Ann Med Res

Current issue list available at AnnMedRes

Annals of Medical Research

journal page: www.annalsmedres.org

Analyzing and detecting risk factors for the diagnosis of angina pectoris with machine learning

Onural Ozhan[a,*], Ipek Balikci Cicek[b], Zeynep Kucukakcali[b]

[a] *Inonu University, Faculty of Medicine, Department of Pharmacology, Malatya, Türkiye*
[b] *Inonu University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Malatya, Türkiye*

ARTICLE INFO

Keywords:
Angina pectoris
Machine learning
Bagged CART
Modeling
Variable importance

Received: Feb 02, 2023
Accepted: Apr 03, 2023
Available Online: 28.04.2023

DOI:
10.5455/annalsmedres.2023.02.043

Abstract

Aim: To classify angina pectoris (AP) in women by applying the Bagged CART approach, which is one of the machine learning (ML) methods, to the open-access AP dataset. Another aim is to reveal the risk factors associated with AP in women through modeling.

Materials and Methods: In the current study, modeling was done with the Bagged CART technique utilizing an open-access data set containing the factors associated with AP. Model results were assessed with accuracy (ACC), sensitivity (Sen), balanced accuracy (BACC), positive predictive value (PPV), specificity (Spe), negative predictive value (NPV), and F1-score performance criteria. In addition, a 5-fold cross-validation approach was applied in the modeling phase. Finally, variable importance was derived with modeling.

Results: ACC, BACC, Sen, Spe, PPV, NPV, and F1-score from Bagged CART modeling were 98.5%, 98.5%, 99.0%, 98.0%, 98.0%, 99.0%, and 98.5%, respectively. Depending on the variable importance values calculated for the input variables investigated in the current study, age, family history of myocardial infarction: yes, the average number of cigarettes smoked per day smoking status: current, family history of angina: yes, hypertensive condition: moderate, smoking status: ex, hypertensive condition: mild, family history of stroke: yes, whether the woman has diabetes: yes were obtained as the most important variables associated with AP.

Conclusion: With the ML model used, the AP dataset was classified successfully, and the associated risk factors were revealed. ML models can be used as clinical decision support systems for early diagnosis and treatment.

Copyright © 2023 The author(s) - Available online at www.annalsmedres.org. This is an Open Access article distributed under the terms of Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

Introduction

Cardiovascular diseases (CVDs) are playing an increasing role in becoming the major reason for mortality and morbidity worldwide and are one of the most important health problems [1]. CVDs problems are believed to be the most common non-communicable illness in the globe. CVDs, in addition to being the most frequent illnesses, are among the leading causes of death worldwide, with the number of deaths attributable to such diseases increasing year after year. In 2012, CVD was responsible for 17.5 million deaths; this number climbed to 17.9 million in 2016 and is projected to reach 22.2 million by 2030. In 2016, 31 percent of all deaths were attributable to CVD. CVDs are predicted to remain the number one cause of death for a long time [2].

Ischemic heart disease (IHD) is the major reason of death among CVDs. IHD is a clinical term that describes the problem of not getting enough blood and oxygen to the myocardium and the resultant clinical images [3]. The most common cause of IHD is the narrowing of the coronary artery lumen due to atherosclerotic plaques resulting in decreased blood flow. The decreased blood flow in the coronary arteries leads to an imbalance between myocardial oxygen demand and blood flow [4]. Angina pectoris (AP) is the most prominent symptom of IHD. AP is a clinical condition accompanied by discomfort or ache in the arm, shoulder, back, chest and jaw [5]. AP is caused by a rise in the heart's oxygen demand at the cell level or a fall in the oxygen concentration in the myocardium, according to AP physiopathology. Although narrowing of the coronary arteries is commonly blamed for decreased oxygen supply, aberrant rise in oxygen consumption, such as raised heart rate, untreated hypertension, and riced myocardial contractility, can also cause AP. AP affects 0.1-20% of the general population aged 45-74, with the frequency increasing with age. AP is expected to affect 20,000-40,000 per-

*Corresponding author:
Email address: onural.ozhan@inonu.edu.tr (Onural Ozhan)

sons per million in most European nations [6].

Since CVD is significantly higher in men than in women, CVD has traditionally been seen to be a "man" disease. But at the same time, CVD is the leading cause of death in women worldwide [7]. Despite having less obstructive coronary artery disease and greater left ventricular function in women, AP is associated with higher morbidity, mortality, and quality-of-life outcomes in women than in men. In the absence of major coronary artery obstructive disease, women suffering chest discomfort and myocardial ischemia face a considerable risk of death and morbidity [8].

Machine learning (ML) is one of the most appropriate techniques for developing models used in the healthcare industry to diagnose diseases [9]. ML is a term in artificial intelligence that is used to develop models or systems that can learn from existing datasets and predict future occurrences [10]. ML is automatically unearthing usable information and detecting hidden patterns in massive data warehouses. From collecting raw data to discovering noteworthy patterns, ML involves several procedures, including data cleansing, transformation, selection, evaluation, and the presentation of researched information to customers. In the healthcare business, ML algorithms have been applied to acquire important insights for enhanced diagnostic decision-making. In addition, it has assisted the systems in learning the diagnosis data, finding key patterns during the learning process, and limiting human interference in decision-making [11].

CART is a continually evolving nonparametric ML approach for regression and classification problems [12]. To study the relationship between response variables and predictors, CART divides data recursively through binary partitioning. Bagged CART is an improvement to the CART algorithm that integrates bagging techniques with CART to increase predictive model performance and decrease overfitting [13]. In this method, each classifier develops and stores its model by classifying a subset of the data. Lastly, the class with the highest votes is chosen as the final classifier based on the voting intentions among these categories [14].

This work aimed to forecast AP in women and determine risk factors associated with AP in women by using the Bagged CART approach to an open-access AP dataset comprising exclusively female patients.

Materials and Methods

Dataset and variables

The open-access "Project Angina Data Set" from https://www.kaggle.com/snehal1409/predict-angina was utilized in the study to explore the prediction of AP in women and to identify risk variables related to AP. In the utilized data set, there are 100 (50%) no responses and 100 (50%) affirmative responses for 200 patients. The study's data set is made up of (whether the woman has angina (no/yes)), smokes (smoking status (current, ex-, non-smoker)), age, angfam (family history of angina (no/yes)), hyper (hypertensive condition (absent, mild, moderate)), cig (average number of cigarettes smoked per day), myofam (family history of myocardial infarction (no/yes)), diabetes (whether the woman has diabetes (no/yes)) variables, strokefam (a family history of stroke (no/yes)).

Bagged classification and regression trees (Bagged CART)

CART is a popular non-parametric classification approach that uses decision trees. Breiman et al. devised this algorithm (1984) [15]. It uses binary trees as its algorithm. This tree, along with others, serves as the foundation for more complex algorithms such as Random Forest (RF). Before generating the decision tree, the CART method separates the input into binary components. The Gini index is used by the CART to specify which variables are given more information for data categorization. When classifying, factors with smaller Gini indices are given more weight. CART uses trial and mistake to get the ideal value for the partition point in every dimension or variable with the lowest Gini index [16].

The bagging strategy can significantly enhance the accuracy of the CART, which is sighted as an unstable model [17]. The bagged CART significantly improves classification performance, eliminates overfitting, and substantially reduces prediction variation. First, CART divides training sample units recursively using a predefined number of variables. It then examines all predictive variables to determine which binary split in a predictive variable is least likely to depart from the anticipated response variable. When generating homogenous final nodes in a hierarchical tree, the process is frequently repeated for each outcome obtained from the initial split. When cross-validation produces the lowest error rate, CART prunes the trees to prevent overfitting [12, 18].

Modeling

Bagged CART was employed in the modeling portion of the study for the aforementioned data set. A 5-fold cross-validation procedure was employed for the analysis. We used ACC, BACC, PPV, NPV, Sen, Spe, and F1-score as performance metrics. Furthermore, the variable importances demonstrate how much the input variables contribute to the output variable. R studio 4.2.1 was used for modeling.

Results

The statistical analysis results of the quantitative independent variables in terms of the target variable (status) are

Table 1. Descriptive statistics for quantitative independent variables.

Variables	Status		p^*
	No	Yes	
	Median (min-max)	Median (min-max)	
Age	49 (29-74)	57 (29-73)	**<0.001**
The average number of cigarettes smoked per day	0 (0-30)	12 (0-40)	**<0.001**

*: Mann Whitney U test, Min: Minimum, Max: Maximum.

Table 2. Descriptive statistics for qualitative independent variables.

Variables	Categories of Variables	Status (Number (%))		*p*
		No	Yes	
Smoke	Current	22 (22)	61 (61)	**<0.001***
	Ex	14 (14)	26 (26)	
	Non-Smoker	64 (64)	13 (13)	
Hypertensive condition	Absent	83(83)	67 (67)	**0.022***
	Mild	14 (14)	23 (23)	
	Moderate	3 (3)	10 (10)	
Family history of angina	No	94 (94)	85 (85)	0.065**
	Yes	6 (6)	15 (15)	
Family history of myocardial infarction	No	88 (88)	47 (47)	**<0.001***
	Yes	12 (12)	53 (53)	
Family history of stroke	No	94 (94)	94 (94)	1**
	Yes	6 (6)	6 (6)	
Diabetes	No	97 (97. 97)	94 (94.9)	0.445***
	Yes	2 (2.02)	5 (5.1)	

*: Pearson chi-square test, **: Continuity Correction test, ***: Fisher's Exact test.

Table 3. Performance measure values derived from the Bagged CART model.

Performance Metrics	Value (%)
ACC	98.5
BACC	98.5
Sen	99.0
Spe	98.0
PPV	98.0
NPV	99.0
F1-score	98.5

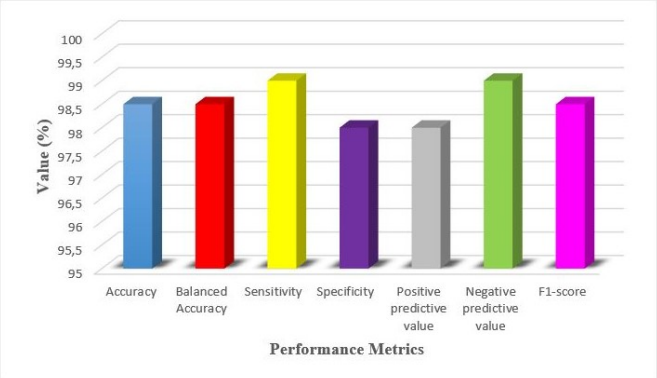

Figure 1. Performance metrics obtained from the Bagged CART method.

Table 4. The variable importance values derived from the Bagged CART method.

Variables Importance	Values
Age	100
The average number of cigarettes smoked per day	70.726
Family history of myocardial infarction/yes	55.291
Smoking status/current	53.518
Family history of angina/yes	15.883
Hypertensive condition/ moderate	15.835
Smoking status/ex	15.227
Hypertensive condition/ mild	12.366
Family history of stroke/yes	10.024
Whether the woman has diabetes/yes	8.344

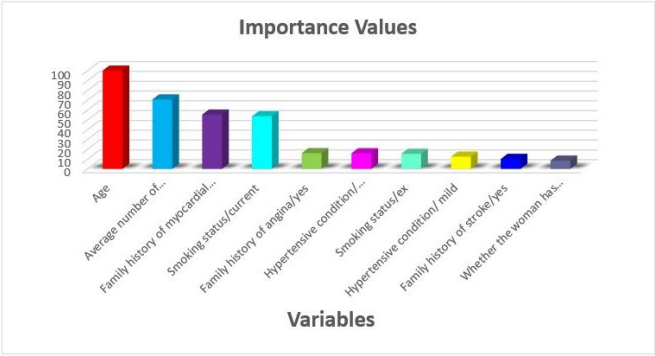

Figure 2. The importance values of the variables are determined by the Bagged CART model.

given in Table 1. Table 1 reveals that the target variable groups differ statistically significantly in age and cigarette variables ($p<0.05$).

The statistical analysis results of the qualitative independent variables in terms of the target variable (status) are given in Table 2. Table 2 reveals a statistically significant relationship between the variables smoking, hypertension, and family history of myocardial infarction and the target variable groups ($p<0.05$).

The results of the performance measures produced from modeling with Bagged CART are shown in Table 3.

As a consequence of the modeling, the ACC, BACC, Sen, Spe, PPV, NPV, and F1-score derived from the Bagged CART model were 98.5%, 98.5%, 99.0%, 98.0%, 98.0%, 99.0%, and 98.5%, respectively.

Figure 1 depicts the results of the Bagged CART model's performance measures.

Table 4 shows the variable importance values computed as a consequence of the Bagged CART method.

The graph of the important values of the variables acquired as a consequence of the Bagged CART model is shown in Figure 2.

Discussion

IHD is the leading cause of death in Western countries. Over the previous 30 years, significant progress has been made in this field. Contrary to common assumption, more women than men die from coronary artery disease in the United States and Europe, despite the fact that the incidence of mortality from IHD has steadily decreased in men and stayed rather steady in women [19, 20].

IHD is a primary cause of morbidity, in addition to being a vital contributor to death. Angina is a relatively prevalent illness, affecting roughly 5% of the general population in Western nations, affecting lifestyle quality and capacity to work, as well as having major economic effects. In recent years, the occurrence of this form of IHD has not decreased but has even grown. Once more, women perform worse than males in this area [19]. With 13.331 cases of angina in women and 11,511 cases in men, a recent meta-analysis of cross-sectional and prospective studies from 31 countries worldwide consistently showed a significant female excess [21].

Women with AP differ significantly from males with this condition. Compared to men, women with IHD are older, have more comorbid conditions, and have a more functional impairment [22]. In addition, IHD in women is both underdiagnosed and undertreated. Early identification is vital for effective treatment [23]. Traditional diagnostic approaches may be insufficient for females. Much research on clinical decision-support systems has been conducted to tackle mentioned challenges by applying various ML approaches [24]. Data processing using ML classifiers may play an important role in predicting CVD [25]. Several studies have been undertaken recently for this goal. All of the mentioned research disclosed that adopting computerized medical decision-support systems can help clinicians make accurate and fast diagnoses of patients. As a result, an effort is made to bridge expert knowledge and expertise to construct a system that equitably supports the diagnostic process [24].

The purpose of this research was to forecast AP in women and to identify the risk variables that may be related to AP in women by using the Bagged CART approach to an open-access dataset of exclusively female patients with AP. Among the performance metrics obtained from the Bagged CART result, ACC, BACC, Sen, Spe, PPV, NPV, and F1-score were obtained as 98.5%, 98.5%, 99.0%, 98.0%, 98.0%, 99.0%, and 98.5%, respectively. The modeling produced successful results in the diagnosis of AP in women. The variables most associated with the AP diagnosis were age, the average number of cigarettes smoked per day, family history of myocardial infarction: yes, smoking status: current, family history of angina: yes, hypertensive condition: moderate, smoking status: ex, hypertensive condition: mild, family history of stroke: yes, whether the woman has diabetes: yes, respectively.

A study has shown that hypertension is a risk factor for AP [26]. This result supports the result of the current study. In another study, patients with AP hypertension, chronic kidney disease, and diabetes were more prevalent among women over 65 years old than among males of the same age. Men and women under the age of 65 were more likely to have a family history of coronary heart disease, and to have a higher body mass index [27]. These results support the conclusion in the present study that having a family history of myocardial infarction, diabetes, age, and hypertension are risk factors for AP. Consistent with the present study, another study showed that positive family history and smoking increase the risk of AP [28].

In another study using the same data set, J48 and RF models from decision trees models were used. RF model gave better results in the study. Among the performance metrics obtained from the RF result, ACC, BACC, Sen, Spe, PPV, NPV, and F1-score were obtained as 92.1%, 92.1%, 89.5%, 94.7%, 94.4%, 90.0%, and 91.9%, respectively. Also, the variable importance was absent in this research [29]. However, the variable importance was added in the current investigation, and the factors related to the disease were evaluated according to their importance.

Conclusion

As a result of the study's findings, Bagged CART model has been found to be successful in predicting AP in women. Furthermore, the risk factors for AP in women were evaluated and their important values were mentioned in this study. With this successful classification performance, it will help in the early diagnosis of the disease for effective treatment and will benefit from taking necessary precautions in the early stages.

Ethical approval

Ethics committee approval is not required in this study.

References

1. Mitka M. Heart disease a global health threat. Jama. 2004;291(21):2533-.
2. Şahin B, İlgün G. Risk factors of deaths related to cardiovascular diseases in World Health Organization (WHO) member countries. Health & Social Care in the Community. 2022;30(1):73-80.
3. Crea F, Camici PG, De Caterina R, Lanza GA. Chronic ischaemic heart disease. In: Camm AJ, Luescher TF, Serruys PW, Eds. The ESC Textbook of Cardiovascular Medicine, Blackwell Publishing Ltd., Oxford, 2006; 391-424.
4. Kim HW, Klem I, Kim RJ. Detection of myocardial ischemia by stress perfusion cardiovascular magnetic resonance. Magnetic resonance imaging clinics of North America. 2007;15(4):527-40.
5. Diamond GA. A clinically relevant classification of chest discomfort. Journal of the American College of Cardiology. 1983;1(2 Part 1):574-5.
6. Timmis AD, Feder G, Hemingway H. Prognosis of stable angina pectoris: why we need larger population studies with higher endpoint resolution. Heart. 2007;93(7):786-91.
7. Möller-Leimkühler AM. Gender differences in cardiovascular disease and comorbid depression. Dialogues in clinical neuroscience. 2022;71-83.

8. Wenger NK. Angina in women. Current cardiology reports. 2010;12(4):307-14.
9. Thabtah F. Machine learning in autistic spectrum disorder behavioral research: A review and ways forward. Informatics for Health and Social Care. 2019;44(3):278-97.
10. Muhammad L, Algehyne EA, Usman SS. Predictive supervised machine learning models for diabetes mellitus. SN Computer Science. 2020;1(5):1-10.
11. Muhammad L, Al-Shourbaji I, Haruna AA, Mohammed IA, Ahmad A, Jibrin MB. Machine learning predictive models for coronary artery disease. SN Computer Science. 2021;2(5):1-11.
12. Hamze-Ziabari S, Bakhshpoori T. Improving the prediction of ground motion parameters based on an efficient bagging ensemble model of M5´ and CART algorithms. Applied Soft Computing. 2018;68:147-61.
13. Deng H, Diao Y, Wu W, Zhang J, Ma M, Zhong X. A high-speed D-CART online fault diagnosis algorithm for rotor systems. Applied Intelligence. 2020;50(1):29-41.
14. Choubin B, Abdolshahnejad M, Moradi E, Querol X, Mosavi A, Shamshirband S, et al. Spatial hazard assessment of the PM10 using machine learning models in Barcelona, Spain. Science of The Total Environment. 2020;701:134474.
15. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees: Chapman & Hall; 1984;358.
16. Timofeev R. Classification and regression trees (CART) theory and applications. Humboldt University, Berlin. 2004;54.
17. Murphree DH, Arabmakki E, Ngufor C, Storlie CB, McCoy RG. Stacked classifiers for individualized prediction of glycemic control following initiation of metformin therapy in type 2 diabetes. Computers in biology and medicine. 2018;103:109-15.
18. Duan H, Deng Z, Deng F, Wang D. Assessment of groundwater potential based on multicriteria decision making model and decision tree algorithms. Mathematical Problems in Engineering. 2016;2016.
19. Rosamond W, Flegal K, Friday G, Furie K, Go A, Greenlund K, et al. Heart disease and stroke statistics—2007 update: a report from the American Heart Association Statistics Committee and Stroke Statistics Subcommittee. Circulation. 2007;115(5):69-171.
20. Allender S, Scarborough P, Peto V, Rayner M, Leal J, Luengo-Fernandez R., & Gray A. European cardiovascular disease statistics. European Heart Network. 2008;3:11-35.
21. Hemingway H, Langenberg C, Damant J, Frost C, Pyörälä K, Barrett-Connor E. Prevalence of angina in women versus men: a systematic review and meta-analysis of international variations across 31 countries. Circulation. 2008;117(12):1526-36.
22. Bairey Merz CN, Shaw LJ, Reis SE, Bittner V, Kelsey SF, Olson M, et al. Insights from the NHLBI-Sponsored Women's Ischemia Syndrome Evaluation (WISE) Study: Part II: gender differences in presentation, diagnosis, and outcome with regard to gender-based pathophysiology of atherosclerosis and macrovascular and microvascular coronary disease. Journal of the American College of Cardiology. 2006;47(3S):21-9.
23. Zuchi C, Tritto I, Ambrosio G. Angina pectoris in women: focus on microvascular disease. International journal of cardiology. 2013;163(2):132-40.
24. Absar N, Das EK, Shoma SN, Khandaker MU, Miraz MH, Faruque MRI, Tamam N, Sulieman A, Pathan RK. The Efficacy of Machine-Learning-Supported Smart System for Heart Disease Prediction. Healthcare (Basel). 2022;10(6):1137.
25. Mohan S, Thirumalai C, Srivastava G. Effective heart disease prediction using hybrid machine learning techniques. IEEE access. 2019;7:81542-54.
26. Xue X, Liu Y, Yang M, Wang S, Huang M, Gao S, Xu Y, Gao S, Li L, Yu C. Effect of hypercholesterolemia alone or combined with hypertension on the degree of coronary artery stenosis in patients with coronary heart disease angina pectoris: A medical records based retrospective study protocol. Medicine (Baltimore). 2020;99(38): e22225.
27. Xu M, Li H-W, Chen H, Guo C-Y. Sex and age differences in patients with unstable angina pectoris: a single-center retrospective study. The American Journal of the Medical Sciences. 2020;360(3):268-78.
28. Merry AH, Boer J, Schouten LJ, Feskens EJ, Verschuren W, Gorgels AP, et al. Smoking, alcohol consumption, physical activity, and family history and the risks of acute myocardial infarction and unstable angina pectoris: a prospective cohort study. BMC cardiovascular disorders. 2011;11(1):1-14.
29. Çiçek İB, Küçükakçali Z, Güldoğan E. Comparison Of Different Decision Tree Models In Classification Of Angina Pectoris Disease. The Journal Of Cognitive Systems.5(2):74-7.