithms

The various classification machine learning algorithms used for the prediction of Framinghamdataset are reviewed below: Logistic Regression is a statistical model that is used as a binary classifier, which classifies each sample into two classes (Yes/No). It is used for predicting the categorical dependent variable using a given set of independent variables [4]. Decision Tree is a nonparametric algorithm and is considered a classical machine learning algorithm. It performs well in situations where there is a single attribute that can easily split the data and helps in decision making [5]. Random Forest is an ensemble algorithm based on DT algorithm and works well with large datasets with high dimensionality. RF runs often on large datasets and it is slow in operation compare to other algorithms. This technique can be used for both regression and classification tasks but generally performs better in classification tasks. This technique is based on the belief that more number of trees would converge to the right decision. For regression, it takes the mean of all the outputs of each of the decision trees whereas in classification it uses a voting system and then decides the class [6]. K-Nearest Neighboris a lazy supervised machine learning algorithm that used to predict and classify and it is easy to implement and understand, requires short training time and whole training set is used for prediction. this algorithm is a nonparametric and used to predict and classify unknown data from known data by measuring the distance between them. The distance metric is using to measure the distance between point from testing data with all the point in training data [7]. Support vector machine is succeeded supervised learning algorithm for classification problems. SVM is a nonparametric algorithm aimed to find the optimal hyper-planes that separate classes on a training dataset.Generally, the main idea of SVM comes from binary classification, namely to find a hyperplane as a segmentation of the two classes to minimize the classification error [8].

Dataset

The Framingham dataset included in this research work has 16 columns with 15 independent variables and one dependent target variable. It has 4240 rows. The variable's description is given in table 1.

Table.1:Selected Framingham dataset attributes.

Attribute & Description		Number of outlier and missing data	
		Missing	Outlier
Age	(32-70)	0	0
Sex	0=Female, 1=Male	0	0
Education	It takes values as: 1=High School, 2=High School or GED, 3=College or Vocational School, 4=College	105	0
Current Smoker	0=No 1=Yes	0	0
Cigs Per Day	Number of Cigarettes smoked Per Day (0-70)	29	0
BP Meds	0=No 1=Yes	53	0
Prevalent Stroke	0=No 1=Yes	0	0
Prevalent Hyp	0=No 1=Yes	0	0
Diabetes	0=No 1=Yes	0	0
Tot Chol	Serum Cholesterol (107-696) (mg/dl)	50	196
Sys BP	(83.5-295) (mm/hg)	0	302
Día BP	(48-142.5) (mm/hg)	0	248
Body Mass Index (BMI)	(15.54-56.8)	19	240
Heart Rate	Heart Rate achieved (44-143)	1	271
Glucose	(40-394) (mg/dl)	388	522
10-year CHD	0=Healthy, 1=Diseases	0	0

Performance Evaluation Criteria

Typically, the performance of the ML prediction algorithms measured by using some metrics based on the classification algorithm. In this work, the prediction results are evaluated by using the metrics such as confusion matrix, accuracy,F-measure geometric mean.

Confusion Matrix

In ML, the algorithmsare basically assessingby a confusion matrix. For a binary class problem, a matrix is a square of two by two as shown in table 2; column represents the algorithms prediction;while the row is the real value of class label where, true positive (TP) is number of positive samples correctly predicted. False negative (FN) is number of positive samples wrongly predicted. False positive (FP) is number of negative samples wrongly predicted as positive. True negative (TN) is number of negative samples correctly predicted [8].

Table (2): Confusion Matrix

Actual	Predicted	
	Positive	Negative
Positive	TP	FP
Negative	FN	TN

Accuracy:Accuracy, the most popular metric for classifier evaluation, it assesses the overall effectiveness of the algorithm by estimating the probability of the true value of the class label [4]. The equation of accuracy as follows:

$$Accuracy = (TP + TN) / (TP + FP + TN + FN).$$

F-measure is defined as the harmonic mean of precision and recall (Saleh et al., 2020).And computed with a next equation:

$$F1\ Score = 2 \times (Precision \times Recall) / (Precision + Recall).$$

Where,

$$Precision = TP/(TP+FP)$$

and

$$Recall = TP/ (TP + FN)$$

Geometric Mean

$$GM = \sqrt[2]{Accuracy \times F1\ score}$$

Application

R programming is used for implementing the classification techniques.The data preprocessing and cleaning process (data imputation-mean technique) handling the missing and outlier data values from the dataset. Outlier is considered as noise in the data and affects the accuracy of the algorithms. We have used Boxplot to find out and treat outliers. Fig. 2 shows the boxplot of 6 attribute namely; age, totChol,sysBP, diaBP, BMI, heartRate and glucose. Missing values in the data can exist because of various reasons including the faults of the measuring instrument, human error or inconsistent measuring unit etc. Before training the model, missing values should be handled as it affects the accuracy of the learning algorithm. [9]. Thus, the first step used to clean the data is finding incomplete data or the null values and dealing with them to improve the performance of model. The missing values found are 645 out of 4240 rows which are approximately 12% of actual data as showed fig.3

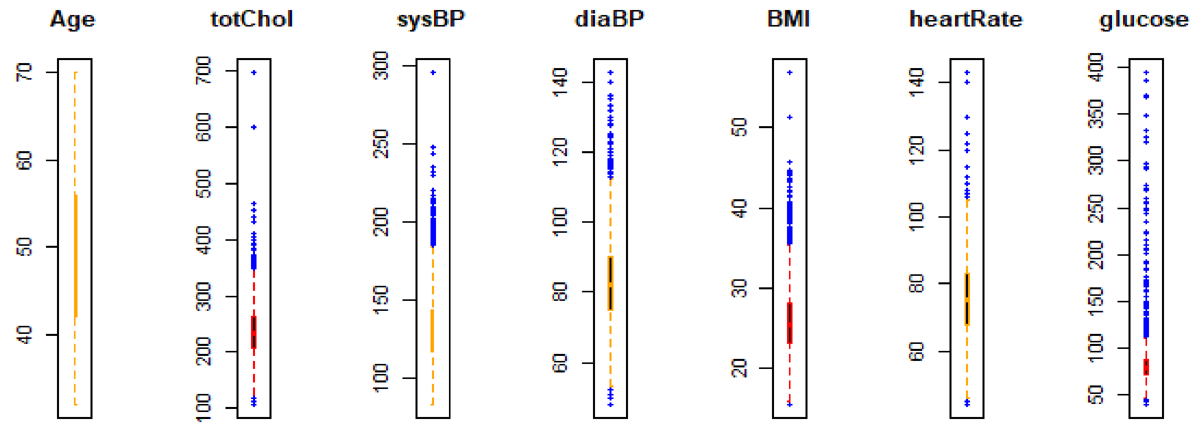

Fig. 2: Boxplots of data set

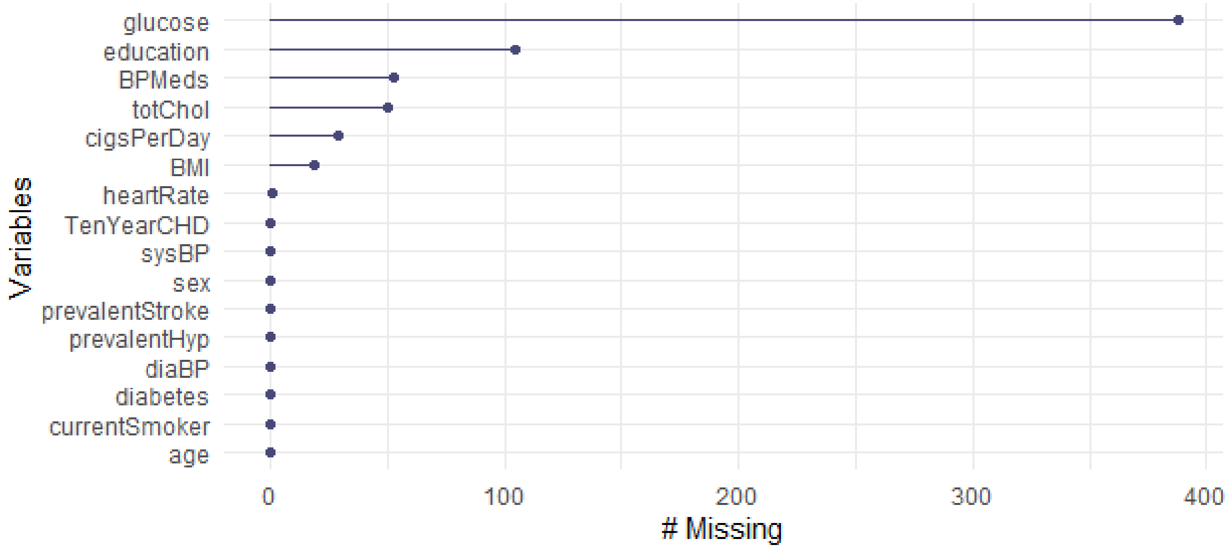

Fig.3: Missing value.

Table 1 shows the attributes along with the missing values against each attribute. As proposed in our framework, all the missing values and outlier of specific attribute (in the dataset) are replaced by the median of all the values of corresponding attribute. The reason for using the median substitution is that it increases the samples in our data without adding further information. In this manner, it contributes towards making more informed prediction / decision.Fig.4 and Fig.5 show that no outlier and missing values are present in the dataset.

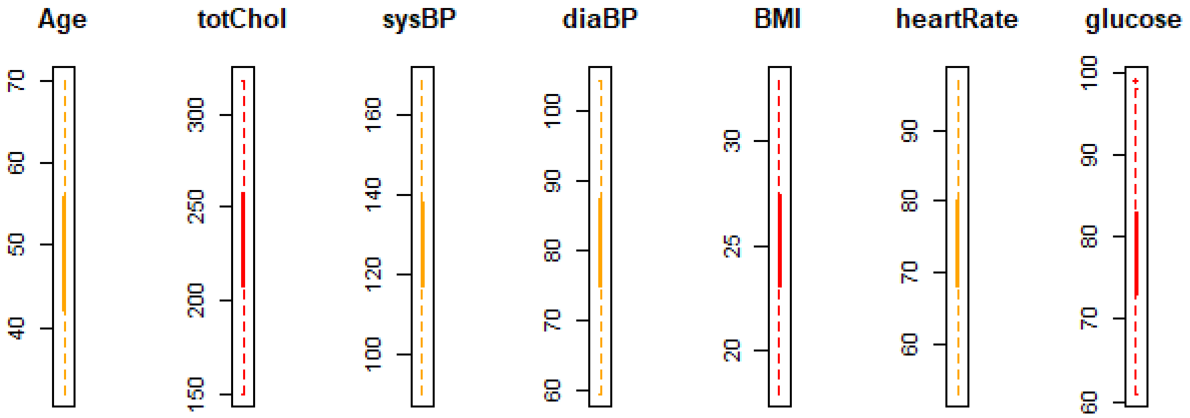

Fig.4: Handling Outlier data

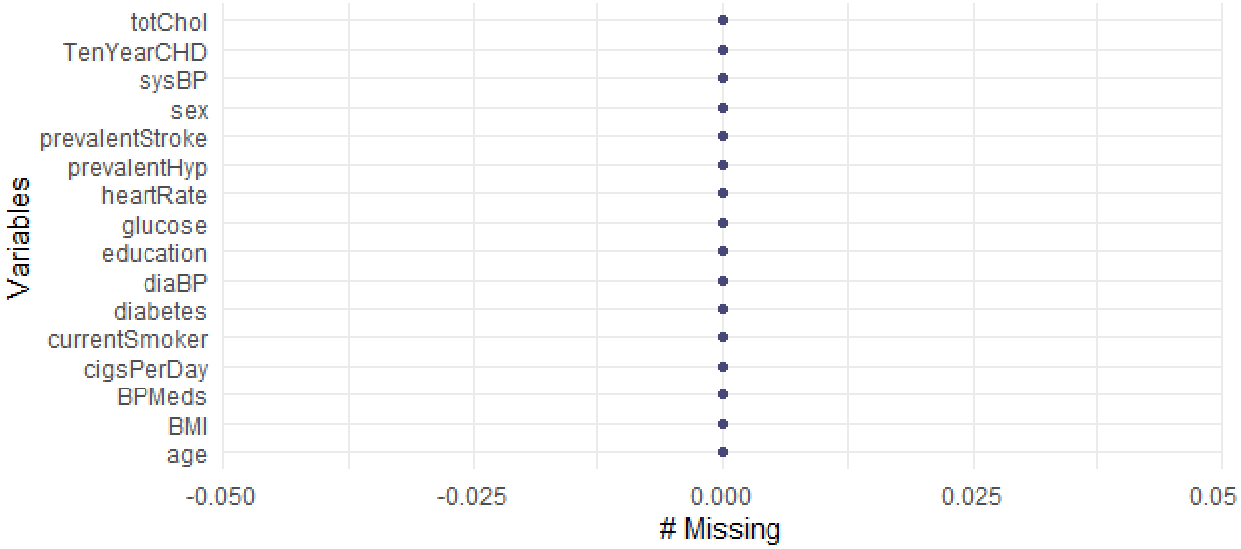

Fig.5: Handling Missing Values.

After imputation to build a classification model, the combined dataset with 16 attributes is divided into training and testing data with a percentage split of 70-30%.In thiscase, data is split below into two subsets: training (70%) and testing (30%). The confusion matrix obtained by five different supervised machine learning algorithms is given below. The performance measures are in accordance with the accuracy of each classification algorithm. Ten-fold cross validation was utilized to evaluate the performance of the classification models.In this approach, the entire dataset is divided into ten subsets and processed ten times where, nine subsets are used as testing sets and the remaining subset is used as training. Finally, the results are obtained by averaging each ten iterations.

Table (3): Split data to 70-30% with set. seed (5020).

	Algorithms	Metric		
		ACC	F1	GM
Performances	SVM	0.845	0.9159	0.8797
	DT	0.8482	0.9178	0.8823
	KNN	0.8395	0.9119	0.8749
	LR	0.8489	0.9180	0.8827
	RF	0.8505	0.9190	0.8840

Table 3 compares the classification metrics of algorithms.In table (3), the dataset is classified, the accuracy rates of SVM, DT, KNN, LR and RF are found in the range of 83.95% - 85.05.

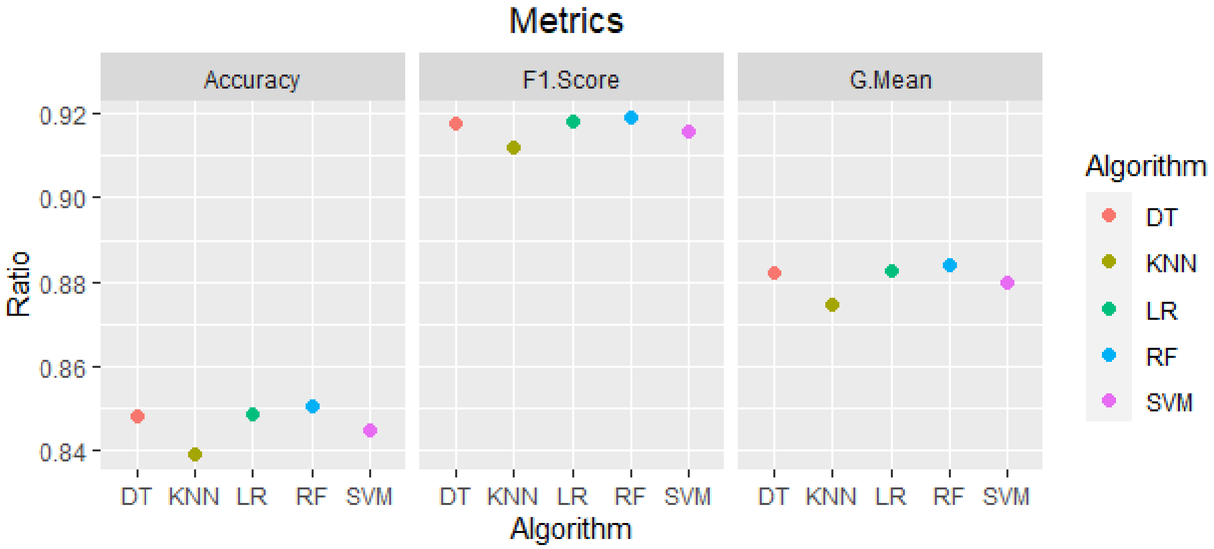

Fig.6:Evaluation of Performance.

The performance parameter measurement in table 4 gives a very promising result by RF algorithm in our dataset. This algorithm reaches 91.90% for F-measure, 88.40% for geometric mean and 85.05% for accuracy. Fig.6 also confirms this conclusion, while the other algorithms have lessdifferent metrics.

CONCLUSION

In this paper, it can be concluded that there is a huge scope for machine learning algorithms in predicting cardiovascular diseases or heart related disease. In this paper various popular machine learning algorithms has been discussed with their basic working mechanism where are applied to various real world datasets (in this case Framingham dataset) and study is carried out to find out the classifier which can perform well on the real world data sets.Finally, several binary classification algorithms that are very useful in detecting cardiovascular diseases or heart related disease are analyzed. RF has proven to be the best classification algorithm to classify the risk of heart disease with 85.05% accuracy.

REFERENCES

1. Ha, D. T., Loan, P. T. T., Giap, C. N., & Huong, N. T. L. (2020). An Empirical Study for Student Academic Performance Prediction Using Machine Learning Techniques. International Journal of Computer Science and Information Security (IJCSIS), 18(3).
2. Kumar, A., Sushil, R., and Tiwari, A. K. (2019). Comparative study of classification techniques for breast cancer diagnosis. International Journal of Computer Sciences and Engineering, 7(1), 234-240.
3. Kumar, G. R., Ramachandra, G. A., and Nagamani, K. (2013). An efficient prediction of breast cancer data using data mining techniques. International Journal of Innovations in Engineering and Technology (IJIET), 2(4), 139.
4. Rahim, A., Rasheed, Y., Azam, F., Anwar, M. W., Rahim, M. A., and Muzaffar, A. W. (2021). An Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases. IEEE Access, 9, 106575-106588.
5. Reddy, N. S. C., Nee, S. S., Min, L. Z., and Ying, C. X. (2019). Classification and feature selection approaches by machine learning techniques: Heart disease prediction. International Journal of Innovative Computing, 9(1).
6. Swain, D., Ballal, P., Dolase, V., Dash, B., andSanthappan, J. (2020). An Efficient Heart Disease

Prediction System Using Machine Learning. In Machine Learning and Information Processing (pp. 39-50). Springer, Singapore.

7. Bekkar, M., Djemaa, H. K., andAlitouche, T. A. (2013). Evaluation measures for models assessment over imbalanced data sets. J Inf Eng Appl, 3(10).
8. Jiao, Y., and Du, P. (2016). Performance measures in evaluating machine learning based bioinformatics predictors for classifications. Quantitative Biology, 4(4), 320-330.
9. Saleh, B., Saedi, A., Al-Aqbi, A., and Salman, L. (2020). Analysis of Weka Data Mining Techniques for Heart Disease Prediction System. International Journal of Medical Reviews, 7(1), 15-24.
10. Singh, D., andSamagh, J. S. (2020). A comprehensive review of heart disease prediction using machine learning. *Journal of Critical Reviews*, 7(12), 281-285.