ODU TIP DERGİSİ/ODU MEDİCAL JOURNAL (ODU MED J) December 2022;9(3):104-109

ARAŞTIMA MAKALES/ RESEARCH ARTICLE

Detection of Coronary Heart Disease by Data Mining Methods Using Clinical Data

Burak Yağın[1](ID)

[1]Department of Biostatistics and Medical Informatics, Faculty of Medicine, İnönü University, Malatya, Turkey

Received: 03 October 2022, Accepted: 19 December 2022, Published online: 31 December 2022

Abstract

Objective: The major cause of death worldwide is anticipated to continue to be heart disease, one of the most prevalent diseases in the globe. Therefore, the aim of this study is to classify and predict coronary heart disease using the Random forest (RF) method, which is one of the predictive algorithms of data mining.

Methods: The dataset was divided into 80% training and 20% test sets to avoid bias. Then the model was trained on the training set and tested on the test set. In the study, the RF algorithm was used for the classification of coronary heart disease. The performance of the RF model was evaluated with balanced accuracy, accuracy, sensitivity, F1-score, negative predictive value, positive predictive value, specificity, and confusion matrix results.

Results: The data set included clinical data of 1190 patients, 281 (23.6%) female, and 909 (76.4%) males. Based on the results from the RF model, balanced accuracy, accuracy, sensitivity, F1-score, negative predictive value, positive predictive value, and specificity for heart disease were 0.945, 0.945, 0.920, 0.941, 0.931, 0.963, and 0.968, respectively.

Conclusion: According to the performance measures obtained in the test set for coronary heart disease (CHD), the RF model performed well. As a result, the proposed model can provide clinicians with clinical decision support for the preliminary diagnosis of CHD patients.

Key Words: Heart disease, artificial intelligence, data mining, random forest, parameter optimization

Klinik Veriler Kullanılarak Veri Madenciliği Yöntemleriyle Koroner Kalp Hastalığının Tespiti

Özet

Amaç: Dünya çapında en önemli ölüm nedeninin, dünyadaki en yaygın hastalıklardan biri olan kalp hastalığı olmasının devam etmesi beklenmektedir. Bu nedenle bu çalışmanın amacı, veri madenciliğinin tahmin edici algoritmalarından biri olan Rastgele Orman (RF) yöntemini kullanarak koroner kalp hastalığının sınıflandırılması ve tahmin edilmesidir.

Yöntemler: Modelin doğrulaması için veri seti %80 eğitim ve %20 test setlerine bölünmüştür. Daha sonra model eğitim seti üzerinde eğitilmiş ve test seti üzerinde test edilmiştir. Çalışmada koroner kalp hastalığı sınıflandırması için RF algoritması kullanılmıştır. RF modelinin performansı, dengeli doğruluk, doğruluk, duyarlılık, F1-skor, negatif tahmin değeri, pozitif tahmin değeri, seçicilik ve karışıklık matrisi sonuçları ile değerlendirildi.

Bulgular: Veri seti 281 (%23.6) kadın ve 909 (%76.4) erkek olmak üzere 1190 hastanın klinik verilerini içermektedir. RF modelinden elde edilen sonuçlara göre kalp hastalığı için dengeli doğruluk, doğruluk, duyarlılık, F1-skor, negatif tahmin değeri, pozitif tahmin değeri ve seçicilik sırasıyla 0.945, 0.945, 0.920, 0.941, 0.931, 0.963 ve 0.968 idi.

Sonuç: Koroner kalp hastalığı (KKH) için test setinde elde edilen performans ölçütlerine göre, RF modeli iyi performans gösterdi. Sonuç olarak önerilen model, klinisyenlere KKH hastalarının ön tanısı için klinik karar desteği sağlayabilir.

Anahtar Kelimeler: Kalp hastalığı, kardiyovasküler hastalık, yapay zeka, veri madenciliği, rastgele orman

Suggested Citation: Yağın B. Detection of Coronary Heart Disease by Data Mining Methods Using Clinical Data. ODU Med J, 2022;9(3): 104-109

Address for correspondence/reprints: **Telephone number:** +90 (506) 777 14 06

Burak Yağın **E-mail:** burak.yagin@inonu.edu.tr

INTRODUCTION

One of the most frequent reasons for a heart attack is a clogged coronary artery. The primary cause of death is still heart disease. For the best treatment, the prevention of poor clinical outcomes, and mortality, early diagnosis of this condition is essential. The risk factors for cardiovascular disease (CVD) include male gender, advanced age, inactivity, low education, unemployment, economic situation, psychological state, and hypertension. Heart attack risk can be decreased by eating a diet high in salt, consuming less fat, cholesterol, alcohol, and smoking, exercising frequently, and decreasing weight (1-3).

Artificial intelligence (AI) methods are often used to quickly, and accurately diagnose heart disease (4). AI is a technology that improves performance and productivity in the field by automating processes or tasks that previously required manpower. AI makes it possible for machines to learn from experience, adapt to new inputs, and perform human-like tasks. Using AI technologies, computers can be trained to perform specific tasks by processing large amounts of data and recognizing patterns in the data. These technologies are used in medicine for purposes such as classifying and predicting diseases and determining the most important factors that cause diseases (5-10).

One of the fastest expanding subfields of AI is data mining. Data mining is the process of analyzing large data sets to reveal hidden critical decision-making information for future analysis. The amount of data obtained in the field of medicine is increasing day by day. Predictive algorithms of data mining can be used to extract useful, important, and relevant information from these data for the relevant situation (11-13).

Data mining methods are needed both in estimating heart attacks and other heart diseases and in determining risk factors. From this point of view, in this study, using the risk factors related to heart attack, a successful model is created to predict the disease by classifying it quickly and accurately with minimum error.

METHODS

Dataset

The heart disease dataset from the IEEE Data Port database (https://ieee-dataport.org/open-access/heart-disease-datasetcomprehensive#files) was obtained. The data set included clinical data of 1190 patients, 281 (23.6%) female, and 909 (76.4%) male (14).

Data processing and modeling

The dataset was divided into 80% training and 20% test sets to avoid bias. Then the model was trained on the training set and tested on the test set. Then, 5-fold cross-validation was used as a resampling method. Balanced accuracy, accuracy, sensitivity, F1-score, negative predictive value, positive predictive value, and specificity values were calculated to evaluate the performance of the model.

Random Forest (RF)

RF algorithm, which can be used for classification and regression analysis, is an ensemble classification method. The goal of this community classifier is to improve classification accuracy by building many decision trees. During the training phase, RFs create various decision trees and labels based on the majority. The main distinction between RFs and decision tree methods is that identifying the root node and splitting the nodes essentially work arbitrarily. The RF approach is taken into consideration in this study since it can resolve the over-learning issue and is effective at detecting noise and outliers. It is also one of the best techniques for identifying the most significant feature among the features of the data set (15-20).

RESULTS

Table 1 contains the confusion matrix for the RF model in the test data set. While the RF model could not distinguish 4 patients with heart disease, it accurately predicted 121 patients.

Table 1. Confusion Matrix for the RF model

Predict	References	
	Control	Heart disease
Control	103	4
Heart disease	9	121

Table 2 shows the performance criteria obtained with the RF model in the test data set. In Figure 1, the learning curve of the model in train, test and resampling is given. Based on the results from the RF model, balanced accuracy, accuracy, sensitivity, F1-score, negative predictive value, positive predictive value, and specificity for heart disease were 0.945, 0.945, 0.920, 0.941, 0.931, 0.963, and 0.968, respectively.

Table 2. Values for performance metrics for the RF model

Table 2. Values for performance metrics for the RF model

Metric	Value
Balanced Accuracy	0.945
Accuracy	0.945
Sensitivity	0.920
F1-score	0.941
Negative predictive value	0.931
Positive predictive value	0.963
Specificity	0.968

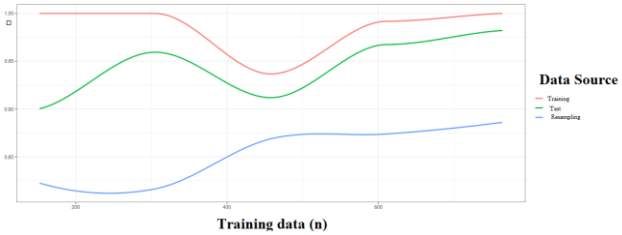

Figure 1. The learning curve for model accuracy performance

In Figure 2, the order of importance and the importance coefficients of the variables included in the model are given according to the contribution of the RF model to the prediction performance. According to the results of the study,

the slope of the peak exercise being the unsloping of the ST segment and the maximum heart rate were the first two features that contributed the most to the estimation performance of the RF model. In addition, the peak exercise ST segment down sloping did not contribute to the determination of heart disease.

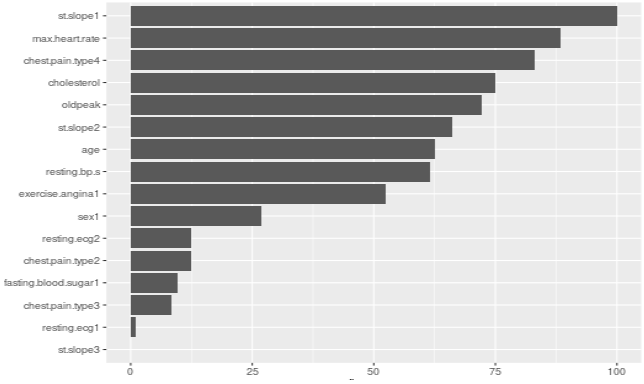

Figure 2. The graphic of feature importance for the RF model

DISCUSSION

Heart diseases are one of the leading causes of death in the world. A heart attack is a physiological condition characterized by severe chest pain and the possibility of death as a result of heart failure caused by a problem with the coronary arteries. A heart attack occurs when the heart's oxygen supply is cut off due to a rapid decrease or interruption of blood flow in the vessels that feed the heart. The clogged vessel can damage or even kill the heart muscle. Heart attack is the most common health problem in rich countries. It is an important health problem that has become more common in underdeveloped countries. The World Health Organization estimates that more than 16 million people die each year due to heart attacks. These results are one-third of all deaths (1, 21-23).

The development of health care has enabled studies such as identifying the variables underlying heart diseases and avoiding its occurrence. Recent studies have been able to pinpoint heart disease risk factors, but many scientists agree that further studies are necessary before we can apply this knowledge to lower the prevalence of heart disease. Different factors may contribute to heart problems. Reducing these heart disease risk factors may really help prevent heart disease, according to certain research studies. The prevention of heart disease risk has been the subject of numerous studies. There will be more options to avoid heart disease as more studies on the condition are conducted (24, 25).

For this reason, early detection and treatment of heart diseases are very important. The use of data mining methods for heart diseases is an important research topic. Based on this, we predicted heart disease with a fast and high-performance model using the heart disease dataset containing clinical data in this study.

The performance criteria obtained from the RF method in the study; balanced accuracy, accuracy, sensitivity, F1-score, negative predictive value, positive predictive value, and specificity for heart disease were 0.945, 0.945, 0.920, 0.941, 0.931, 0.963, and 0.968, respectively.

CONCLUSION

In conclusion, this study can suggest an RF algorithm for the development of predictive models for heart diseases and the developed model can help clinicians in the early diagnosis of cardiac patients.

Ethics Committee Approval: Ethics committee approval is not required in this study.

Peer-review: Externally peer-reviewed.

Author Contributions:

Concept: Design: Literature search: Data Collection and Processing: Analysis or Interpretation: Written by: BY

Conflict of Interest: The author declared no conflict of interest

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

REFERENCES

1. Yilmaz R, Yağin FH. Early detection of coronary heart disease based on machine learning methods. Medical Records. 2022;4(1):1-6.
2. Absar N, Das EK, Shoma SN, Khandaker MU, Miraz MH, Faruque M, et al., editors. The Efficacy of Machine-Learning-Supported Smart System for Heart Disease Prediction. Healthcare; 2022: MDPI.
3. Demir De. Classification Of Heart Attack Risks with Artificial Intelligence Methods. Artificial Intelligence Applications and Their Economic Effects on The Field Of Health Care. 2022:9.
4. Yilmaz R, Yağin FH. A comparative study for the prediction of heart attack risk and associated factors using MLP and RBF neural networks. The Journal of Cognitive Systems. 2021;6(2):51-4.
5. Paksoy N, Yağin FH. Artificial Intelligence-based Colon Cancer Prediction by Identifying Genomic Biomarkers. Medical Records.4(2):196-202.
6. Akbulut S, Yagin FH, Colak C. Prediction of COVID-19 Based on Genomic Biomarkers of Metagenomic Next. Biomed Eng. 2020;14:4-15.
7. Hamet P, Tremblay J. Artificial intelligence in medicine. Metabolism. 2017;69:S36-S40.
8. Haick H, Tang N. Artificial intelligence in medical sensors for clinical decisions. ACS nano. 2021;15(3):3557-67.
9. Koteluk O, Wartecki A, Mazurek S, Kołodziejczak I, Mackiewicz A. How do machines learn? artificial intelligence as a new era in medicine. Journal of Personalized Medicine. 2021;11(1):32.
10. Bhinder B, Gilvary C, Madhukar NS, Elemento O. Artificial intelligence in cancer research and precision medicine. Cancer Discovery. 2021;11(4):900-15.
11. Yağin FH, Yağin B, Arslan AK, Çolak C. Comparison of Performances of Associative Classification Methods for Cervical Cancer Prediction: Observational Study. Turkiye Klinikleri Journal of Biostatistics. 2021;13(3).
12. Akbulut S, Yagin FH, Colak C. Prediction of Breast Cancer Distant Metastasis by Artificial Intelligence Methods from an Epidemiological Perspective. Istanbul Medical Journal. 2022;23(3).

13. Han J, Pei J, Tong H. Data mining: concepts and techniques: Morgan kaufmann; 2022.
14. Siddhartha M. Heart disease dataset (comprehensive). IEEE Dataport. 2020.
15. Perçin İ, Yağin FH, Arslan AK, Çolak C, editors. An interactive web tool for classification problems based on machine learning algorithms using java programming language: data classification software. 2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT); 2019: IEEE.
16. Cicek İB, İlhami S, Yağin FH, Colak C. Development of a Python-Based Classification Web Interface for Independent Datasets. Balkan Journal of Electrical and Computer Engineering.10(1):91-6.
17. Oshiro TM, Perez PS, Baranauskas JA, editors. How many trees in a random forest? International workshop on machine learning and data mining in pattern recognition; 2012: Springer.
18. Khan MA, Memon SA, Farooq F, Javed MF, Aslam F, Alyousef R. Compressive strength of fly-ash-based geopolymer concrete by gene expression programming and random forest. Advances in Civil Engineering. 2021;2021.
19. Gupta VK, Gupta A, Kumar D, Sardana A. Prediction of COVID-19 confirmed, death, and cured cases in India using random forest model. Big Data Mining and Analytics. 2021;4(2):116-23.
20. Mohana RM, Reddy CKK, Anisha P, Murthy BR. Random forest algorithms for the classification of tree-based ensemble. Materials Today: Proceedings. 2021.
21. Bhuyan SK, Sahoo H, Habada SK, chandra Pradhan P. Prevalence of Coronary Heart Disease (CHD) risk factors-a hospital based cross sectional survey from South Odisha.
22. Aravinda C, Lin M, Kumar U, Reddy K, Prabhu GA. Deep Learning Techniques for Data Analysis Prediction in the Prevention of Heart Attacks. Internet of Healthcare Things: Machine Learning for Security and Privacy. 2022:217-40.
23. Kavitha M, Gnaneswar G, Dinesh R, Sai YR, Suraj RS, editors. Heart disease prediction using hybrid machine learning model. 2021 6th International Conference on Inventive Computation Technologies (ICICT); 2021: IEEE.
24. Humphries SE, Drenos F, Ken-Dror G, Talmud PJ. Coronary heart disease risk prediction in the era of genome-wide association studies: current status and what the future holds. Circulation. 2010;121(20):2235-48.
25. Kim J, Lee J, Lee Y. Data-mining-based coronary heart disease risk prediction model using fuzzy logic and decision tree. Healthcare informatics research. 2015;21(3):167-74.