Ann Med Res

Current issue list available at AnnMedRes

Annals of Medical Research

journal page: www.annalsmedres.org

Machine learning-based forecasting of coronary artery disease risk factors and diagnostic insights

Zekeriya Dogan[a], Ipek Balikci Cicek[b,*]

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

ARTICLE INFO

Keywords:
Coronary artery disease
Classification
Tree-based machine learning
Risk factor

Received: Aug 18, 2023
Accepted: Sep 13, 2023
Available Online: 27.09.2023

DOI:
10.5455/annalsmedres.2023.08.204

Abstract

Aim: As people's quality of life and habits have changed, Coronary Artery Disease (CAD) has become the leading cause of death globally. It is a complicated cardiac disease with various risk factors and a wide range of symptoms. An early and accurate diagnosis of CAD allows for the quick administration of appropriate treatment, which contributes to a decreased mortality rate. Machine learning (ML) algorithms for CAD prediction and treatment decisions are quickly being developed and implemented in clinical practice. Predictive models based on machine learning algorithms may aid health personnel in the early diagnosis of CAD, lowering mortality. Thus, this study goal is to forecast the elements that may be connected with CAD using tree-based approaches, which are one of the machine learning methods, and to discover which factor is more effective on CAD.

Materials and Methods: The open-access heart disease dataset was used within the scope of the study to investigate the risk factors related with CAD. The data set used contains the values of 333 patients, as well as 20 input and 1 target variables. The 10-fold cross validation approach was employed in the modeling, and the data set was divided as 80%: 20% as training and test datasets. For model assessment, the measures of accuracy (ACC), balanced accuracy (b-ACC), sensitivity (SE), specificity (SP), positive predictive value (ppv), negative predictive value (npv), and F1-score were utilized.

Results: The values of ACC, b-ACC, SE, SP, ppv, npv, and F1-score performance metrics were 9 98.5%, 98.8%, 97.7%, 100%, 100%, 95.8% and 98.8%, respectively, as a consequence of the estimate model results created with the XGBoost approach, which has the best performance among tree-based models. When the groups with or without CAD were compared, a statistically significant difference was found in terms of the age variable. There is also a significant relationship between the active, lifestyle, ihd, dm, ecgpatt, qwave variables and the presence/absence of the CAD variable. When the variable significance values obtained as a result of modeling with the highest performing XGBoost are examined, it is seen that the variables that most associated with CAD are ekgpatt: normal, ekgpatt: ST-depression, ekgpatt: T-inversion, qwave: yes, age, bpdias, height, LDL, HR, IVSD: with LVH, bpsyDM.

Conclusion: According to the performance criteria of the forecasting models used, CAD gave distinctively successful results in forecasting. By identifying risk factors associated with CAD, the proposed machine learning models can provide clinicians with practical, cost-effective and beneficial assistance in making accurate predictive decisions.

Introduction

Heart disease is the leading cause of death in both men and women across the world, accounting for more than half of all male fatalities. Heart illness encompasses a wide range of cardiac problems. The most common type of heart disease is CAD, which can lead to heart attacks, which kill about 370.000 people each year [1,2].

CAD is a kind of heart disease characterized by a narrowing or blockage of the coronary arteries, which transport blood and oxygen to the heart muscle. CAD is often associated with a process known as atherosclerosis. Atherosclerosis is a condition characterized by the accumulation of fat and cholesterol inside blood vessels. These deposits form plaques on the walls of the arteries and can cause narrowing or blockage of the arteries. As a result, decreased blood flow to the heart can lead to serious complications such as heart attack and heart failure [3,4]. According to

*Corresponding author:
Email address: ipek.balikci@inonu.edu.tr (Ipek Balikci Cicek)

World Health Organization (WHO) statistics, around 17.9 million people die each year as a result of CAD. According to WHO data, whereas CAD accounts for 38% of all fatalities worldwide, it ranks first in Turkey with 44% [5]. As a result, CAD is one of the leading causes of morbidity and death in both developed and developing nations [6].

Three major risk factors for heart disease are smoking, high cholesterol, and high blood pressure. Some other medical conditions and lifestyle choices, such as obesity, physical inactivity, diabetes, excessive alcohol use, and poor diet can also increase the risk of heart disease [7].

Although various methods are used for the diagnosis of the disease, the method to be used varies according to the patient's symptoms, risk factors and various medical conditions. To estimate the severity of heart illness in patients, four fundamental techniques are being employed. Among these are chest X-rays, coronary angiograms, electrocardiograms (ECG), and exercise stress testing [8]. Diagnosing heart disease in the early stages is crucial to saving patients' lives. Early identification of coronary heart disease assists physicians in selecting the best course of therapy and increases patients' chances of survival. However, specialists are not generally available in many impoverished nations and places to undertake these diagnostic tests. Furthermore, in many circumstances, incorrect diagnoses and improperly performed medical procedures might threaten a patient's health and increase the economic burden on society [9]. As a result, accurate and early diagnosis of heart disease has become critical in enhancing patients' prospects of long-term survival.

The diagnosis of CAD is challenging; nevertheless, computer assisted detection has been created to predict heart disease in individuals automatically. ML, as one of the most modern computer-aided detection technologies, is an emerging technique for analyzing medical data and making predictions about early detection discoveries. ML is an artificial intelligence (AI) technique that is used to develop models or systems that can learn from existing datasets to predict future events [10]. ML discovers important information and identifies hidden patterns in massive data warehouses automatically [11].

Therefore, in this study, various tree-based machine learning models were applied to predict the presence and absence of CAD in patients. In addition, this study aimed to compare the performance of Stochastic gradient boosting (SGB), XGBoost and Bagged CART machine learning methods used in CAD estimation and to determine the risk factors associated with CAD.

Materials and Methods

Dataset and variables

The Erbil Heart Disease Dataset used in this study is open access and was collected from a private hospital in Erbil, Iraq, in order to predict patients with and without CAD. The dataset includes some demographic information about the patients, some physical examination data of the patients, and medical laboratory tests [12]. Table 1 displays all twenty one variables in the dataset, along with their related datatypes and brief descriptions.

Biostatistics analysis phase

In the present study, Kolmogorov-Smirnov test was used to examine the conformity of numerical variables to normal distribution. Data that were not normally distributed were summarized as the median (minimum-maximum). Normally distributed data are given as mean ± standard deviation. Mann Whitney U test, Independent sample t-test, Pearson chi-square test and Yates' corrected correction chi-square test were used to compare the data in terms of groups in biostatistical analyses. In statistical analyses, $p<0.05$ was considered statistically significant in all comparisons. All analyzes were performed using IBM SPSS Statistics 26.0 for Windows (New York; USA).

Modelling

In the modeling phase, tree-based SGB, XGBoost and Bagged CART methods were used to predict and classify patients with and without CAD. These machine learning methods used to build predictive models are briefly explained. SGB is a strategy that combines bagging and boosting. To begin, rather than using the entire data set to do the boosting, a random sample of the data is picked at each stage of the procedure. Second, boosting employs deviance as a proxy for misclassification rates, with the gradient determined by this proxy being the steepest gradient approach. Finally, rather than completely formed classification trees being produced at each phase of the boosting procedure, relatively small trees with 6 terminal nodes are constructed [13]. As with earlier ensemble procedures, no larger trees are constructed; instead, each tree created throughout the process—often between 100 and 200 trees—is totaled up, and each observation is classified according to the categorization that appears most frequently in the trees [14].

Gradient Boosting is a useful machine learning strategy for regression and classification issues in which weak prediction models commonly yield decision tree ensembles. Gradient Boost, which is based on the boosting approach, tries to build several weak learners sequentially and integrate them into a complicated model [15]. Extreme Gradient Boosting (XGBoost) is a supervised learning approach that uses gradient boosting machines. Its foundation is built on gradient boosting and decision tree methods. XGBoost offers a substantial speed and performance boost over other algorithms. XGBoost is ten times quicker than previous algorithms, highly performance predictive, and includes a number of regularizations that enhance performance overall while lowering overfitting and over-learning [16].

CART (Classification and Regression Trees) is a non-parametric decision tree logging method extensively used as a classifier. It uses binary trees as its algorithm [17]. Bagging is a popular ensemble strategy for improving decision tree prediction accuracy. To ensure the high quality of the CART model, the bagging strategy was used. Each classifier in this technique generates and stores its model by categorizing a subset of the data. Finally, depending on vote intention among these categories, the class with the most votes is picked as the final classifier [18].

The dataset will be divided into k subsets, with one of the k subsets serving as the test set and the remaining k−1 subsets acting as the training set. As a consequence, each data point appears exactly once in the test set and k−1 times in the training set. The results of the k folds will be averaged to create a single guess. Because it is the most commonly used standard value in research, k=10 was picked [19].

The performance of the model was assessed using ACC, b-ACC, SE, SP, ppv, npv, and F1-score performance measurements.

Results

The dataset used in this study consists of a total of 333 people, 118 (35.4%) with CAD and 215 (64.6%) without CAD. There were 178 (53.5%) female patients and 155 (46.5%) male patients. The mean age of all patients is 55.12±14.159 years and the median age is 57 (20-90).

Biostatistical analysis results

It was determined whether there was a statistically significant difference between the two groups in terms of the presence/absence of CAD, which is the target variable of the quantitative independent variables. When the analysis results were examined, a difference was found between the two groups in terms of age variable ($p<0.05$). There was no significant difference between the two groups in terms of years, LDL, height, weight, HR, BPSYS, BPDIAS variables ($p>0.05$). Table 2 shows the results of the statistical analyses of the quantitative independent variables in relation to the target variable.

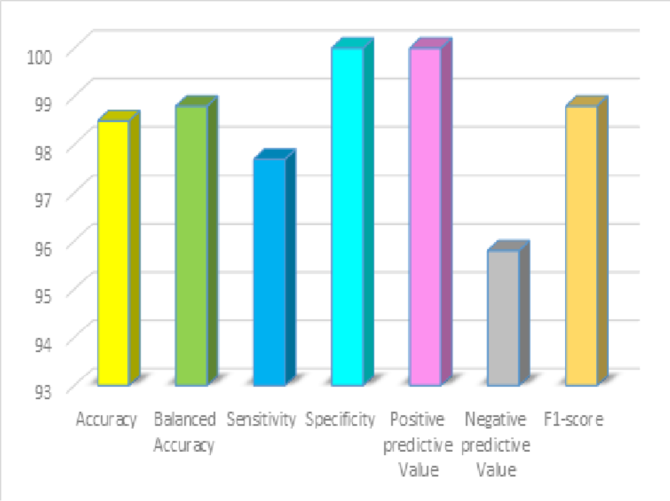

Figure 1. Performance metrics graph for the XGBoost model.

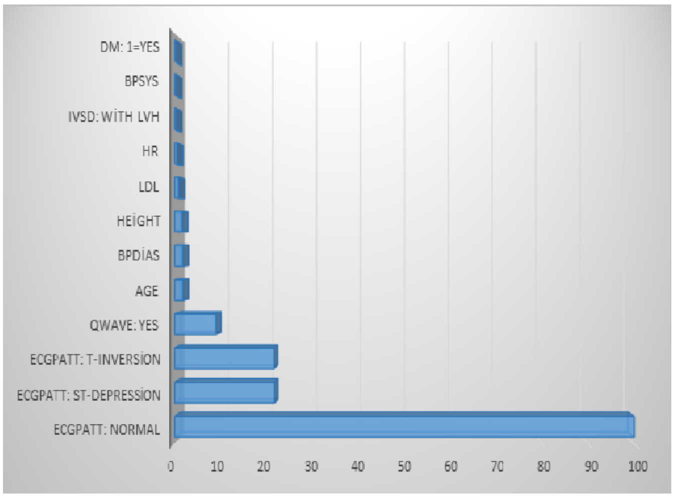

Figure 2. Variable Importance Graph for XGBoost model.

When the groups with or without CAD were compared, a statistically significant difference was found in terms of the age variable. There is also a significant relationship between the active, lifestyle, ihd, dm, ecgpatt, qwave variables and the presence/absence of the CAD variable. However, there is no significant relationship between sex, smoke, chp, fh, htn, ivsd variables and presence/absence of CAD variable ($p>0.05$). Table 3 shows the statistical analysis results of the qualitative independent variables according to the target variable.

Modelling results

Table 3 displays performance measures of tree-based machine learning prediction models used to predict the presence and absence of coronary heart disease using test datasets.

As seen in Table 4, the test results of the model accuracy of the SGB, Bagged CART and XGBoost prediction models were as follows: 97.2% for SGB, 95.5% for Bagged CART, 98.5% for XGBoost. In addition, b-ACC, SE, SP, ppv,npv test results for the XGBoost model showing the best accuracy performance are 98.5%, 98.8%, 97.7% 100%, 100%, 95.8%, 98.8%, respectively. The results of the performance measures of the XGBoost model are given in the Figure 1.

The results of the variable importance value obtained from the XGBoost estimation model, which gave the highest accuracy result as a result of the modeling, are given in Table 5. Figure 2 depicts a graph of variable importance values derived from the XGBoost model.

Discussion

For the past 15 years, cardiovascular diseases have been the leading cause of death globally. In 2019, 17.9 million people are estimated to die from cardiovascular diseases, accounting for 32% of all fatalities worldwide. The global prevalence of cardiovascular disease is increasing at an alarming rate, with yearly deaths expected to exceed 20 million by 2030. The most common kind of cardiovascular disease is CAD. It is the top cause of death worldwide. According to WHO figures, around 17.7 million people died from CAD in 2015, accounting for 31% of total mortality. Many of these CAD deaths may have been averted if CAD was correctly identified and treated quickly.

The preventability of heart and cardiovascular illnesses is the most effective factor in reducing these fatalities worldwide. The development of health care has enabled studies such as identifying the variables underlying disease and avoiding its occurrence. Although recent research has identified risk factors for heart disease, many researchers believe that further study is required to apply this information to minimize the occurrence of heart disease. Heart

Table 1. The names, types, and explanations of dataset* variables.

Variable	Description	Type
Age	Age of patients (in years)	Continuous
Sex	Patient's gender (1= female, 0= male)	Categorical
Smoke	Whether the patient smokes (0=No, 1=Yes)	Categorical
Years	If a smoker, the number of years they have smoked.	Continuous
Idl	The patient's LDL-Cholesterol ratio.	Continuous
Chp	Type of chest pain (1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic).	Categorical
Height	The patient's height in cm.	Continuous
Weight	The patients' weight in kg.	Continuous
Fh	A family history of heart disease.	Categorical
Active	Whether or if the patient is active (0=No, 1=Yes).	Categorical
Lifestyle	The location of living (1=City, 2=Town, 3=Village).	Categorical
Ihd	Is the patient undergoing any cardiac catheterization or heart intervention? (0=No, 1=Yes).	Categorical
Hr	Heart Rate ratio.	Continuous
Dm	Is the patient diabetic ? (0=No, 1=Yes)	Continuous
Bpsys	The Systolic Blood Pressure Ratio.	Continuous
Bpdias	The blood pressure diastolic ratio.	Categorical
Htn	Is the patient hypertensive? (0=No, 1=Yes).	Categorical
Ivsd	IVSD is a measurement used to diagnose Left Ventricular Hypertrophy (LVH).	Categorical
Ecgpatt	Contains four ECG test categories: (1=ST-Elevation, 2=ST-Depression, 3=T-Inversion, 4=Normal).	Categorical
Qwave	The existence or absence of the Q wave (0=No, 1=Yes).	Categorical
Target	If the patient has heart disease or not (0=no heart disease, 1=has heart disease).	Categorical

*: The Erbil Heart Disease Dataset was collected from a private hospital in Erbil, Iraq, in order to predict patients with and without CAD.

Table 2. The statistical analysis of the quantitative independent variables.

Variable	Target				p
	Without CAD		With CAD		
	Mean± Standard deviation	Median(min-max)	Mean± Standard deviation	Median(min-max)	
Age	52.12±13.93	52(23-90)	60.58±12.94	62(20-87)	**<0.001***
Years	4.31±10.55	0(0-50)	5.69±12.43	0(0-50)	0.344*
LDL	114.47±35.33	112(29-213)	110.11±42.38	106.5(26-260)	0.204**
Height	162.85±10.9	163(137-192)	160.74±11.91	160(128-188)	0.115*
Weight	83.12±15.33	81(47-134)	80.4±15.41	80(41-119)	0.359*
HR	83.99±13.53	84(54-130)	83.69±16.51	84(40-140)	0.988*
BPSYS	122.88±21.25	120(80-200)	124.96±21.53	120(90-220)	0.515*
BPDIAS	75.14±12.27	70(40-120)	74.41±13.45	70(45-140)	0.419*

HR: Heart rate ratio; BPSYS: The systolic blood pressure ratio; BPDIAS: The blood pressure diastolic ratio; *:Mann-Whitney U test, **:Independent sample t-test.

disease can be caused by a variety of factors. According to some research, decreasing these risk factors for heart disease may actually help avoid heart disease [20]. Furthermore, while the CAD mortality rate is high, the chances of survival are increased if the diagnosis is established early enough.

For these reasons, healthcare practitioners employ early illness detection and therapy or intervention options. Several approaches for early detection of coronary artery disease have been presented in recent decades. Many experts from many scientific fields have suggested that AI, which may enhance diagnostic accuracy, is one of the most successful options for the early identification of illnesses, which has been increasingly widespread in recent years. Human intelligence is used by AI to discover data correlations and emulate human problem-solving behaviors [21-23]. ML, which have played a critical role in illness diagnosis, are an important area of AI. ML, a subset of AI, has several uses in detecting and forecasting various illnesses. ML approaches have been employed successfully in a variety of health research domains [24]. ML algorithms are a standard technique in knowledge exploration, used to develop high-accuracy prediction and diagnostic models. By monitoring and detecting links between data, ML classification systems learn to interpret it. Monitoring via learning from labeled data enables us to detect previously unseen data output. This is a common ML approach in which a system is given a list of input-output pairs and the system attempts to identify a function from input to output [25,26].

Therefore, this study aimed to determine the tree-based machine learning method that is effective in predicting the

Table 3. The statistical analysis of the qualitative independent variables.

Variables		Target		p
		without CAD	with CAD	
Sex	Male	99 (46.05)	56 (47.46)	0.805*
	Female	116 (53.95)	62 (52.54)	
Smoke	No	176 (81.86)	92 (77.97)	0.391*
	Yes	39 (18.14)	26 (22.03)	
CHPFH	Typical Angina	26 (12.09)	21 (17.80)	0.433*
	Atypical Angina	40 (18.60)	17 (14.41)	
	Non-Anginal Pain	77 (35.81)	39 (33.05)	
	Asymptomatic	72 (33.49)	41 (34.75)	
FH	No	168 (78.14)	86 (72.88)	0.281*
	Yes	47 (21.86)	32 (27.12)	
Active	No	120 (55.81)	90 (76.27)	**<0.001***
	Yes	95 (44.19)	28 (23.73)	
Lifestyle	City	97 (45.12)	73 (61.86)	**0.007***
	Town	88 (40.93)	29 (24.58)	
	Village	30 (13.95)	16 (13.56)	
IHD	No	175 (81.40)	73 (61.86)	**<0.001***
	Yes	40 (18.60)	45 (38.14)	
DM	No	173 (80.47)	83 (70.34)	**0.036***
	Yes	42 (19.53)	35 (29.66)	
Htn	No	106 (49.30)	55 (46.61)	0.638*
	Yes	109 (50.70)	63 (53.39)	
IVSD	No	152 (70.70)	87 (73.73)	0.557*
	Yes	63 (29.30)	31 (26.27)	
Ecgpatt	ST-Elevation	8 (3.72)	15 (12.71)	**<0.001***
	ST-Depression	3 (1.40)	52 (44.07)	
	T-Inversion	1 (0.47)	48 (40.68)	
	Normal	203 (94.42)	3 (2.54)	
Qwave	No	215 (100.00)	92 (77.97)	**<0.001****
	Yes	0 (0.00)	26 (22.03)	

*:Pearson chi-square test,**: ***: Yates' Correction chi-square test.

Table 4. Metrics of the performance of the ML models used on the Erbil Heart Disease Dataset.

Model	ACC	b-ACC	SE	SP	ppv	npv	F1 score
SGB	0.972	0.979	1	0.957	0.926	1	0.962
Bagged CART	0.955	0.945	0.977	0.913	0.955	0.955	0.966
XG Boost	**0.985**	**0.988**	**0.977**	**1**	**1**	**0.958**	**0.988**

The values in boldface show the best performance group.

Table 5. The results of the variable importance value from XGBoost model.

Variables	Importance value
Ecgpatt: Normal	100
Ecgpatt: ST-Depression	21.755
Ecgpatt: T-Inversion	21.681
Qwave: Yes	9.035
Age	1.837
Bpdias	1.814
Height	1.653
LDL	0.881
HR	0.546
IVSD: with LVH	0.319
Bpsys	0.293
Dm: 1=Yes	0.289

presence and absence of patients with CAD based on data including patient demographics, patient history, physical examination and symptomatic features, medical laboratory tests, and diagnostic features . For this purpose, SGB, XGBoost and Bagged CART algorithms were evaluated and compared. In addition, determining the risk factors that may be associated with CAD as a result of the evaluations of the prediction models is another aim of the study.

Considering the results of the statistical analysis, there was a statistically significant difference between the without CAD and with CAD groups with the age variable, but no statistically significant difference was observed in the other variables. Also without CAD and with CAD groups and active, lifestyle, ihd, dm, ecgpatt, qwave variables, while there was no statistically significant relationship between sex, smoke, chp, fh, htn, ivsd variables.

As a result of the estimation model findings made with the XGBoost method, which has the best performance among tree-based models, the values of ACC, b-ACC, SE, SP, ppv, npv and F1-score performance metrics were 98.5%, 98.8%, 97.7%, 100%, 100%, 95.8% and 98.8%. In addition, when the variable importance values obtained as a result of modeling with the highest performing XGBoost were examined, it was seen that the important factors associated with CAD were ekgpatt: normal, ekgpatt: ST-depression, ekgpatt: T-inversion, qwave: yes, age, bpdias, height, LDL, HR, IVSD: with LVH, bpsys, DM: 1=yes.

ML method has been widely used in the literature for CAD detection. ML method has been widely used in the literature for CAD detection. Some of these studies are briefly described below.

Kurt and et al., investigated five different algorithms for the prediction of CAD: Classification and Regression Tree, Logistic Regression, Radial Bias Function, Artificial Neural Network and Self Organising Feature Maps. The algorithms were tested on a dataset that included 1245 patient records and various predictor parameters such as age, gender, BMI, and so on. The results of the tests demonstrated that the Neural Network outperformed the other algorithms [27].

An another study, several classification algorithms were used to the Z-Alzadeh Sani data, which comprised 303 patients from the medical and research center by Alizadehsani et al. The data collection contains 216 people with CAD, with the remaining patients being disease-free, as well as 54 features. The researcher also created a feature selection algorithm that used Naive Bayes, Sequential Minimal Optimisation (SMO), Bagging with SMO, and Neural

Network to produce three different features. The accuracy of SMO reached the highest value of 94.08% when assessed using the three established criteria, according to the test results [28].

In 2015, Akila et al., created a hybrid model for the identification and prediction of CAD, which was validated using patient data from occupational drivers obtained from a medical college and hospital. The model was split into two halves. The decision tree algorithm was used in the first step to detect danger by categorizing physical and biological factors. The second involved analyzing CAD risk flagged cases using decision tree with Multi-Layer Perceptron employing habitation and medical history variables. Decision tree and Multi-Layer Perceptron have classification accuracy of 98.66% and 96.66%, respectively [29].

Using 10-fold cross-validation on the dataset, Nassif et al. (2018) conducted a study to examine the effectiveness of three machine learning algorithms for the prediction of CAD. The data set was subjected to three different feature selection techniques, feature versus risk score graphs were produced to find the traits that are most closely connected to the risk of CAD, and the top seven features were selected as input variables for the algorithms. The Naive Bayes Algorithm outperformed all others in the study, obtaining an accuracy score of 84% [30].

According to the predictive model performance success obtained in this study, a machine learning-assisted risk classification approach for the diagnosis of CAD appears to have predictive value in terms of classifying patients and guiding further studies.

Conclusion

In conclusion, recent advances in AI and ML have facilitated the identification of individuals at high risk for disease appearance, and they are used in the current study to to design a methodology for risk prediction of CAD occurrence based on several risk factors. We attempted to comprehend and study the relationship of characteristics with CAD using data analysis, and we discovered hidden patterns involving CAD-related symptoms. Thus, the proposed method might aid healthcare providers and clinical specialists in their attempts to avert devastating CAD accumulations in both individual patients and populations. Furthermore, by monitoring and analyzing risk variables, individualized guidelines and actions to avoid CAD may be proposed.

Ethical approval

Ethics committee approval is not required in this study.

References

1. Heron M. Deaths: leading causes for 2008. National Vital Statistics Reports: From the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System. 2012;60(6):1-94.
2. Miao KH, Miao JH, Miao GJ. Diagnosing coronary heart disease using ensemble machine learning. International Journal of Advanced Computer Science and Applications. 2016;7(10).
3. Malakar AK, Choudhury D, Halder B, Paul P, Uddin A, Chakraborty S. A review on coronary artery disease, its risk factors, and therapeutics. Journal of cellular physiology. 2019;234(10):16812-23.
4. Dipto IC, Islam T, Rahman HM, Rahman MA. Comparison of Different Machine Learning Algorithms for the Prediction of Coronary Artery Disease. Journal of Data Analysis and Information Processing. 2020;8(2):41-68.
5. Virani SS, Alonso A, Benjamin EJ, Bittencourt MS, Callaway CW, Carson AP, et al. Heart disease and stroke statistics—2020 update: a report from the American Heart Association. Circulation. 2020;141(9):e139-e596.
6. Alizadehsani R, Roshanzamir M, Abdar M, Beykikhoshk A, Khosravi A, Panahiazar M, et al. A database for using machine learning and data mining techniques for coronary artery disease diagnosis. Scientific data. 2019;6(1):227.
7. Hajar R. Risk factors for coronary artery disease: historical perspectives. Heart views: the official journal of the Gulf Heart Association. 2017;18(3):109.
8. Alizadehsani R, Zangooei MH, Hosseini MJ, Habibi J, Khosravi A, Roshanzamir M, et al. Coronary artery disease detection using computational intelligence methods. Knowledge-Based Systems. 2016;109:187-97.
9. Mahesh T, Dhilip Kumar V, Vinoth Kumar V, Asghar J, Geman O, Arulkumaran G, et al. AdaBoost ensemble methods using K-fold cross validation for survivability with the early detection of heart disease. Computational Intelligence and Neuroscience. 2022;2022.
10. Muhammad LJ, Algehyne EA, Usman SS. Predictive supervised machine learning models for diabetes mellitus. SN Computer Science. 2020;1(5):240.
11. Muhammad L, Algehyne EA, Usman SS, Ahmad A, Chakraborty C, Mohammed IA. Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset. SN computer science. 2021;2:1-13.
12. [cited 2023 2022]. Available from: https://doi.org/10.34740/KAGGLE/DSV/3989065.
13. 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.
14. Moisen GG, Freeman EA, Blackard JA, Frescino TS, Zimmermann NE, Edwards Jr TC. Predicting tree species presence and basal area in Utah: a comparison of stochastic gradient boosting, generalized additive models, and tree-based methods. Ecological modelling. 2006;199(2):176-87.
15. Wang J, Li P, Ran R, Che Y, Zhou Y. A short-term photovoltaic power prediction model based on the gradient boost decision tree. Applied Sciences. 2018;8(5):689.
16. Salam Patrous Z. Evaluating XGBoost for user classification by using behavioral features extracted from smartphone sensors. 2018.
17. Timofeev R. Classification and regression trees (CART) theory and applications. Humboldt University, Berlin. 2004;54.
18. 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.
19. Sonkar P. Application of supervised machine learning to predict the mortality risk in elderly using biomarkers. 2017.
20. Taşçı ME, Şamlı R. Veri madenciliği ile kalp hastalığı teşhisi. Avrupa Bilim ve Teknoloji Dergisi. 2020:88-95.
21. Finegold JA, Asaria P, Francis DP. Mortality from ischaemic heart disease by country, region, and age: statistics from World Health Organisation and United Nations. International journal of cardiology. 2013;168(2):934-45.
22. Cassar A, Holmes Jr DR, Rihal CS, Gersh BJ, editors. Chronic coronary artery disease: diagnosis and management. Mayo Clinic Proceedings; 2009: Elsevier.
23. Mirbabaie M, Stieglitz S, Frick NR. Artificial intelligence in disease diagnostics: A critical review and classification on the current state of research guiding future direction. Health and Technology. 2021;11(4):693-731.
24. Hamet P, Tremblay J. Artificial intelligence in medicine. Metabolism. 2017;69:S36-S40.
25. Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in medicine. 2001;23(1):89-109.
26. Kumar Y, Koul A, Sisodia PS, Shafi J, Kavita V, Gheisari M, et al. Heart failure detection using quantum-enhanced machine learning and traditional machine learning techniques for internet of artificially intelligent medical things. Wireless Communications and Mobile Computing. 2021;2021:1-16.

27. Kurt I, Ture M, Kurum AT. Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. Expert systems with applications. 2008;34(1):366-74.
28. Alizadehsani R, Habibi J, Hosseini MJ, Mashayekhi H, Boghrati R, Ghandeharioun A, et al. A data mining approach for diagnosis of coronary artery disease. Computer methods and programs in biomedicine. 2013;111(1):52-61.
29. Akila S, Chandramathi S. A hybrid method for coronary heart disease risk prediction using decision tree and multi layer perceptron. Indian Journal of Science and Technology. 2015;8(34):1-7.
30. Nassif AB, Mahdi O, Nasir Q, Talib MA, Azzeh M, editors. Machine learning classifications of coronary artery disease. 2018 International Joint Symposium on artificial intelligence and natural language processing (iSAI-NLP); 2018: IEEE.