ODU TIP DERGİSİ/ODU MEDİCAL JOURNAL (ODU MED J) October 2023;10(3):111-120

ARAŞTIMA MAKALES/ RESEARCH ARTICLE DOI: 10.56941/odutip.1345551

Establishing a Model for the Classification of Heart Attack and Identification of Associated Risk Factors with Machine Learning Methods

Zekeriya Doğan[1(ID)], Zeynep Küçükakçalı[2(ID)]

[1]Marmara University School of Medicine, Department of Cardiology, Istanbul
[2]Inonu University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Malatya, Turkey.

Received: 18 August 2023, Accepted: 20 September 2023, Published online: 29 October 2023

Abstract

Object: Increased survival rates in heart attacks (HAs) depend on early intervention and treatment. In this study, it is aimed to predict the factors that may be associated with HA and to determine which factor is more effective by using Stochastic Gradient Boosting (SGB) method, one of the machine learning methods.

Methods: An open access data set was used in the study. The 5-fold cross-validation method was used in modeling and the data set was divided into training and test data sets as 80%:20%. Accuracy (ACC), balanced accuracy (b-ACC), sensitivity (SE), specificity (SP), positive predictive value (ppv), negative predictive value (npv) and F1 score metrics were used for model evaluation.

Results: The results obtained from the performance metrics with the modeling were 98.9%, 98.7%, 99.4%, 98.0%, 98.8%, 99%, and 99.1% for ACC, b-ACC, SE, SP, ppv, npv, and F1-score, respectively. According to variable importance values, troponin and CK-MB appear to be associated with HA, respectively.

Conclusion: According to the modeling results, factors that may be associated with heart attack were determined with high accuracy by machine learning method. Thanks to these two enzymes, early diagnosis can be made in individuals at risk of having a heart attack, and poor prognosis and deaths can be prevented.

Key Words: Heart attack, classification, machine learning, risk factor

Makine öğrenimi yöntemleri ile kalp krizinin sınıflandırılması ve ilişkili risk faktörlerinin belirlenmesi için bir model oluşturulması

Özet

Amaç: Kalp krizlerinde (KK) hayatta kalma oranlarının artması, erken müdahale ve tedaviye bağlıdır. Bu çalışmada, makine öğrenmesi yöntemlerinden biri olan Stokastik Gradient Boosting (SGB) yöntemi kullanılarak KK ile ilişkili olabilecek faktörlerin tahmin edilmesi ve hangi faktörün daha etkili olduğunun belirlenmesi amaçlanmaktadır.

Yöntemler: Araştırmada açık erişimli veri seti kullanıldı. Modellemede 5 katlı çapraz doğrulama yöntemi kullanılmış ve veri seti %80:%20 olacak şekilde eğitim ve test veri setlerine bölünmüştür. Model değerlendirmesi için doğruluk (ACC), dengeli doğruluk (b-ACC), duyarlılık (SE), özgüllük (SP), pozitif tahmin değeri (ppv), negatif tahmin değeri (npv) ve F1 skoru metrikleri kullanıldı.

Bulgular: Modelleme ile performans metriklerinden elde edilen sonuçlar ACC, b-ACC, SE, SP, ppv, npv, F1 puanı çin %98,9, %98,7, %99,4, %98,0, %98,8, %99 ve %99,1 olmuştur. Değişken önem değerlerine göre sırasıyla troponin ve CK-MB'nin KK ile ilişkili olduğu görülmektedir.

Sonuç: Modelleme sonuçlarına göre kalp kriziyle ilişkili olabilecek faktörler makine öğrenmesi yöntemiyle yüksek doğrulukla belirlendi. Bu iki enzim sayesinde kalp krizi geçirme riski taşıyan bireylerde erken tanı yapılabilmekte, kötü gidişat ve ölümlerin önüne geçilebilmektedir.

Anahtar kelimeler: Kalp krizi, sınıflandırma, makine öğrenmesi, risk faktörü

Suggested Citation: Zekeriya D, Küçükakçalı Z. Establishing a model for the classification of heart attack and identification of associated risk factors with machine learning methods. ODU Med J, 2023;10(3): 111-120.

Address for correspondence/reprints:

Zeynep Küçükakçali

Telephone number: +90 (553) 373 24 04

E-mail: zeynep.tunc@inonu.edu.tr

INTRODUCTION

HA is one of the leading causes of death worldwide and will cause millions of deaths each year and there will be no end to it. According to the World Health Organization, approximately 17.7 million deaths from cardiovascular disease were estimated in 2015, almost 31% of all deaths worldwide (1, 2). The HA situation is the sudden stop of the heart without any warning, and early intervention is seen as the only way to prevent the mortality and morbidity associated with this condition (3-5). Therefore, it will be able to intervene immediately in patients who have had HA; in order to perform the life-saving treatment that patients need, a system that is very fast, has high accuracy and sensitivity, and most importantly, can diagnose with cheap costs and less equipment is needed.

Physicians and specialists often use electrocardiograms (ECG), echocardiograms, and blood tests to diagnose a HA. The most preferred diagnostic method is ECG and electrical signals passing through the human heart are recorded with the electrodes attached to the patient's chest in the ECG, and these signals are abnormal if there is an unhealthy heart. Therefore, if the patient is having a HA, the signals will be abnormal and this is a late intervention and will be ineffective to save lives. In addition, the specificity of the ECG is affected by individual variations in the anatomy of the heart, as well as by pre-existing heart diseases, injuries, and surgeries such as coronary artery bypass surgery. Therefore, it cannot be an early diagnosis argument (1, 6, 7). Echocardiogram, on the other hand, is used to determine whether any part of the heart is damaged using sound waves and creating images. It has almost the same disadvantages as ECG. Therefore, these disadvantages prevent an echocardiogram from being an early diagnosis method for HA detection (8). Both of these techniques are not preferred because of their disadvantages and the accuracy of identifying HA depends entirely on the doctor's knowledge, and experience with these methods.

Comparatively, detecting HA indicators in the blood is less expensive, faster, and more objective. Some proteins and enzymes, such as brain natriuretic peptide (BNP), troponin myoglobin, and creatine kinase isoenzymes, which can be identified by blood tests, seep into the blood slowly before a HA. CK-MB isoenzyme, which is a type of creatine kinase, which is especially located in heart muscle cells, increases in the blood, especially in heart diseases. In addition, another cardiac biomarker, troponin, is released when the heart muscle is damaged, as in a HA, and the more damage occurs to the heart, the more it increases in the blood. From the differences in these, it may be possible to determine the risk of HA with high accuracy. Thus, a diagnosis can be made to detect a HA early and to initiate treatment (1).

For this reason, in the current study, modeling was done with ML methods in order to determine the factors associated with HA using an open-access data set of patients with demographic characteristics and blood values. With the modeling, it was aimed to classify the patients with and without a HA and to determine the factors associated with HA.

METHODS

Dataset and Variables

The dataset used in the current study, which consists of the information of individuals who have had and have not had a HA, is a data set collected in the cardiology center of the Erbil region in Iraq in 2018. The dataset includes 1319 patients, and there are eight input and one output variable in the dataset. The variable, which is the output variable, has 2 categories, the negative category indicates no HA, and the positive category indicates a HA. Input variables consist of age, blood glucose, heart rate (impulse), systolic blood pressure, diastolic blood pressure, CK-MB (kcm), and troponin variables (9).

Biostatistics Analysis Phase

In the study, data were summarized as median (95 percent confidence intervals), and number (percentage). The Kolmogorov-Smirnov test was utilized to evaluate if the data was normal or not. The Mann-Whitney U test was used for statistical analysis of non-normally distributed data. $p<0.05$ was considered statistically significant. Analyzes were performed using IBM SPSS Statistics 25.

Modelling Phase

In the modeling phase with the data set, SGB, one of the tree-based methods among the ML techniques, was used to model patients with and without HA and to investigate the effect of input variables on the output variable. SGB is a method invented by Fridman by integrating randomization into the gradient boosting approach. In each iteration of this method, a sub-sample is randomly selected by using the permutation sampling approach. This sub-sample is used to calculate the current state of the model instead of all students, thus reducing the correlation between the established trees (10, 11). Unlike other ensemble learning methods, this method summarizes each tree (approximately 100 to 200 trees) generated as the process runs, rather than creating huge huge trees, and each observation is categorized according to the most common classification among trees. This form of separation distinguishes the SGB model from other augmentation techniques. In addition, this discrimination method reduces the sensitivity to outliers and unbalanced datasets. This method, which has a very high predictive power compared to other known algorithms, is also 5 times faster. Another and one of the most important features of the model is that it includes a set of regularization methods that can improve overall performance and reduce over-fitting and over-learning (10, 12). The data are separated as 80% training and 20% test data. The n-fold cross-validation method, one of the resampling methods, was utilized in this work to ensure model validity. In The n-fold cross-validation method; The dataset is first separated into n pieces, and the model is then applied to those pieces.

- In the second step, one of the n parts is used for testing, while the remaining n-1 parts are used for training.

- In the last stage, the cross-validation approach is evaluated using the average of the values collected from the models.

ACC, b-ACC, SE, SP, ppv, npv, and F1-score measures were utilized to assess the modeling performance.

Graphical summary

The graphical summary showing the biostatistical analyses, and modeling process applied in the study is shown in Figure 1.

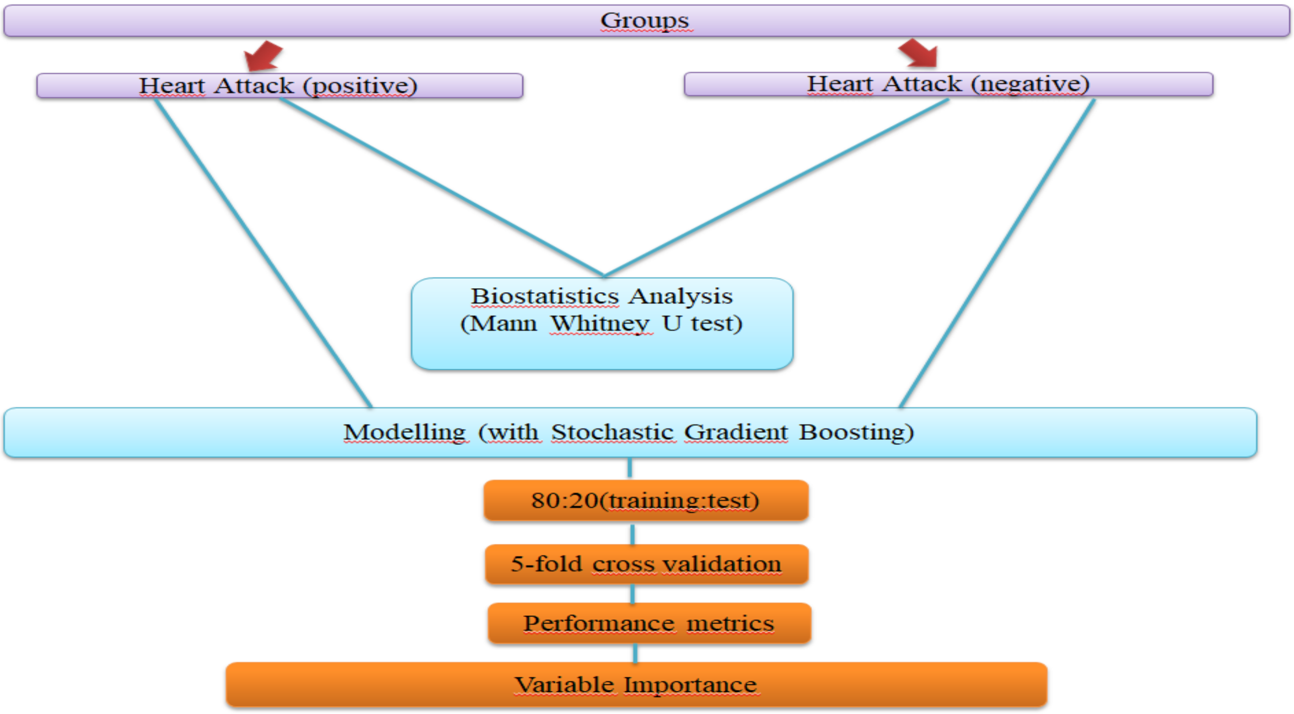

Figure 1. Biostatistical analysis and modeling process

RESULTS

The data set used in the study consists of 1306 individuals, of which 806 had a heart attack while 500 did not. The characteristics of 444 female and 862 male individuals are available. The average age is 56.22.

Biostatistical analysis results

When patients with and without HA were compared in terms of input variables such as age, heart rate, systolic blood pressure, diastolic blood pressure, blood glucose, CK-MB (kcm), and troponin; there was a statistically significant difference between the 2 groups in age, CK-MB (kcm) and troponin variables. However, statistical significance was not found in other variables. The results of the analyzes are given in Table 1

Modelling Results

The values of the performance metrics obtained by modeling with SGB using individuals with and without a heart attack are given in Table 2. Graphics of performance metrics are given in Figure 2. The graph of the variable importance obtained as a result of the modeling is given in Figure 3.

Table 1. Comparison of output variable in terms of input variables

Variables	Group		p*
	HA (-)	HA (+)	
	Median (95,0% Lower CL for Median; 95,0% Upper CL for Median		
age	52(50-55)	60(60-62)	0.000
impluse	75(74-78)	74(74-76)	0.779
systolic blood pressure	125(124-129)	122(120-125)	0.275
diastolic blood pressure	72(71-75)	71(70-74)	0.902
glucose	116.5(111-122)	116(114-122)	0.554
CK-MB	2.31(2.11-2.53)	3.76(3.28-4.29)	0.000
troponin	0.006(0.006-0.007)	0.044(0.037-0.053)	0.000

*:Mann Whitney U test

Table 2. Performance metrics values obtained after modeling

Performance Metrics	Performance Metrics Value (%)
ACC	98.9
b-ACC	98.7
SE	99.4
SP	98.0
ppv	98.8
npv	99
F1-score	99.1

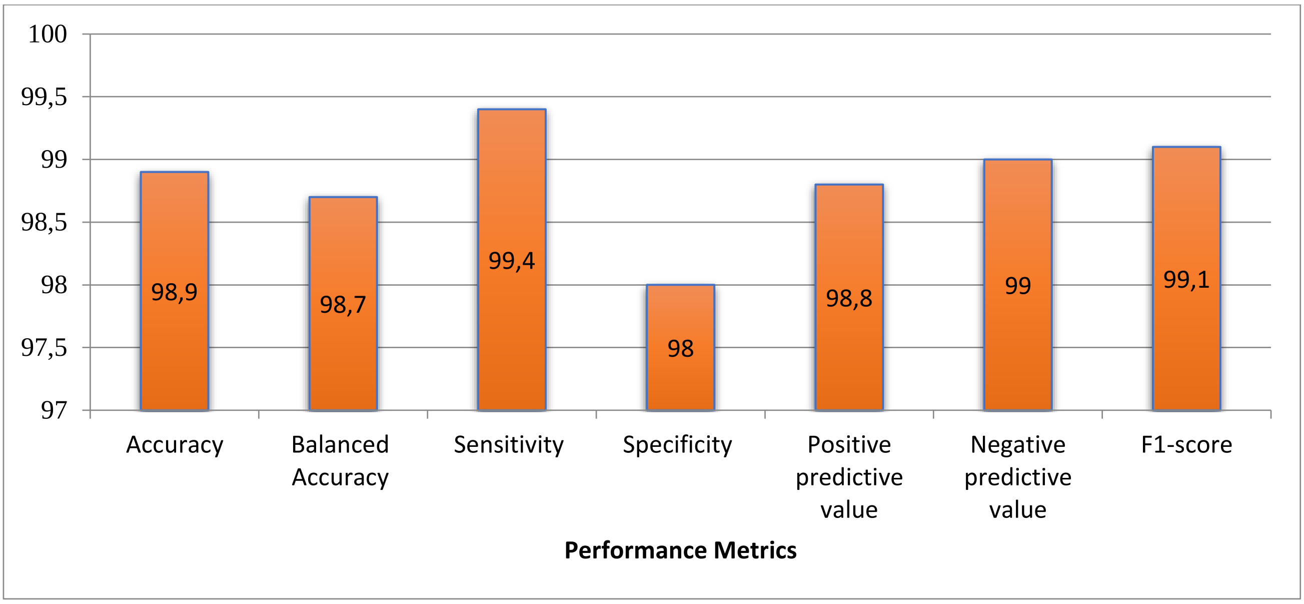

Figure 2. Graph of performance metrics

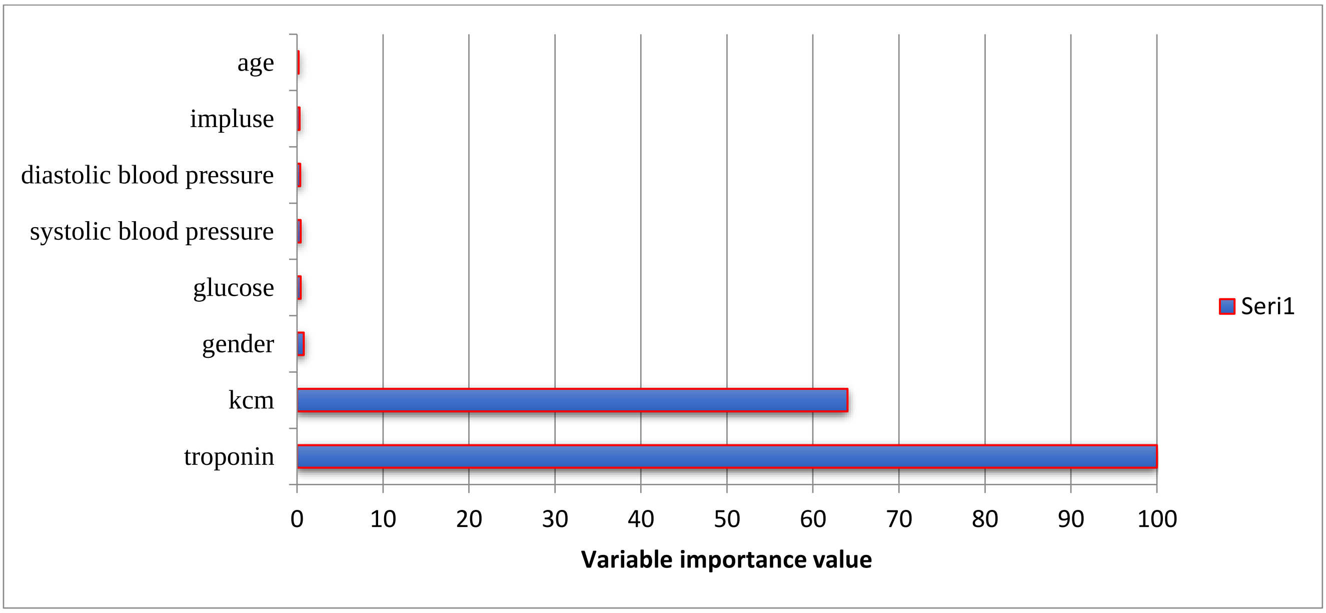

Figure 3. Variable Importance Graph

DISCUSSION

Cardiovascular diseases are among the most common causes of death today and are quite common in Western countries. Studies have shown that the death rate from cardiovascular diseases in the world will increase from 28.9% to 36.3% between 1990 and 2020 (13). The most common cardiovascular disease is a HA. HA is defined as the damage and deterioration of heart muscle cells that cannot receive sufficient oxygen due to the deterioration of blood supply in any part of the heart (14). In the case of a HA, if the heart muscle is without oxygen for a long time, the outcome will be worse and death will occur. Almost half of deaths from heart attacks occur within the first hour. This mortality rate reaches 80% in the first 24 hours following the HA (15). HA, which is very common, especially in the productive age group of the society and causes serious problems due to complications in the post-acute period, and can even result in death in advanced stages, is an important public health problem. Although there have been developments in diagnosis and treatment methods related to the disease in recent years, it is one of the most important causes of morbidity and mortality in our country and industrialized societies (13).

Evaluation of the incidence and case mortality of HA, which is known as one of the most important components of the cardiovascular disease burden, will be decisive in the reduction of coronary disease mortality (16). For this reason, the need for diagnostic markers to predict HA is increasing. It is known that some isoenzymes tend to increase in the blood before a HA occurs. Therefore, the presence of a diagnostic model based on these enzymes will be able to detect the risk of HA at an early stage and reduce the mortality that may occur. The most commonly used diagnostic methods in the diagnosis of HA are physical examination, Electrocardiography (ECG) containing Q waves, and the results of tests such as creatinine kinase, Myoglobin and Troponin (17). Creatine kinase and troponin enzymes in the data set used in the study have been used reliably in the diagnosis of HA for many years (18). Creatine kinase enzyme starts to rise 4 to 9 hours after myocardial injury and reaches its peak at 24 hours. For this reason, patients with chest pain and signs of HA can be diagnosed by the value of this enzyme in the blood. This enzyme returns to its normal range between 48-72 hours. (19). Troponin, a protein specific to skeletal and cardiac muscle fibers, mixes with blood from the muscle due to damage to the heart in unexpected situations such as a HA. The troponin value begins to rise at the blood level in relation to cardiac, or in other words, damage to the heart. The troponin level, which can be detected in the laboratory test performed in the first hour after the injury, reaches the maximum level in the 24th hour and maintains its positive value for 1 week (20).

It may be possible to detect HA with these cardiac biomarkers. For this reason, a diagnosis system based on these markers can be used for early diagnosis and to predict risk in patients who apply to the hospital with chest pain. For this purpose, in the present study, it was aimed to determine the risk factors that may be associated with HA by making a ML-based model that can detect HA by using the data set of 1306 patients' blood values. In this context, the variable importance values obtained as a result of the modeling and the variables that most explain the HA were obtained and their relations with the HA were confirmed.

According to the results of the statistical analysis, a statistically significant difference was found between the heart attack (+) and heart attack (-) groups in age, troponin, and CK-MB variables, and no statistical differences were observed in other variables. In the heart attack (+) group, the 0.038-unit increase in the troponin variable and the 1.45-unit increase in the CK-MB variable were found to be significant compared to the heart attack (-) group.

The values of ACC, b-ACC, SE, SP, ppv, npv, and F1-score performance metrics obtained according to the modeling results made with the SGB method were 98.9%, 98.7%, 99.4%, 98.0%, 98.8%, 99%, and 99.1%. According to the results obtained here, the modeling method used classifies the HA situation with a very high rate, and these results showed that the model used was successful in the prediction of HA. In addition, when the variable importance values obtained as a result of the modeling were examined, it was seen that the most important parameters associated with HA were troponin and CK-MB (kcm). Other variables, on the other hand, seem to have a low effect on HA. These results support the literature and the risk of HA can be evaluated with troponin, and CK-MB an isoenzyme of creatin kinase. With this evaluation, the plight of individuals can be prevented and possible deaths can be prevented by intervening early in HA.

Ethics Committee Approval: No required

Peer-review: Externally peer-reviewed

Author Contributions: Concept: ZD, ZK, Design: ZD, ZK, Literature search: ZD, ZK, Data Collection and Processing: ZD, ZK, Analysis or Interpretation: ZD, ZK, Writing: ZD, ZK,

Conflict of Interest: The authors declared no conflict of interest.

Financial Disclosure: The authors declared that this study has not received no financial support.

REFERENCES

1. Liu Z, Meng D, Su G, Hu P, Song B, Wang Y, et al. Ultrafast Early Warning of Heart Attacks through Plasmon-Enhanced Raman Spectroscopy using Collapsible Nanofingers and Machine Learning. Small (Weinheim an der Bergstrasse, Germany). 2023;19(2):e2204719.
2. Maghdid SS, Rashid T, Ahmed S, Zaman K, Rabbani M. Analysis and prediction of heart attacks based on design of intelligent systems. J Mech Contin Math Sci. 2019;14(4):628-45.
3. Organization WH. Cardiovascular diseases (cvds). http://www who int/mediacentre/factsheets/fs317/en/index html. 2009.
4. https://www.nhs.uk/conditions/heart-

attack/diagnosis/#:~:text=An%20ECG%20machine%20records%20these,about%205%20minutes%20to%20do. [cited 2023 07.04].

5. Arslankaya S, Çelik MT. Prediction of heart attack using fuzzy logic method and determination of factors affecting heart attacks. International Journal of Computational and Experimental Science and Engineering. 2021;7(1):1-8.
6. Zimetbaum PJ, Josephson ME. Use of the electrocardiogram in acute myocardial infarction. New England Journal of Medicine. 2003;348(10):933-40.
7. Gupta V, Mittal M. A novel method of cardiac arrhythmia detection in electrocardiogram signal. International Journal of Medical Engineering and Informatics. 2020;12(5):489-99.
8. Leung DY, Davidson PM, Cranney GB, Walsh WF. Thromboembolic risks of left atrial thrombus detected by transesophageal echocardiogram. The American journal of cardiology. 1997;79(5):626-9.
9. Anshori M, Haris MS. Predicting Heart Disease using Logistic Regression. Knowledge Engineering and Data Science. 2022;5(2):188-96.
10. Friedman JH. Stochastic gradient boosting. Computational statistics & data analysis. 2002;38(4):367-78.
11. Ye J, Chow J-H, Chen J, Zheng Z, editors. Stochastic gradient boosted distributed decision trees. Proceedings of the 18th ACM conference on Information and knowledge management; 2009.
12. Lawrence R, Bunn A, Powell S, Zambon M. Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis. Remote sensing of environment. 2004;90(3):331-6.
13. Zeynep T, Çiçek İB, Güldoğan E. Performance evaluation of the deep learning models in the classification of heart attack and determination of related factors. The Journal of Cognitive Systems. 2020;5(2):99-103.
14. Halıcı Z, Yasin Bayır HS, Çadırcı E, Keleş MS, Bayram E. Investigation of the Effects of Amiodarone on Serum Eritropoietin Levels in İsoproterenol-induced Acute and Chronic Myocardial Infarct Model of Rats. Avrasya J Med, 2006;38(3): 68-72.
15. Storrow AB, Gibler WB. Chest pain centers: diagnosis of acute coronary syndromes. Annals of emergency medicine. 2000;35(5):449-61.
16. Roger VL. Epidemiology of myocardial infarction. The Medical clinics of North America. 2007;91(4):537-52; ix.
17. Özdemir G, Bilen Ö, Ateş SC. Establishment of a Decision Support System for Determining the Risk Probability of Heart Attack in Hospital Emergency Visitors. Düzce Üniv Bilim ve Tek Der. 2022;10(4):2093-106.

18. Yöntem M, Erdoğdu BS, Akdoğan M, Kaleli S. The Importance of Cardiac Markers in Diagnosis of Acute Myocardial Infarction. Online Türk Sağlık Bilimleri Dergisi. Online Türk Sağlık Bilimleri Der. 2017;2(4):11-7.
19. Lewandrowski K, Chen A, Januzzi J. Cardiac markers for myocardial infarction: a brief review. Pathology Patterns Reviews. 2002;118(suppl_1):S93-S9.
20. Filatov V, Katrukha A, Bulargina T, Gusev N. Troponin: structure, properties, and mechanism of functioning. Biochemistry c/c of Biokhimiia. 1999;64:969-85.