DOI: 10.37990/medr.1011924 Med Records 2022;4(1):1-6

MEDICAL RECORDS-International Medical Journal

Araştırma Makalesi / Research Article

Early Detection of Coronary Heart Disease Based on Machine Learning Methods

Makine Öğrenme Yöntemlerine Dayali Kroner Kalp Hastaliğinin Erken Tespiti

Rüstem Yilmaz[1], Fatma Hilal Yağın[2]

[1]Samsun Gazi State Hospital, Department of Cardiology, Ilkadim, Samsun, Turkey
[2]Inonu University Department of Biostatistics and Medical Informatics, Faculty of Medicine, Malatya, Turkey

Abstract

Aim: Heart disease detection using machine learning methods has been an outstanding research topic as heart diseases continue to be a burden on healthcare systems around the world. Therefore, in this study, the performances of machine learning methods for predictive classification of coronary heart disease were compared.

Material and Method: In the study, three different models were created with Random Forest (RF), Logistic Regression (LR), and Support Vector Machine (SVM) algorithms for the classification of coronary heart disease. For hyper parameter optimization, 3-repeats 10-fold repeated cross validation method was used. The performance of the models was evaluated based on Accuracy, F1 Score, Specificity, Sensitivity, Positive Predictive Value, Negative Predictive Value, and Confusion Matrix (Classification matrix).

Results: RF 0.929, SVM 0.897 and LR 0.861 classified coronary heart disease with accuracy. Specificity, Sensitivity, F1-score, Negative predictive and Positive predictive values of the RF model were calculated as 0.929, 0.928, 0.928, 0.929 and 0.928, respectively. The Sensitivity value of the SVM model was higher compared to the RF.

Conclusion: Considering the accurate classification rates of Coronary Heart disease, the RF model outperformed the SVM and LR models. Also, the RF model had the highest sensitivity value. We think that this result, which has a high sensitivity criterion in order to minimize overlooked heart patients, is clinically very important.

Keywords: Heart disease, machine learning, classification, random forest, parameter optimization

Öz

Amaç: Kalp hastalıklarının dünya çapında sağlık sistemleri üzerinde bir yük olmaya devam etmesi nedeniyle, makine öğrenme yöntemlerini kullanarak kalp hastalığı tespiti olağanüstü bir araştırma konusu olmuştur. Bu nedenle, bu çalışmada, koroner kalp hastalığının tahmin edici sınıflandırması için makine öğrenme yöntemlerinin performansları karşılaştırılmıştır.

Materyal ve Metot: Çalışmada koroner kalp hastalığının sınıflandırılması için Rasgele Orman (RF), Lojistik Regresyon (LR) ve Destek Vektör Makinesi (SVM) algoritmaları ile üç farklı model oluşturulmuştur. Hiperparametre optimizasyonu için 3 tekrarlı 10 katlı tekrarlı çapraz doğrulama yöntemi kullanıldı. Modellerin performansı Doğruluk, F1 Skoru, Seçicilik, Duyarlılık, Pozitif Tahmin Değeri, Negatif Tahmin Değeri ve Karışıklık Matrisi (Sınıflandırma matrisi) temel alınarak değerlendirilmiştir.

Bulgular: Koroner kalp hastalığını RF 0.929, SVM 0.897 ve LR 0.861 doğrulukla sınıflandırdı. RF modelinin seçicilik, duyarlılık, F1-skor, negatif tahmin ve pozitif tahmin değerleri sırasıyla 0.929, 0.928, 0.928, 0.929 ve 0.928 olarak hesaplanmıştır. Ek olarak SVM modelinin duyarlılık değeri RF'ye göre daha yüksek çıkmıştır.

Sonuç: Koroner Kalp hastalığının doğru sınıflandırma oranları göz önüne alındığında, RF modeli SVM ve LR modellerinden daha iyi performans göstermiştir. Ayrıca RF modeli en yüksek duyarlılık değerine sahipti. Gözden kaçırılan kalp hastalarını en aza indirmek için yüksek bir duyarlılık kriterine sahip olan bu sonucun klinik açıdan oldukça önemli olduğunu düşünmekteyiz.

Anahtar Kelimeler : Kalp hastalığı, makine öğrenmesi, sınıflandırma, rastgele orman, parametre optimizasyonu

INTRODUCTION

Coronary heart disease (CHD) is the world's leading cause of death. CHD is often referred to as ischemic heart disease or coronary artery disease. Coronary heart disease arises when fatty deposits in the coronary arteries impede or disrupt blood flow to the heart. Over time, the walls of coronary arteries may become furrowed with fatty deposits. Atheroma is the term for fatty deposits, and

Geliş Tarihi / Received: 19.10.2021 **Kabul Tarihi / Accepted:** 14.11.2021
Sorumlu Yazar /Corresponding Author: Fatma Hilal Yağın, Inonu University Department of Biostatistics and Medical Informatics, Faculty of Medicine, Malatya, Turkey, E-mail: hilal.yagin@inonu.edu.tr

atherosclerosis is the term for the process. Smoking and ingesting excessive amounts of alcohol on a regular basis are two lifestyle factors that induce atherosclerosis. CHD is caused by excessive cholesterol, high blood pressure (hypertension), and diabetes (1, 2).

The most frequent symptoms of coronary heart disease are chest pain (angina) and shortness of breath. Patients' medical/family history and risk factors are requested if a doctor believes you are at risk of coronary. Although coronary heart disease cannot be cured, medicines can help manage the symptoms and extend life by reducing the risk of complications such as heart attacks and heart failure. As a result, early detection and treatment of the disease are critical for lowering the mortality rate (3-5).

Artificial intelligence (AI) are used for the diagnosis of many diseases such as heart, diabetes and cancer, and thus is becoming more and more popular in healthcare. AI is a broad term that encompasses analytical algorithms that iteratively learn from data, allowing computers to discover hidden insights without being explicitly instructed where to seek. These are a group of operations that include terminology like machine learning, cognitive learning, deep learning, and reinforcement learning-based methods for integrating and interpreting complicated biological and healthcare data in situations where traditional statistical methods fail (6).

Machine learning (ML) is an area of AI that involves using mathematical models to assist a computer in learning without being given explicit instructions. Algorithms are used in machine learning to find patterns in data. These patterns are also utilized to build a data model that predicts the future. Machine learning algorithms are used in prospective clinical trials to compare existing standard of care procedures with the goal of introducing precision diagnostics, risk stratification, and personalized medicines.

Cardiovascular diseases are a group of disorders that can benefit tremendously from proactive care, prevention, and prediction, and hence AI approaches. Understanding the intricate individual risk factors, behavioral variables, and treatment pathways predictive of illness outcomes in specific patient cohorts, as well as establishing early therapeutic interventions, will require a variety of AI algorithms (7-14).

The aim of this study is to classify coronary artery disease with machine learning methods and to compare the classification performances of Logistic Regression, Random Forest, and Support Vector Machine methods.

MATERIAL AND METHOD

Dataset

The heart disease dataset used in this study was obtained from the IEEEDataPort database (https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive#files). The dataset was created by combining Cleveland, Hungarian, Switzerland, Statlog (Heart) Data Set, and Long Beach VA datasets. Combining was performed using 11 covariates from these 5 heart disease datasets. In this way, a rather large data set was obtained compared to the existing heart disease datasets. In the data set, 281 (23.6%) of the patients were female and 909 (76.4%) were male. The mean age of female was 53±10 and the mean age of male was 54±9. Detailed information about the data set is as in Table I and Table II (15).

Logistic Regression (LR)

Logistic Regression Analysis (LR) is a method used to determine the cause-effect relationship between the dependent variable and the independent variables, without being dependent on a certain distribution assumption, when the dependent variable is categorical and the independent variables are mixed-scale. Using the maximum likelihood estimation method, LR estimates the unknown parameter values that maximize the probability obtained from the data set. Thus, the parameter estimates that maximize the likelihood function are selected and the parameter estimates that best match the observed data are obtained (16, 17).

Table 1. Heart disease dataset attribute description

Attribute	Code given	Unit	Data type
Age	age	in years	Numeric
Sex	sex	1,0	Binary
Chest pain type	chest pain type	1, 2, 3, 4	Nominal
Resting blood pressure	resting bp s	in mm Hg	Numeric
Serum cholesterol	cholesterol	in mg/dl	Numeric
Fasting blood sugar	fasting blood sugar	1,0>120 mg/dl	Binary
Resting electrocardiogram results	resting ecg	0, 1, 2	Nominal
Maximum heart rate achieved	max heart rate	71-202	Numeric
Exercise induced angina	angina	0,1	Binary
Oldpeak=ST	oldpeak	depression	Numeric
The slope of the peak exercise ST segment	ST slope	0, 1, 2	Nominal
Class	target	0,1	Binary

Table 2. Description of nominal attributes in dataset	
Attribute	**Description**
Sex	1=male, 0=female
Chest Pain Type	Value1: typical angina Value 2: atypical angina Value 3: non-anginal pain Value 4: asymptomatic
Fasting blood sugar	fasting blood sugar >120 mg/dl (1=true;0=false) Value 0: normal
Resting electrocardiogram results	Value1: having ST-T wawe abnormality (T wawe inversions and/or ST elevation or depression of>0.05 Mv Value 2: showing probable or definite left ventricular hypertrophy by Estes criteria
Exercise induced angina	1=yes,0=no
The slope of the peak exercise ST segment	Value1: unsloping Value2: flat Value3: downsloping
Class	1=heart disease,0=Normal

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a set of supervised learning algorithms that detect patterns. It is a type of classification method that estimates the classification function and analyzes the data used for classification. When compared to other approaches, it often delivers better categorization results. It is a nonlinear classification method that has been reported. SVM's main concept is to use a hyperplane as a decision surface to optimize the margin of separation between positive and negative samples (Figure 1). This method converts non-linear input sample data into a high-dimensional space where the data may be separated linearly, resulting in improved classification (or regression) accuracy. SVMs are unique in that they have a strong theoretical base as well as cutting-edge success in real-world applications, especially in bioinformatics (18).

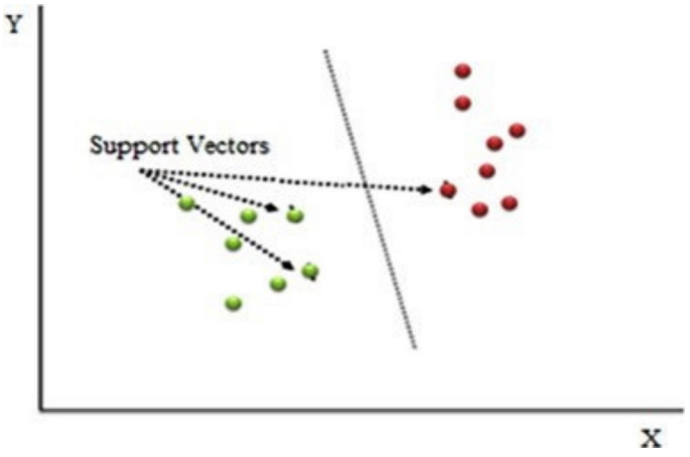

Figure 1. Support Vector distribution

Random Forest (RF)

Random Forest (RF) is community classification used for classification and regression analysis. In this community classifier, it is aimed to increase the classification success by creating more than one decision tree. RFs work by creating various decision trees and labels according to the majority during the training phase. The difference of RFs from decision tree algorithms is that basically finding the root node and splitting the nodes work randomly. The reason why the RF method is also considered in this study is that it is good at detecting noise and outliers and can solve the over-learning problem. It is also one of the most appropriate methods to define the most important feature among the data set features. Thus, feature extraction is applied in the most accurate way and the success rate is achieved to reach the highest rates. The classification logic of the random forest algorithm is as in Figure 2 (19, 20).

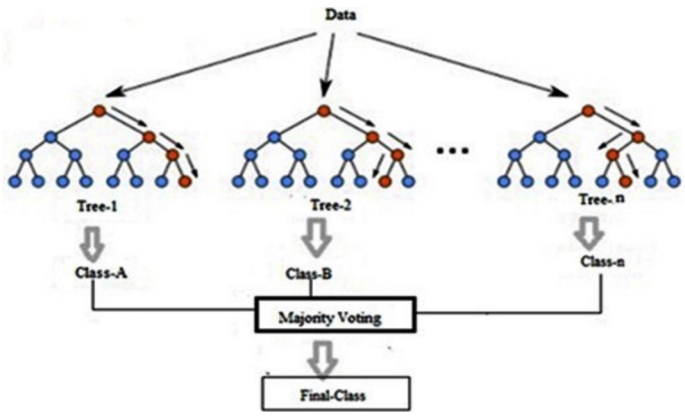

Figure 2. Random Forest algorithm

Data Preprocessing, Modeling and Evaluation of Predictive Models

The SVM-SMOTE method was first applied to the data set. After SVM-SMOTE, a total of 1258 (initially 1190) samples were obtained, 629 in each class. The dataset was then split with 80% for training and 20% for testing. LR, SVM and RF algorithms were used for the classification task. The optimal hyper-parameters of each model were determined by Grid Search with 3 repeats and 10-fold Repeated k-Fold Cross- Validation. The created models were evaluated with Accuracy, Specificity, Sensitivity, F1-score, Negative predictive value, and Positive predictive value.

RESULTS

In Table 3, hyper-parameters and their values determined by grid search for each model are given.

Figure 3, Figure 4, and Figure 5 show the confusion matrices for the LR, RF and SVM algorithms, respectively.

Table 3. Optimal Hyper- Parameters Determined By Grid Search

Algorithm	Parameter	Optimal Hyper- Parameters
RF	criterion	entropy
	max_depth	10
	min_samples_leaf	1
	min_samples_split	2
LR	C	1
	penalty	l2
	solver	newton-cg
SVM	C	10
	Gamma	1

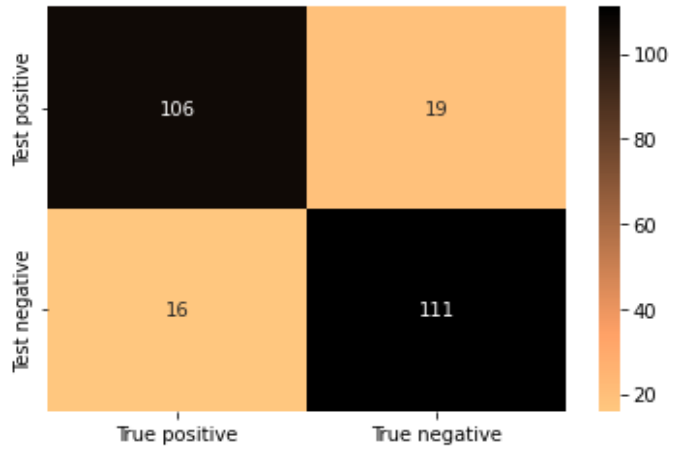

Figure 3. Confusion matrices for LR algorithm

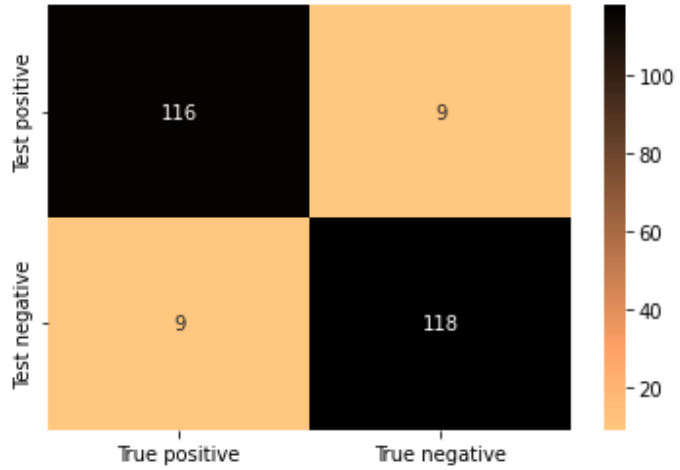

Figure 4. Confusion matrices for RF algorithm

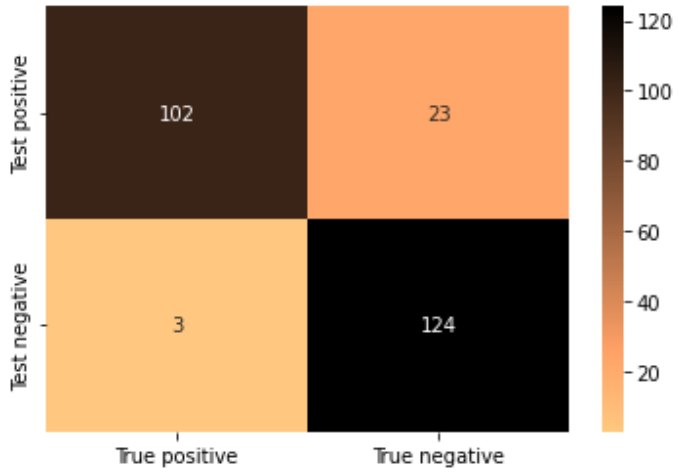

Figure 5. Confusion matrices for SVM algorithm

In Table 4, the performance criteria of the classification algorithms used in the study such as Accuracy, Specificity, Sensitivity, F1-score, Negative predictive value, and Positive predictive value are given.

- The values of Accuracy, Specificity, F1-score, Sensitivity, Negative predictive value, and Positive predictive value criteria obtained from the RF model were calculated as 0.929, 0.929, 0.928, 0.928, 0.929, and 0.928 respectively.
- The values of Accuracy, Specificity, F1-score, Sensitivity, Negative predictive value, and Positive predictive value criteria obtained from the SVM model were calculated as 0.897, 0.844, 0.887, 0.971, 0.976, and 0.816 respectively.
- In addition; the values of Accuracy, Specificity, F1-score, Sensitivity, Negative predictive value, and Positive

predictive value criteria obtained from the LR model were calculated as 0.861, 0.854, 0.858, 0.869, 0.874, and 0.848 respectively.

• As a result; The RF method offers the highest performance compared to SVM and LR.

Table 4. Performance Metrics Results for Classification Models

Score/Model	LR	RF	SVM
Accuracy	0.861	0.929	0.897
Specificity	0.854	0.929	0.844
Sensitivity	0.869	0.928	0.971
F1-score	0.858	0.928	0.887
Negative predictive value	0.874	0.929	0.976
Positive predictive value	0.848	0.928	0.816

DISCUSSION

Early detection of anomalies aids in the long-term saving of human life. The processing of raw healthcare data of heart disease lead to the discovery of this procedure. Machine learning algorithms can be used to process the raw data, resulting in a new and original recommendation for heart disease. Heart disease prognosis is regarded as one of the most difficult and significant issues in medicine. If the condition is detected early on, the mortality rate can be managed, and preventive measures can be implemented as soon as feasible (21, 22).

In this study, a model for heart disease prediction with the help of machine learning is proposed. For this purpose, three machine learning algorithms, RF, SVM and LR, were used. The results were that the RF algorithm performed better in heart disease prediction compared to other methods (LR, SVM).

There are many studies in the literature for the prediction of heart diseases. In a study, algorithms such as J48, K Nearest Neighbors (KNN), Decision Tree and Naive Bayes (NB) were used for heart disease detection and the highest accuracy (83.732%) was obtained with J48 (23). Another article estimated whether a person has heart disease as a percentage using Data Mining classification techniques. In the study, Decision Tree, KNN, and Naive Bayes (NB) algorithms were used and heart diseases were estimated. NB achieved the highest accuracy (73.7%) in classifying heart diseases (24). In 2017, Hend Mansoor et al. looked examined the performance of LR and RF classification algorithms for assessing CVD patients' risk exposure. They demonstrated that the LR Model outperformed the RF classification technique. The LR Model had an accuracy of 89 percent, whereas the RF Model had an accuracy of 88 percent (25). A different paper Random Forest algorithm was used to classify heart disease. They classified new and unknown patients with 84.448% accuracy in the test dataset (26).

The performance measures obtained in most of the studies mentioned above are lower than the current study. In this study, hyperparameter optimization for LR, SVM and RF algorithms used to classify heart disease helped to create models with higher performance by choosing the most optimal model. In other words, hyper parameter optimization is an important step to create the most optimal model in machine learning models. In the study, many of the classical machine learning algorithms were tried and the three algorithms with the highest performance were continued to work. The current study predicted coronary artery heart disease more successfully than the literature. The values of Accuracy, Specificity, F1-score, Sensitivity, Negative predictive value, and Positive predictive value criteria obtained from the RF model were calculated as 0.929, 0.929, 0.928, 0.928, 0.929, and 0.928, respectively.

In conclusion, the present study aimed to find the best ML technique among the ML algorithms that are well accepted and easy to implement, and found that the proposed RF algorithm performs well, at least for this dataset. Therefore, the RF algorithm can be recommended for the development of prediction models for heart and different diseases in the future.

CONCLUSION

In conclusion, the RF model may be useful for early detection of coronary heart disease.

Financial disclosures: *The authors declared that this study hasn't received no financial support.*

Conflict of Interest: *The authors declare that they have no competing interest.*

Ethical approval: *Ethics committee approval is not required in this study.*

REFERENCES

1. Watanabe T, Ando K, Daidoji H, et al. A randomized controlled trial of eicosapentaenoic acid in patients with coronary heart disease on statins. J Cardiol. 2017;70:537-44.
2. Paynter NP, Balasubramanian R, Giulianini F, et al. Metabolic predictors of incident coronary heart disease in women. Circulation. 2018;137:841-53.
3. Wolters FJ, Segufa RA, Darweesh SK, et al. Coronary heart disease, heart failure, and the risk of dementia: a systematic

review and meta-analysis. Alzheimer's & Dementia. 2018;14:1493-504.

4. Dogan MV, Grumbach IM, Michaelson JJ, et al. Integrated genetic and epigenetic prediction of coronary heart disease in the Framingham Heart Study. PloS One. 2018;13:0190549.
5. Kaya ÖÜMO. Performance evaluation of multilayer perceptron artificial neural network model in the classification of heart failure. J Cognitive Systems. 2021;6:35-8.
6. Shameer K, Johnson KW, Glicksberg BS, et al. Machine learning in cardiovascular medicine: are we there yet? Heart. 2018;104:1156-64.
7. Birjandi SM, Khasteh SH. A survey on data mining techniques used in medicine. J Diabetes Metabol Disorders. 2021:1-17.
8. Kucukakcali Z, Balikçi Çiçek İ, Güldoğan E, et al. Assessment of associative classification approach for predicting mortality by heart failure. J Cognitive Systems. 2020;5:41-5.
9. Balikçi Çiçek İ, Küçükakçali Z, Çolak C. Associative classification approach can predict prostate cancer based on the extracted association rules. J Cognitive Systems. 2020;5:51-4.
10. Balikçi Çiçek İ, Küçükakçali Z. Classification Of Hypothyroid Disease With Extreme Learning Machine Model. The Journal of Cognitive Systems. 2020;5:64-8.
11. Küçükakçali Z, Balikçi Çiçek İ. Performance evaluation of the ensemble learning models in the classification of chronic kidney failure. J Cognitive Systems. 2020;5:55-9.
12. Arslan AK, Kucukakcali Z, Balikçi Çiçek İ, et al. A novel interpretable web-based tool on the associative classification methods: an application on breast cancer dataset. J Cognitive Systems. 2020;5:33-40.
13. Balikçi Çiçek İ, Küçükakçali Z. Classification of prostate cancer and determination of related factors with different artificial neural network. Middle Black Sea J Health Sci. 2020;6:325-32.
14. Küçükakçali Z, Balikçi Çiçek İ, Güldoğan E. Performance evaluation of the deep learning models in the classification of heart attack and determination of related factors. J Cognitive Systems. 2020;5:99-103.
15. Siddhartha M. Heart Disease Dataset (Comprehensive). Kaggle Inc. 2019.
16. Shah K, Patel H, Sanghvi D, et al. A comparative analysis of logistic regression, random forest and KNN models for the text classification. Augmented Human Res. 2020;5:1-16.
17. Kirasich K, Smith T, Sadler B. Random forest vs logistic regression: binary classification for heterogeneous datasets. SMU Data Science Review. 2018;1:9.
18. Pisner DA, Schnyer DM. Support vector machine. Machine Learning: Elsevier; 2020. p.101-21.
19. Probst P, Wright MN, Boulesteix AL. Hyperparameters and tuning strategies for random forest. wiley interdisciplinary reviews: data mining and knowledge discovery. 2019;9:1301.
20. Iwendi C, Bashir AK, Peshkar A, et al. COVID-19 patient health prediction using boosted random forest algorithm. Frontiers in Public Health. 2020;8:357.
21. Shah D, Patel S, Bharti SK. Heart disease prediction using machine learning techniques. SN Computer Science. 2020;1:1-6.
22. Kondababu A, Siddhartha V, Kumar BB, et al. A comparative study on machine learning based heart disease prediction. Materials Today: Proceedings. 2021.
23. Bahrami B, Shirvani MH. Prediction and diagnosis of heart disease by data mining techniques. JMultidisciplinary Engineering Sci Technol (JMEST). 2015;2:164-8.
24. Ashari A, Paryudi I, Tjoa AM. Performance comparison between Naïve Bayes, decision tree and k-nearest neighbor in searching alternative design in an energy simulation tool. International Journal of Advanced Computer Science and Applications (IJACSA). 2013;4:33-9.
25. Islam H, Elgendy Y, Segal R, et al. Risk prediction model for in-hospital mortality in women with ST-elevation myocardial infarction: A machine learning approach. J Heart & Lung. 2017:1-7.
26. Buettner R, Schunter M, editors. Efficient machine learning based detection of heart disease. 2019 IEEE International Conference on E-health Networking, Application & Services (HealthCom); 2019: IEEE.