Journal of Autonomous Intelligence (2023) Volume 6 Issue 3
doi: 10.32629/jai.v6i3.628

ORIGINAL RESEARCH ARTICLE

A precise coronary artery disease prediction using Boosted C5.0 decision tree model

Surjeet Dalal[1], Umesh Kumar Lilhore[2,*], Sarita Simaiya[2], Vivek Jaglan[3], Anand Mohan[4], Sachin Ahuja[5], Akshat Agrawal[1], Martin Margala[6], Prasun Chakrabarti[7]

[1] *Department of Computer Science and Engineering, Amity University Haryana, Gurugram 122413, Haryana, India*

[2] *Department of Computer Science and Engineering, Chandigarh University, Mohali 140413, India*

[3] *Amity School of Engineering and Technology, Amity University Madhya Pradesh, Gwalior 474020, India*

[4] *Department of Physics, Kunwar Singh College, Darbhanga 846004, India*

[5] *Department of Computer Science and Engineering, Chandigarh University, Mohali1 40413, India*

[6] *School of Computing and Informatics, University of Louisiana USA, Lafayette LA 70504, United States of America*

[7] *ITM SLS Baroda University, Vadodara 391510, India*

*** Corresponding author:** Umesh Kumar Lilhore, umeshlilhore@gmail.com

ABSTRACT

In coronary artery disease, plaque builds up in the arteries that carry oxygen-rich blood to the heart. Having plaque in the arteries can constrict or impede blood flow, leading to a heart attack. Shortness of breath and soreness in the chest are common symptoms. Lifestyle modifications, medication, and potentially surgery are all options for treatment. In coronary artery disease, plaque builds up in the arteries that carry oxygen-rich blood to the heart. Having plaque in the arteries can constrict or impede blood flow, leading to a heart attack. Shortness of breath and soreness in the chest are common symptoms. Lifestyle modifications, medication, and potentially surgery are all options for treatment. This paper presents a Hybrid Boosted C5.0 model to predict coronary artery disease more precisely. A Hybrid Boosted C5.0 model is formed by combining the C5.0 decision tree and boosting methods. Boosting is a supervised machine learning method that leverages numerous inadequate models to construct a more robust and powerful model. The proposed model and some well-known existing machine learning models, i.e., decision tree, AdaBoost, and random forest, were implemented using an online coronary artery disease dataset of 6611 patients and compared based on various performance measuring parameters. Experimental analysis shows that the proposed model achieved an accuracy of 91.62% at training and 81.33% at the testing phase. The AUC value achieved in the training and testing phase is 0.957 and 0.88, respectively. The Gini value achieved in the training and testing phase is 0.914 and 0.759, respectively, far better than the proposed method.

Keywords: machine learning; C5.0 decision tree algorithm; coronary artery disease; prediction; over-fitting; boosting

ARTICLE INFO

Received: 11 May 2023
Accepted: 17 July 2023
Available online: 1 September 2023

1. Introduction

A constriction or blockage of their coronary arteries is caused by an accumulation of fatty material known as plaque, which describes the same condition. To put it another way, atherosclerosis leads to coronary artery disease. Plaque clogs their arteries, resulting in atherosclerosis. Blood clot-forming fibrin is found in plaque, cholesterol, fats, and other waste materials. Their arteries shrink and harden when plaque builds up on their walls[1]. As a result of plaque build-up, their heart's blood flow might be slowed or even halted, causing cardiac arrest. Their heart can't function correctly if it doesn't receive the oxygen and nutrition it requires. This is known as

ischemia. Because of insufficient blood flow to the heart muscle, you may have chest pain or discomfort when exercising (called angina). As a result, it raises their chance of developing heart disease[2].

Heart disease affects everyone, regardless of age or ethnicity. A person's success may vary depending on their unique circumstances. In the beginning, it's common for children to begin the process. The walls of their blood vessels begin to show symptoms of fat at age 10. As their arteries get clogged, their bodies release white blood cells to clear the cholesterol. The assault, on the other hand, exacerbates the inflammation. An additional layer of blood vessels is formed around the plaque to protect it from damage. Breaking apart the plaque's thin top is possible (due to blood pressure or other causes). Platelets, blood cell fragments that cling to the location of "the damage," create a clot. The clot narrows the arteries even further. A blood clot can spontaneously disintegrate. When the clot stops blood flow, the heart is starved of oxygen, resulting in angina[3].

One word used to describe many health disorders that impact the heart's structure and function is "heart disease". Those with coronary heart disease have arteries that can not supply the heart with adequate oxygen-rich blood. Deaths from this disease are at an all-time high in the United States of America. According to the Centres for Disease Control and Prevention, coronary artery disease in the United States is the most prevalent kind of heart disease, which affects 18.2 million individuals[4].

The more prominent coronary arteries on the surface of the heart are affected by coronary artery disease. Coronary microvascular disease, on the other hand, involves the heart's tiny arteries. Women are more likely than males to develop coronary microvascular disease. The kind of coronary heart disease determines the source of the condition. As a result, the heart's major arteries might become partially or entirely blocked by this deposit. Small blood veins in the heart malfunction, causing coronary microvascular disease. A heart-healthy lifestyle can help avoid coronary heart disease in most people. Though two people have the same type of coronary heart disease, their symptoms may differ, even if they both have it. As a result, many persons with coronary heart disease go undiagnosed until they experience chest discomfort, a heart attack, or cardiac arrest, all indicative of impeded blood flow to the heart[5].

The signs and symptoms of coronary artery disease change with the progression of the disease. Damage can exist even if no visible symptoms of it are present. When you push yourself physically, you may experience shortness of breath or chest aches for the first time. These symptoms only appear for a small percentage age of the population on rare occasions. Chest discomfort or an actual heart attack may be the initial symptom for some. According to Park et al.[6], doctors use signs like blood pressure, cholesterol levels, and glucose levels from a blood test to determine if someone has coronary artery disease. This information may be used to predict their 10-year cardiovascular risk-their chances of a heart attack or stroke[6].

There are a variety of tests that may be performed to get more information about your condition, depending on your symptoms[7]:

Coronary calcium testing: A CT scan shows calcium and plaque accumulation between heartbeats. The patients can see the damage known as the hardening of the arteries. Treating with a starting plus aspirin may be unsure in patients with no documented coronary heart disease.

High sensitivity C-reactive protein blood test: This tells you if your inflammation levels are more significant than usual.

Electrocardiogram (EKG or ECG): Relaxed heartbeats can be measured using an electrocardiogram (ECG).

Exercise stress test ("treadmill test"): A treadmill test that measures a person's heart rate while the heart pumps harder.

Echocardiogram: An X-ray of the heart using an ultrasound.

Chest X-ray: In this shot, you can see their lungs and heart.

Cardiac catheterization: Using a catheter, a small tube is introduced into a coronary artery to detect obstructions.

Coronary angioplasty: To widen a constricted artery using an expanding balloon. When an artery narrows, nearly 90% of the time, a stent (a metal scaffold) is used to repair it[8].

In this paper, we provide a machine-learning model for predicting individual cases of coronary artery disease using the decision tree. Most of our time is spent delving into the specifics of each prediction. This research aims to construct a predictive model that can effectively estimate the probability of coronary artery disease by utilizing patient data. Providing an automatic method enables healthcare professionals to make well-informed decisions regarding patient care and treatment alternatives. Also, to enhance patients' quality of life through early identification of coronary artery disease (CAD) and implementation of preventive measures. The main contributions of this paper are as follows:

- The primary aims of utilizing a C5.0 decision tree in detecting coronary artery disease (CAD) encompass identifying the most significant risk factors linked to CAD.
- Using the C5.0 decision tree model, we are trying to develop a very accurate model. Various features from the given prescribed dataset are used for coronary artery disease.
- To explain individual predictions, AUC and Gini values are used. The hybrid C5.0 decision tree algorithm is designed to predict coronary artery disease. In addition, ophthalmologists may benefit from the explanation's clinical relevance.
- The proposed model and some well-known existing machine learning models, i.e., decision tree, AdaBoost, and random forest, were implemented using an online coronary artery disease dataset of 6611 patients and compared based on various performance measuring parameters. Experimental analysis shows that the proposed model achieved an accuracy of 91.62% at training and 81.33% at the testing phase.

The complete article is divided into various sections, which include. Section 2 covers related work, section 3 covers materials and methods, section 4 covers experimental results and analysis and section 5 covers conclusions and future work.

2. Related work

Coronary artery disease significantly contributes to adult morbidity and disability in industrialized countries, leading to various illnesses, impairments and fatalities. This serves as a driving force for investigators to seek a highly effective resolution for the issue mentioned earlier. Several significant contributions can be identified.

Ghosh et al.[9] created AI calculations for the ID of sickness and the guess of mortality risk to decide if such models perform better than traditional factual investigations. Zeroing in on fringe supply route sickness (PAD), patient information was obtained from an imminent, observational analysis of 1755 patients. Both machine-learned models were notably preferred and adjusted over the stepwise strategic relapse models, giving more exact illness and mortality risk gauges. AI approaches can create more precise infection arrangements and expectation models. These instruments might demonstrate clinically valuable for the mechanized distinguishing proof of patients with exceptionally horrible sicknesses for which forceful gambling factor the executives can further develop results.

Goswami et al.[10] examines the chance of foreseeing ACS utilizing AI calculations at the beginning phase using clinical and research center information in patients who introduced chest torment at confirmation of trauma center or short-term patient facility. In light of the component example and channel qualities, the

authors can dissect which highlights have discriminative data while different elements assist with further developing precision by diminishing the clamour of highlights with discriminatory data. The order accuracy is average at 0.81 for non-prepared information, and the authors could choose significant elements of characterization and the highlights without discriminative data yet diminishing the commotion of applicable highlights.

Similarly, in the research by Nakanishi et al.[11], when the authors kill the essential elements picked at the first stage, the less-educational highlights have discriminatory data accomplishing a precision of 81%. The gamble gathering of ACS patients can be chosen before the crisis utilizing expectation calculation, and the accuracy is high. There were a few missing pieces of information connected with foresee ACS in this review. In this way, later on, an extensive companion study without missing information ought to be expected to obtain an eventual outcome as described by Bom et al.[12].

Ramirez et al.[13] conducted the coronary artery disease learning and algorithm development (CADLAD) study to determine the demonstrated execution of cPSTA in surveying individuals with chest pain for coronary angiography, which is associated with coronary artery disease (ANGIO). For obese and older people, standard CAD diagnosis procedures may be less accurate. This examination focuses on these patients. This multi-centre, non-huge gamble study was designed to construct and evaluate machine-learned estimates for surveying the existence of CAD (defined as at least one 70% stenosis or fragmented stream hold 0.80). CADLAD's broad break effects have been exhibited elsewhere. This inquiry focuses on CADLAD's elderly and obese (BMI > 30) members.

According to Collet et al.[14], preceding ANGIO, cPSTA signals were acquired exceptionally slowly. An approval partner was used to aimlessly and tentatively try highlights (numerical and topographic) extracted from the signs. Machine-learned angiography findings were matched with cPSTA data from 513 participants to construct a CAD survey. For preliminary testing, 94 people were used from a separately approved partner. This inquiry focused on those over 65 and those with a BMI of over 30. Resting cPSTA imaging appears to be effective across the board, including in the older and larger subpopulations, despite the limited power of the underlying data.

Howard et al.[15] aimed to study AI to predict major adverse cardiovascular events (MACE) using clinical data and obtained image variables from pressure and rest examinations. A fast SPECT scan on 2619 individuals (48% men, 16 to 65 years old) was performed in succession. The stress-just ML, the stress/rest ML, the master added pressure/distinction scores (SSS, SDS), and the scheduled pressure/ischemic complete perfusion shortfalls (TPD) MACE forecasts were examined by region under the beneficial working quality bend (AUC), and a MACE occurred in 320 patients (12%). Two hundred thirty-five patients (10%) developed MACE in their partners who had never had a MI or CABG.

An AI-based model developed by Kim et al.[16] estimates the number of models on display, and the area under the recipient working trademark (AUROC) was used. One thousand nine hundred fifty-seven individuals were divided into sCAD (n = 1442) and non-sCAD (n = 515) groups according to whether the primary epicardial coronary vein showed 50% stenosis. Eighty-seven research facility markers were used in the forecast model. There were six ideal mixtures (T1, T2, ..., T6), each with a different number of the selected lab markers (ranging from 1 to 6). In every category, 77.47%, 85.21%, 8563%, 85.21%, 85.21%, 85.21%, 85.21%, and 84.65%, respectively, were the most accurate. The AI model and its outcomes ought to be tried in an imminent observational concentration from here on out.

Kawasaki et al.[17] proposed an original innovation in light of PC AI, which dissects natural thoracic signals and surveys discharge division in no time. Ventriculograms acquired during the coronary artery disease learning and algorithm development (CADLAD) study were utilized to decide analytic execution for figuring LVEF. Another goal of the CADLAD multicenter trial was to use machine learning to determine

LVEF to detect clinically significant CAD. Heart phase space tomography analysis (cPSTA) was used to acquire resting-state signal information from people with symptoms resembling CAD before ventriculography.

Vazquez et al.[18], presented a machine-learned model to calculate the LVEF. The highlights used are depictions of stage space transformed into a three-layered image of the heart. The resulting stage space signals were fed into a machine-gained algorithm that examined LVEF. One would consider an LVEF advantage of 50% to be uncommon. A machine-learned method based on 96 participants' ventriculography data and stage-space signals was used to measure LVEF. Tests were conducted on 29 unrelated symptoms. One hundred two people (81.6%) had LVEFs greater than 50%, whereas only 23 people (18.4%) had LVEFs lower than or equal to 50%. Using stage-space tomography, the authors found that the responsiveness was 86%, and the specificity was 66%. According to Sandhu et al.[19], ML strategy promptly recognizes ordinary from strange EF (<50%) in this minor accomplice of subjects. Further, AI and enlistment are expected to build the exactness of the appraisal. Likewise, future examinations will incorporate matching stage signal information with the highest quality level (CMR) estimation of LVEF.

Schwalm et al.[20] used the multi-ethnic study of atherosclerosis (MESA), to see if AI (ML) progress would allow for more developed expectations in the forecast of coronary illness (CHD) and atherosclerotic cardiovascular disease (ASCVD) events (MESA). A total of 6814 asymptomatic individuals who had undergone CAC examinations and had been monitored for CHD and ASCVD events for more than ten years were included in the review. ML's analysis included medical imaging data such as CAC scores, CAC volumes, extracardiac scores, and pericardial fat volume. Beneficiary administrator bends examined the AUC region by comparing it to clinical information, the CAC Agatston score, and ML's mix of all clinical and CT criteria. ASCVD risk (0.688, p0.001) and CAC score (0.742, p0.001) were both stronger predictors of CHD occurrences than ML with all covariates (0.765, p0.001). CAC score was more accurate than ASCVD risk. ML with all components (0.763) had a superior AUC for predicting ASCVD events than either the ASCVD risk score (0.710) or the CAC score (0.714).

Swathy et al.[21] evaluated CCTA performed on 203 patients with suspected CAD. An artificial intelligence (AI) model developed by the researchers had an AUC (Application under the Bend) of 0.79, which outperformed the expected AUC (AUC = 0.65 + 0.04, P = 0.01) given easily accessible clinical variables. Another group of 34 proteins could predict the lack of CAD (AUC = 0.85 + 0.05, P = 0.05), again outflanking expectations with accessible characteristics (AUC = 0.70 + 0.04, p = 0.01). Ahmad et al.[22] developed two reciprocal protein markers using AI models based on designated proteomics. These encouraging findings support specialized proteomics in identifying cardiovascular risk factors in outcome studies.

Chang et al.[23] utilizes an original AI-driven clinical and proteomic way to foresee clinically giant PAD. A cross-sectional study of 131 short-term patients was conducted under their direction (controls, 41; PAD, 90). Because of its clinical appearance, claudication, and lower leg brachial record of less than 0.9, the cushion was examined by a board-certified vascular specialist. Ineligible were those who had appendage-specific ischemia or a history of revascularization. The atherosclerotic disease had not been diagnosed in the control group, and the lower leg brachial file was less than 0.90. A blood sample was taken for a plasma proteome analysis.

Hossain et al.[24] identified a board prophetic of PAD using the least point relapse and a final model with a minor outright shrinkage and determination administrator using five clinical criteria and 35 protein biomarkers. They went from a 1 to a three on the Rutherford scale in patients with PAD who had hypertension compared to controls (P = 0.05) (1, 32%; 2, 35%; 3, 33%). Indicative proteins were renal injury atom 1, aspiratory surfactant-related protein D, and interleukin receptor adversary, all linked to hypertension

as the sole clinical variable. The model judged cross-approval and in-example regions under the bend of 0.81 and 0.84 as PAD's existence.

Considering all factors in the study of Liu et al.[25], the optimal score was 63% awareness, 93% particularity, and 95% positive prophetic value, bringing about an in-example region under the collector working trademark bend of 0.84. This clever AI-driven clinical and proteomic symptomatic device accurately distinguished PAD from an exhaustive clinical evaluation finished by a vascular specialist. As innovation's job in helping doctors keeps developing, AI might acquire a more boundless job supporting the conclusion of persistent illnesses like PAD.

Li et al.[26] depicts coronary atherosclerosis' advancement in youthful patients and recognizes the gamble elements of unfortunate results. Members with severe or stable obstructive CAD under 45 were randomly recruited and closely monitored. The most important outcomes were all-cause mortality, MI, recalcitrant angina needing coronary revascularization, and ischemic stroke. While thinking about every repetitive occasion, similar variables and Asian nationality anticipated unfortunate results, yet relentless smoking greatly affected visualization. Untimely CAD is a forceful sickness, regardless of the suggested avoidance measures, with high paces of intermittent occasions and mortality. Identity and accompanying provocative illness are related to unfortunate anticipations, alongside lacking control of chance elements.

Huang et al.[27] fostered a brain organization to perform mechanized pressure waveform investigation and permit continuous precise ID of damping. The neural network was built and tested based on two independent datasets of well-qualified feelings from the centre's research lab. Waveforms of 5709 distinct heartbeats were extracted and grouped. The review developed an intermittent convolutional brain organization to group beats as one or the other usual, indicating damping or artefacts. Assessments from two independent labs were used to adjust for any inaccuracies. The brain network was 99.4% accurate (95% certainty stretch: 98.8% to 99.6%) while deciding against the judgments of the interior centre research facility when describing beats from the testing dataset. It was 98.7% exact (95% certainty stretch: 98.0% to 99.2%) when decided against the assessments of an outer center research facility not engaged with brain network preparation.

Gharleghi et al.[28] foster an ML model, using clinical factors to foresee stable obstructive coronary course illness (CAD). From August 2014 to January 2016, the authors analyzed 4906 individuals with stable angina or angina-like symptoms and underwent coronary angiography. Preparation (80%) and approval (20%) sets were generated from the dataset using the most prescient computation among five ML algorithms (20%). The authors compared and contrasted the pre-test probabilities of CAD scores models in the ML model (updated Diamond-Forrester and CAD consortium models). Consequences on coronary angiography, 861 of the 1312 chosen patients had obstructive CAD. In the ML model, 78.6% of the predictions were correct. For the most part, the essential elements in situating were: age; haemoglobin A1c; direction; HDL cholesterol; and trooping T. Stable obstructive CAD may be predicted with high accuracy using ML models, and novel relationships between variables can be discovered using these models.

Uma et al.[29] examined the impacts of consolidated appraisal based on CT-FFR data; six anatomic CCTA indicators were calculated (Agatston score, degree of stenosis severity, mean plaque CT decrementing esteem, the volume of non- and calcified plaques, renovating file). The characteristics are most helpful in identifying ischemia-related damage were separated using arbitrary woodland. One model, model-1, had physical CT descriptors, whereas the other model, model 2, had both physical CT descriptors and CT-FFR as part of its ROC bending. Ischemia-related sores had significantly more dangerous and non-calcified plaque volume than non-ischemic-related sores and a higher rate of rebuilding files (1.04 0.12 vs 1.11 0.13), according to the results of this study. In ischemia-related and non-ischemic-related injuries, the CT-FFR was 0.84–0.14 and 0.71–0.14, respectively. Model-1 and model-2 had ROC bend regions of 0.738 and 0.835,

respectively. Adding CT-FFR to the ischemia sore gamble resulted in significant improvements in renaming and coordinated segregation, respectively, with net renaming improvements of 0.297 and 0.254. It was possible to differentiate painful explicit ischemia using a combined evaluation of physical CCTA highlights and utilitarian CT-FFR.

Mehta and Shukla[30] investigated AI (ML) models using eXtreme Gradient Boosting (XGBoost). For a woman to be at risk for a STEMI, she must have persistent renal failure, a high pulse, and be over 70 years old. Elevated trooping T levels, severe renal failure, and age 75 and above were the most common indicators for males. Low trooping levels, high urea levels, and age over or equal to 80 years were the most critical indicators in women with NSTEMI. It had common for males to have an elevated pulse, high creatinine levels, and chronological age of more than 70 years. Results from their analysis of EHR-based mortality models for several subpopulations of the ACS suggested possible crucial and intelligent sex-explicit gamble signals. Men and women had different risk indicators, highlighting the importance of considering sex-explicit risk factors when developing treatment plans and achieving better clinical outcomes.

Numerous difficulties exist associated with the current methods for predicting artery disease. One significant challenge pertains to using conventional risk factors, including age, gender, and smoking status, in various scenarios. It is worth noting that these factors may not consistently serve as accurate predictors of disease across all individuals. Furthermore, it is worth noting that these methodologies may not adequately consider more contemporary risk factors, such as genetic predisposition or lifestyle variables, including dietary patterns and physical activity. One additional challenge pertains to the reliance on population-level data for many of these methods, which may not account for the unique characteristics of individual patients. Consequently, there is a potential for either overestimating or underestimating the associated risk. Ultimately, it is essential to acknowledge that potential challenges may arise concerning the quality or accessibility of data, thereby potentially compromising the precision of these predictive techniques.

3. Materials and methods

3.1. Dataset

This research utilizes an online Kaggle dataset for CAD disease[31]. According to the Singapore Heart Foundation, 31.7% of all fatalities in Singapore in 2020 will be caused by cardiovascular disease. When fat deposits accumulate in the coronary arteries, the arteries that feed blood to the heart muscle stiffen and constrict, resulting in coronary artery disease (CAD). It is said that CAD is due to lifestyle, which, if detected early, might allow monitoring and lifestyle changes to reduce the risk of cardiovascular disease. This prompted me to focus on predicting CAD risk[32].

To facilitate the understanding of the diseases in the datasets, it has been mapped out the diseases. Many diseases, such as heart failure, heart attack, CHD, stroke, cardiogenic shock, and AV heart block, are caused by CAD directly/indirectly. As such, it has been excluded as the feature to predict CAD because it aims to have earlier detection of CAD to reduce the risk of those diseases. Those with stroke/heart failure would not need a prediction as they might have ended up in the hospital when they learned about it.

3.2. Data pre-processing

Outliers pose a severe threat. They have a significant impact on the model's outcome. Researchers often examine outliers to determine if a given record is the consequence of a mistake in data collection or a unique phenomenon that should be considered when processing data[33].

3.2.1. Handling missing values

The imputation process generates reasonable assumptions to fill up the data gaps. When the 10% of missing data is low, it is the most useful. Insufficient natural variation can prevent a helpful model from

emerging if there is excessive missing data[22]. Alternatively, you can purge the system of any previously stored information. Removing related data while dealing with data that is absent at random is possible. A reliable analysis may not be possible if there aren't enough data points to conclude. In some cases, it may be necessary to keep track of certain occurrences or elements[33].

3.2.2. Handling outliers

In this paper, an outlier is identified using the Z-score approach. When the distribution of a variable closely resembles that of a Gaussian, this approach is commonly employed. The Z-score measures how far a variable's value is from its mean in terms of standard deviations[34].

$$\text{Z-score} = (\text{X-mean})/\text{Standarddeviation} \quad (1)$$

When the standard normal distribution is the distribution of a variable, if the values of a variable are transformed to Z-scores and the standard deviation is equal to 1. To identify outliers, the Z-score approach relies on a user-specified cut-off. The most common lower and higher cut-off values are −3 and +3. According to a conventional normal distribution, 99.7% of all values fall inside a range of −3 and +3[35].

3.3. C5.0 decision tree algorithm

Robust classifiers use a tree structure to describe connections between features and possible outcomes, such as decision tree learners tree-like structure was named because it mimics how a tree grows from a large trunk to smaller and smaller branches as one ascends[36]. There is no better choice for decision trees than C5.0, an algorithm that can handle most issues immediately. C5.0's decision trees outperform other complex machine learning models but are simpler to comprehend and use. The following table illustrates that the algorithm's vulnerabilities are minor and can be avoided in most cases[37].

Strengths

- Highly automated learning technique that can accept numeric or nominal characteristics and missing data. An all-purpose classifier.
- It removes features that aren't necessary.
- It may be applied to small and large datasets alike.
- Results in a model that is understandable to those with no prior math knowledge (for relatively small trees).
- Better than other sophisticated models in terms of efficiency.

Weaknesses

- Models based on decision trees tend to favour characteristics with many levels for splitting.
- The model may easily be over- or under-fitted.
- Because of the dependency on axis-parallel splits, it may be challenging to represent some interactions.
- Even little adjustments to the training data can significantly impact the reasoning of the choice.
- The judgments made by large trees may appear illogical because of the difficulty in interpreting their actions.

For the sake of brevity, our prior decision tree examples omitted the mathematics needed in a machine's divide and conquer method. Let's take a closer look at this to see how this heuristic is put to use in the real world. More trials lead to better results when using C5.0's boost feature, allowing any number of attempts to be used. Boosted classifiers are more time-consuming, but the benefits are worth the effort! Maximizing predicted accuracy using a boosted classifier is always a good idea, even if the unboosted classifiers are already highly accurate[38].

Figure 1 depicts the three nodes of the total of 145 nodes generated for the current problem. The particular node shows the category (0 or 1), % (% age), and n (no of patients) of specific types. Variable misclassification costs are one of the many new features in C5.0. In theory, all classification errors are considered the same in C4.5, but specific ones are more significant in practice. To reduce expected misclassification costs rather than error rates, C5.0 builds classifiers with the ability to designate a different charge for each predicted/actual class combination. In addition, the cases themselves may be of varying interest to the public. For example, the value of each instance may differ depending on the account size in an application that defines persons as "churn-like" or "non-lurking." A property in C5.0 that indicates the relevance of a case can be used to reduce the weighted predictive error rate[39].

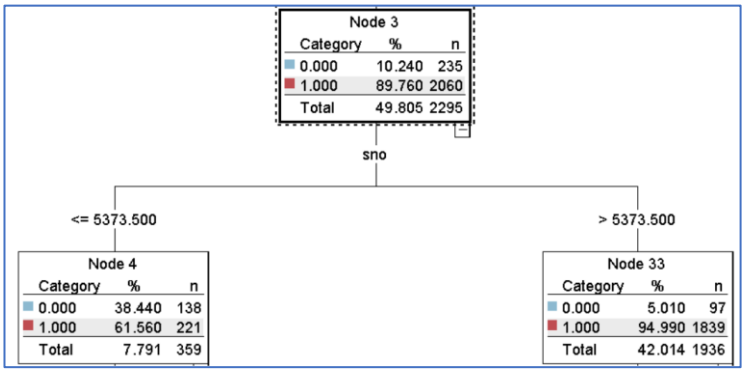

Figure 1. Node in C5.0 decision tree.

Dates, timings, timestamps, ordered discrete characteristics, and case labels are just some new data types in C5.0. C5.0 also allows values to be marked as inapplicable in addition to missing values. In addition, unique attributes may be defined as functions of existing attributes thanks to C5.0. With hundreds or thousands of characteristics, some modern data mining applications are incredibly high-dimensional. Only slightly relevant attributes can be automatically discarded before a classifier is formed in C5.0. Winnowing may minimize the size of classifiers and improve predicted accuracy in high-dimensional applications. It can also shorten the time constructing rule sets[19].

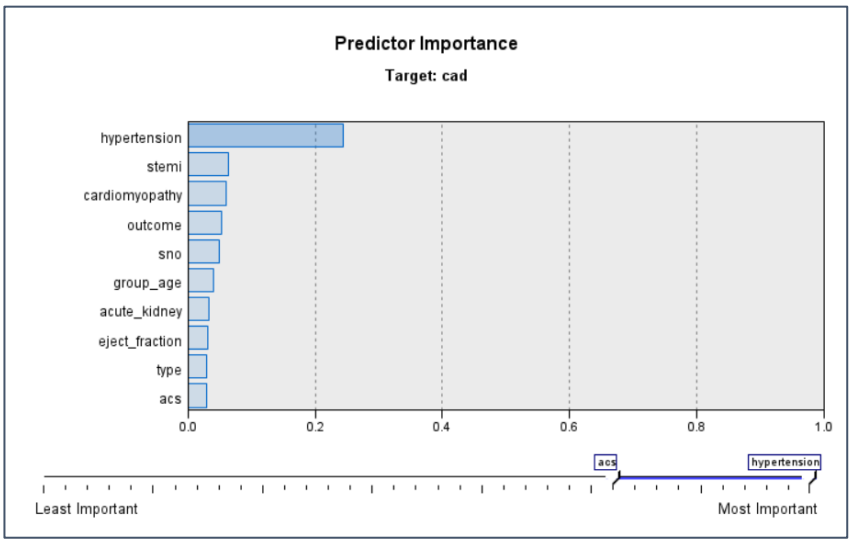

Figure 2. Predictor Importance in C5.0 decision tree.

Figure 2 depicts the predicator importance in the C5.0 decision tree generated for the current problem. The following section represents the structure of the rules used for prediction of CAD prediction as given below:

```
Rule 1—Estimated accuracy 85.07% [boost 91.6%]
        infective_endocarditis = 1 [Mode: 0] => 0
        infective_endocarditis = 0 [Mode: 1]
                hypertension = 1 [Mode: 1]
                        sno <= 5373.500 [Mode: 1]
                                valvular = 1 [Mode: 0] => 0
                                valvular = 0 [Mode: 1]
                                        ventricular = 1 [Mode: 1] => 1
                                        ventricular = 0 [Mode: 1]
                                                stemi = 1 [Mode: 1] => 1
                                                stemi = 0 [Mode: 1]
                                                        eject_fraction <= 59 [Mode: 1]
                                                                outcome = DAMA [Mode: 1] => 1
                                                                outcome = DISCHARGE [Mode: 1]
                                                group_age in [“0–30” “76–150”] [Mode: 1] => 1
                                                group_age in [“31–45”] [Mode: 0] => 0
                                                group_age in [“46–60”] [Mode: 1]
                                                group_plate = high [Mode: 1] => 1
                                                        group_plate = low [Mode: 0]
                                                        leuk_count <= 13.800 [Mode: 0] => 0
                                                        leuk_count > 13.800 [Mode: 1] => 1
                                                        group_plate = normal [Mode: 1]
                                                cardiomyopathy = 1 [Mode: 0]
                                                        age <= 54.500 [Mode: 0] => 0
                                                        age > 54.500 [Mode: 1]
                        urea <= 24 [Mode: 0] => 0
                urea > 24 [Mode: 1] => 1
                                                        cardiomyopathy = 0 [Mode: 1] => 1
                                                group_age in [“61-75”] [Mode: 1]
                                                        atrial_fibril = 1 [Mode: 0] => 0
                                                                atrial_fibril = 0 [Mode: 1] => 1
                                                        outcome = EXPIRY [Mode: 0] => 0
                                                        eject_fraction> 59 [Mode: 0] => 0
                        sno > 5373.500 [Mode: 1]
                                outcome in [“DAMA” “DISCHARGE”] [Mode: 1] => 1
                                outcome in [“EXPIRY”] [Mode: 0]
                                        group_age in [“0–30”] [Mode: 0] => 0
                                        group_age in [“31–45” “61–75”] [Mode: 0] => 0
                                        group_age in [“46–60”] [Mode: 1] => 1
                                        group_age in [“76–150”] [Mode: 0]
                                                acute_kidney = 1 [Mode: 0] => 0
                                                acute_kidney = 0 [Mode: 1]
                                                        haemoglobin <= 12.300 [Mode: 1] => 1
                                                        haemoglobin > 12.300 [Mode: 0] => 0
                                        to be continued for other features.
```

C5.0, on the other hand, is more user-friendly. C4.5’s tools for producing decision trees and rule sets have been unified into a single program, simplifying and extending the available options.

3.4. Hybrid C5.0 decision tree algorithm with boosting

Boosting is a technique used by the C5.0 algorithm to improve its accuracy rate. It operates by sequentially constructing several different models. Construction of the first model proceeds as normal. The records incorrectly categorized by the first model are then used to build a second model. Once the second model’s problems have been discovered, a third model is constructed to focus on those errors. A weighted voting technique integrates individual forecasts into a single overall prediction before instances may be

categorized using all available models. Although a C5.0 model's accuracy may be significantly improved by boosting, the training time required is higher.

Boosting algorithms are unique algorithms that enhance the data model's current results and correct faults. A weighted average and higher voting values anticipate the dialogue between weak and robust learners. A decision stamp and margin-maximizing classification are both used in these methods. Algorithms include AdaBoost or Adaptive boosting algorithm, Gradient, and XG boosting algorithm, just a few examples. These machine learning algorithms undertake a training phase to forecast and fine-tune the outcome.

The boosting method generates many weak learners and combines their predictions into a single strong one. Machine learning algorithms are applied to the data set in various ways to develop these shaky rules. Each of the iteration of these algorithms generates a new set of weak regulations. Weak learners are merged to create a stronger learner, which can predict results more accurately. The working of the proposed algorithm is explained below (Algorithm 1).

Step 1. The data is read in the first step, and the base algorithm gives each sample observation equal weight.

Step 2. The primary learner's incorrect predictions are recognized. The following iteration places considerable weight on these inaccurate predictions in the base learner.

Step 3. Repeat step 2 until the algorithm can categorize the output appropriately.

Algorithm 1 Hybrid C5.0 decision tree algorithm with boosting//Pseudo code for proposed method

1. Initialize a set of CAD training data D
2. Set the number of iterations T
3. Initialize the weights for each instance in D to 1/n, where n is the number of the cases in D
4. For t = 1 to T:
    a. Train a C5.0 decision tree on D with instance weights
    b. Calculate the error rate of the decision tree on D
    c. Calculate the weight of the decision tree as ln ((1-error rate)/error rate)
    d. Update the weights for each instance in D based on whether the decision tree correctly or incorrectly classified it
5. Output the final Boosted C5.0 decision tree

Mathematical modelling for the proposed model

A mathematical model for the proposed hybrid model is created. Let CAD dataset *D* and a loss function *L*: R2→R, the proposed boosting algorithm iteratively constructs a model *F*: X→R to minimize the empirical risk ED [*L*(*F*(*x*), *Y*)] of CAD disease[40].

At each iteration *t*, the model is updated as:

$$F^t(x) = F^{t-1}(x) + \varepsilon H^t(x) \tag{2}$$

where $H^t(x)$ is a weak learner. This weak learner is selected to approximate the negative gradient.

$$-G^t(x, y) = \frac{\sigma L(Y, S)}{\sigma(S)} F^{t-1}(x) \tag{3}$$

Equations (2) and (3) are utilized to set each booster parameter in decision tree C5.0.

4. Experimental results analysis and discussion

This section presents the experimental details, setup requirements, comparison parameters, results analysis, and discussion.

4.1. Experimental setup and comparison parameters

The proposed hybrid C5.0 method and some well-known existing machine learning models, i.e., decision tree, AdaBoost, and random forest, were implemented using Python programming under Anaconda

distribution. Various machine learning libraries, i.e., Matplot, PyTorch and Tensor flow. The hardware details include RAM: 8 GB, HDD 50 GB, processor P5 and above. To compare the existing methods and the proposed method following performance measuring parameters were calculated[38–40].

Precision: The metric evaluates the precision of affirmative predictions generated by the model. The calculation involves determining the proportion of accurate positive forecasts concerning the combined number of accurate positive and inaccurate positive predictions. A high level of precision signifies that the model effectively generates correct positive predictions, whereas a low level of precision suggests that the model produces a significant number of false positive predictions.

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \tag{4}$$

Recall: The metric assesses the model's capacity to classify positive instances accurately. The calculation involves determining the proportion of accurate positive predictions concerning the combined number of accurate positive predictions and inaccurate negative predictions.

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{5}$$

F-measure/F1-score: The metric in question is a composite measure that integrates precision and recall, offering a unified evaluation of a model's performance. The F-measure incorporates the consideration of both false positives and false negatives, thereby offering a balanced evaluation of a model's precision and recall. A high F-measure signifies that the model exhibits elevated levels of precision and recall, whereas a low F-measure suggests that the model is deficient in either precision or recall.

$$\text{F}-\text{score} = 2 \times \frac{(\text{precision} \times \text{recall})}{(\text{precision} + \text{recall})} \tag{6}$$

AUC: The metric known as AUC, which stands for "area under the receiver operating characteristic curve" (ROC), is commonly employed to assess the effectiveness of a binary classification model. The ROC curve is constructed by graphing the TPR concerning the FPR across various threshold values. The AUC metric quantifies the extent of the Area beneath the curve, which spans from 0 to 1. Higher values of AUC correspond to superior model performance. The AUC metric comprehensively evaluates a model's capacity to classify instances, irrespective of the threshold value employed accurately. A ROC curve with an AUC value of 0.5 signifies that the model's predictive ability is equivalent to random chance, whereas an AUC of 1 signifies flawless classification performance.

Gini value: It is referred to as the Gini coefficient or Gini index and is a metric employed in economics and machine learning to quantify levels of inequality. The Gini value is utilized in machine learning to assess the efficacy of a split within a decision tree. The metric quantifies the likelihood that a randomly selected instance from a dataset would be misclassified after being randomly assigned a label based on the label distribution within the dataset. A Gini coefficient of 0 signifies flawless classification, whereas a Gini coefficient of 1 suggests that the tags are distributed randomly, rendering the model incapable of making precise predictions. The split with the lowest Gini value is typically selected as the optimal split in decision trees.

Where True Positive Rate: (TPR), False Positive Rate: (FPR), True Positive: (TP), False Positive: (FP), False Negative: (FN), True Negative (TN).

4.2. Performance of C5.0 decision tree algorithm

The C5.0 decision tree algorithm has been implemented in Python in this research work. First, the simple C5.0 decision tree algorithm has been executed with the CAD Kaggle dataset[31]. The training and testing partition is taken at 70:30 ratios during the execution of both decision tree models.

The performance of the C5.0 is evaluated in terms of accuracy, AUC and Gini values. **Table 1** represents the coincidence matrix for the current domain. This table shows the correlation between observed and predicted values generated by the C5.0 model.

Table 1. Coincidence matrix for basic C5.0. model.

'Partition' = 1_Training	**0**	**1**
0	1049	367
1	321	2871
'Partition' = 2_Testing	**0**	**1**
0	360	226
1	185	1232

Figure 3 highlights the performance of the C5.0 decision tree. The accuracy level achieved in the training and testing phase is 85.07% and 79.48%, respectively. The AUC value reached in the training and testing phase is 0.913 and 0.864, respectively. The Gini value achieved in the training and testing phase is 0.827 and 0.727, respectively. This means that the C5.0 model learns concepts from the noise or random oscillations in the training data.

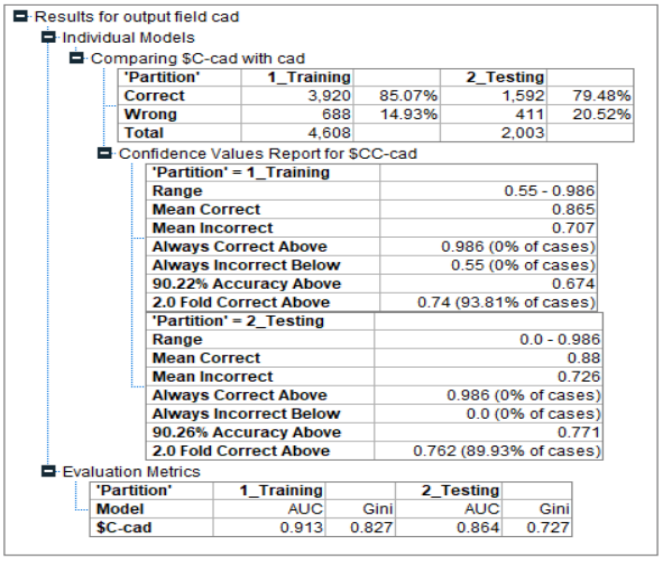

Figure 3. Results of C5.0 decision tree.

4.3. Performance of hybrid C5.0 decision tree algorithm with boosting

Table 2 represents the coincidence matrix for Boosted C5.0 model applied for the current domain. This table shows the correlation between observed and predicted values generated by Boosted C5.0 model.

Table 2. Coincidence matrix for Boosted C5.0. model.

'Partition' = 1_Training	**0**	**1**
0	1186	230
1	156	3036
'Partition' = 2_Testing	**0**	**1**
0	367	219
1	155	1262

Figure 4 highlights the performance of the C5.0 decision tree. The accuracy level achieved in the training and testing phase is 91.62% and 81.33%, respectively. The AUC value gained in the training and

testing phase is 0.957 and 0.88, respectively. The Gini value achieved in the training and testing phase is 0.914 and 0.759, respectively.

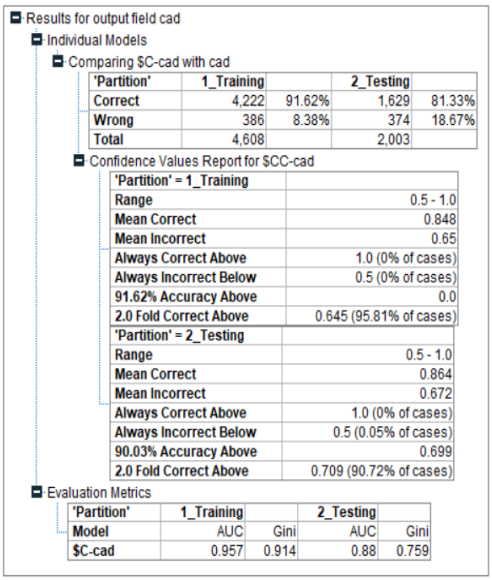

Figure 4. Results of Boosted C5.0 decision tree.

4.4. Comparison of the existing and proposed method

The proposed method and some well-known existing machine learning models, i.e., decision tree, AdaBoost, and random forest, were implemented on the CAD dataset[31] with 10-fold cross-validation. **Table 3** and **Figure 5** show an experimental results comparison of existing and proposed methods.

Table 3. Results comparison of existing and proposed methods.

Methods	Precision	Recall	F-measure	AUC	Gini value
Random forest[2]	81.45%	86.34%	85.86%	85.41%	84.62%
AdaBoost[3]	78.96%	81.24%	80.12%	80.74%	79.89%
C5.0 decision tree	85.07%	86.98%	88.78%	91.3%	89.32%
Proposed model	91.62%	93.65%	95.65%	95.7%	91.4%

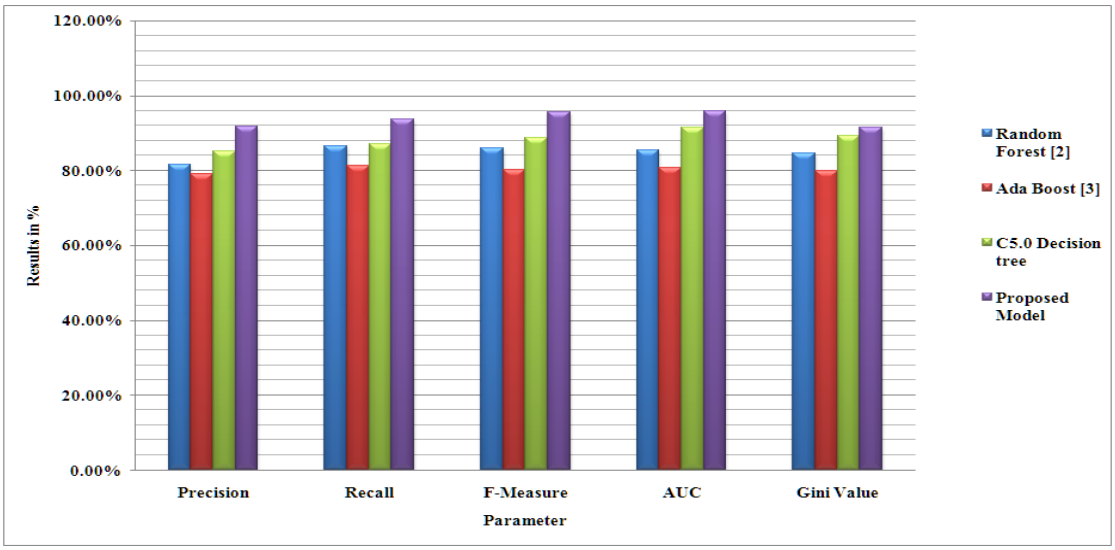

Figure 5. Results comparison of existing and proposed methods.

In the above **Figure 5**, random forest[2] method achieved a precision of 81.45%, recall of 86.34%, F-measure of 85.86%, AUC of 85.41% and Gini value of 84.62%. Another existing method, AdaBoost[3], achieved a precision of 78.96%, recall of 81.24%, F-measure of 80.12%, AUC of 80.74% and Gini value of 79.89%. Like another method, the C5.0 decision tree achieved a precision of 85.07%, recall of 86.98%, F-measure of 88.78%, AUC of 91.3% and Gini value of 89.32%. The proposed C5.0 hybrid boosting method achieved a precision of 91.62%, recall of 93.65%, F-measure of 95.7%, AUC of 95.7%, and Gini value of 91.4%. Based on the experimental analysis, the proposed method has achieved better results than all the existing methods.

5. Conclusions and future scope

Medical professionals have always benefited from clinical decision support systems when making diagnoses. Many deaths today are caused by coronary artery disease (CAD), prompting researchers to suggest more precise forecasting models. Standard C5.0 decision tree models are outperformed by machine learning algorithms when classifying and predicting illness accurately. Methodologies like this can be used to identify at-risk patients in cases where good risk prediction models do not exist or if they have been proven to have poor performance.

In this research, we developed a Hybrid Boosted C5.0 model by combining the C5.0 decision tree and boosting methods. Boosting is a supervised machine learning method that leverages numerous inadequate models to construct a more robust and powerful model. The proposed model and some well-known existing machine learning models, i.e., decision tree, AdaBoost, and random forest, were implemented using an online coronary artery disease dataset of 6611 patients and compared based on various performance measuring parameters. Experimental analysis shows that the proposed model achieved an accuracy of 91.62% at training and 81.33% at the testing phase. The AUC value achieved in the training and testing phase is 0.957 and 0.88, respectively. The Gini value achieved in the training and testing phase is 0.914 and 0.759, respectively, far better than the proposed method.

Aiming to reduce the frequency of undetected illnesses and the burden of unfavourable clinical outcomes due to delays in preventative measures, future studies should focus on testing, automating, and prospectively validating local models. An alternative strategy for enhancing the diagnosis of CAD disease on the data set being used, as well as other real datasets, involves the application of deep learning models. Integrating deep learning methodologies with distributed architectural and design data can further augment diagnostic capabilities.

Author contributions

Conceptualization, SD and UKL; methodology, SS and SA; software, VJ; validation, AM, VJ and PC; formal analysis, UKL; investigation, SD; resources, AA; data curation, SD; writing—original draft preparation, SA; writing—review and editing, AA and MM; visualization, PC; supervision, AM; project administration, MM; funding acquisition, UKL. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Conflict of interest

The authors declare no conflict of interest.

Ethical approval and consent to participate

No ethical approval is required, and the authors consent to participate in the paper.

Consent for publication

Authors provide support for publication.

Data availability statement

The datasets used and analyzed during the current study are available from the corresponding author upon reasonable request.

References

1. Wang G, Gao Y, Xu F, et al. GW28-e0388 A novel machine-learning model for identification of significant coronary artery disease. *Journal of the American College of Cardiology* 2017; 70(16): C113. doi: 10.1016/j.jacc.2017.07.400
2. Stuckey T, Singh N, Goswami R, et al. TCT-177 Assessing coronary artery disease by cardiac phase tomography using machine-learned algorithms in obese and elderly subjects. *Journal of the American College of Cardiology* 2017; 70(18): B75–B76. doi: 10.1016/j.jacc.2017.09.245
3. Griffin WF, Choi AD, Riess JS, et al. AI evaluation of stenosis on coronary CT angiography, comparison with quantitative coronary angiography and fractional flow reserve. *JACC: Cardiovasc Imaging* 2022; 16(2): 193–205. doi: 10.1016/j.jcmg.2021.10.020
4. Rahman F, Finkelstein N, Alyakin A, et al. Using machine learning for early prediction of cardiogenic shock in patients with acute heart failure. *Journal of the Society for Cardiovascular Angiography & Interventions* 2022; 1(3): 100308. doi: 10.1016/j.jscai.2022.100308
5. Ross EG, Shah NH, Dalman RL, et al. The use of machine learning for the identification of peripheral artery disease and future mortality risk. *Journal of Vascular Surgery* 2016; 64(5): 1515–1522.e3. doi: 10.1016/j.jvs.2016.04.026
6. Park JY, Noh YK, Choi BG, et al. TCTAP A-010 A machine learning-based approach to prediction of acute coronary syndrome. *Journal of the American College of Cardiology* 2015; 65(17): S6. doi: 10.1016/j.jacc.2015.03.057
7. Stuckey T, Singh N, Goswami R, et al. TCT-154 Gender based assessment of coronary artery disease by cardiac phase tomography using machine-learned algorithms. *Journal of the American College of Cardiology* 2017; 70(18): B66. doi: 10.1016/j.jacc.2017.09.218
8. Betancur JA, Otaki Y, Fish M, et al. Rest scan does not improve automatic machine learning prediction of major adverse coronary events after high speed myocardial perfusion imaging. *Journal of the American College of Cardiology* 2017; 69(11): 1590. doi: 10.1016/s0735-1097(17)34979-3
9. Ghosh P, Lilhore UK, Simaiya S, et al. Prediction of the risk of heart attack using machine learning techniques. In: Sharma S, Peng SL, Agrawal J, et al. (editors). *Data, Engineering and Applications*. Springer, Singapore; 2022. Volume 907. pp. 613–621.
10. Goswami R, Stuckey T, Meine F, et al. Coronary artery disease learning and algorithm development study: Early analysis of ejection fraction evaluation. *Journal of the American College of Cardiology* 2017; 69(11): 953. doi: 10.1016/s0735-1097(17)34342-5
11. Nakanishi R, Dey D, Commandeur F, et al. Machine learning in predicting coronary heart disease and cardiovascular disease events: Results from the multi-ethnic study of atherosclerosis (Mesa). *Journal of the American College of Cardiology*2018; 71(11): A1483. doi: 10.1016/s0735-1097(18)32024-2
12. Bom MJ, Levin E, Driessen RS, et al. Predictive value of targeted proteomics for coronary plaque morphology in patients with suspected coronary artery disease. *eBioMedicine* 2019; 39: 109–117. doi: 10.1016/j.ebiom.2018.12.033
13. Ramirez JL, Magaret CA, Khetani SA, et al. PC102. A novel machine learning-driven clinical and proteomic tool for the diagnosis of peripheral artery disease. *Journal of Vascular Surgery* 2019; 69(6): e233–e234. doi: 10.1016/j.jvs.2019.04.344
14. Collet JP, Zeitouni M, Procopi N, et al. Long-term evolution of premature coronary artery disease. *Journal of the American College of Cardiology* 2019; 74(15): 1868–1878. doi: 10.1016/j.jacc.2019.08.1002
15. Howard JP, Cook CM, van de Hoef TP, et al. Artificial Intelligence for aortic pressure waveform analysis during coronary angiography: Machine learning for patient safety. *JACC: Cardiovascular Interventions* 2019; 12(20): 2093–2101. doi: 10.1016/j.jcin.2019.06.036

16. Kim JT, Cho S, Lee SY, et al. The use of machine learning algorithms for the identification of stable obstructive coronary artery disease. *Journal of the American College of Cardiology* 2020; 75(11): 254. doi: 10.1016/S0735-1097(20)30881-0
17. Kawasaki T, Kidoh M, Kido T, et al. Evaluation of significant coronary artery disease based on CT fractional flow reserve and plaque characteristics using random forest analysis in machine learning. *Academic Radiology* 2020; 27(12): 1700–1708. doi: 10.1016/j.acra.2019.12.013
18. Vazquez B, Fuentes-Pineda G, Garcia F, et al. Risk markers by sex for in-hospital mortality in patients with acute coronary syndrome: A machine learning approach. *Informatics in Medicine Unlocked* 2021; 27: 100791. doi: 10.1016/j.imu.2021.100791
19. Sandhu JK, Lilhore UK, Poongodi M, et al. Predicting the risk of heart failure based on clinical data. *Human-centric Computing and Information Sciences* 2022; 12: 57. doi: 10.22967/HCIS.2022.12.057
20. Schwalm JD, Di S, Sheth T, et al. A machine learning-based clinical decision support algorithm for reducing unnecessary coronary angiograms. *Cardiovascular Digital Health Journal* 2022; 3(1): 21–30. doi: 10.1016/j.cvdhj.2021.12.001
21. Swathy M, Saruladha K. A comparative study of classification and prediction of Cardio-Vascular Diseases (CVD) using machine learning and deep learning techniques. *ICT Express* 2022; 8(1): 109–116. doi: 10.1016/j.icte.2021.08.021
22. Ahmad A, Corban MT, Moriarty JP, et al. Coronary reactivity assessment is associated with lower health care—Associated costs in patients presenting with angina and nonobstructive coronary artery disease. *Circulation: Cardiovascular Interventions* 2023; 16(7): e012387. doi: 10.1161/CIRCINTERVENTIONS.122.012387
23. Chang V, Bhavani VR, Xu AQ, Hossain MA. An artificial intelligence model for heart disease detection using machine learning algorithms. *Healthcare Analytics* 2022; 2: 100016. doi: 10.1016/j.health.2022.100016
24. Hossain MM, Swarna RA, Mostafiz R. Analysis of the performance of feature optimization techniques for the diagnosis of machine learning-based chronic kidney disease. *Machine Learning with Applications* 2022; 9: 100330. doi: 10.1016/j.mlwa.2022.100330
25. Liu Y, Ren H, Fanous H, et al. A machine learning model in predicting hemodynamically significant coronary artery disease: A prospective cohort study. *Cardiovascular Digital Health Journal* 2022; 3(3): 112–117. doi: 10.1016/j.cvdhj.2022.02.002
26. Li Q, Campan A, Ren A, Eid WE. Automating and improving cardiovascular disease prediction using machine learning and EMR data features from a regional healthcare system. *International Journal of Medical Informatics* 2022; 163: 104786. doi: 10.1016/j.ijmedinf.2022.104786
27. Huang Z, Xiao J, Wang X, et al. Clinical evaluation of the automatic coronary artery disease reporting and data system (CAD-RADS) in coronary computed tomography angiography using convolutional neural networks. *Academic Radiology* 2023; 30(4): 698–706. doi: 10.1016/j.acra.2022.05.015
28. Gharleghi R, Adikari D, Ellenberger K, et al. Automated segmentation of normal and diseased coronary arteries—The ASOCA challenge. *Computerized Medical Imaging and Graphics* 2022; 97: 102049. doi: 10.1016/j.compmedimag.2022.102049
29. Uma KV, Pudumalar S, Sharon blessie E. A combined classification algorithm based on C5.0 and NB to predict chronic obstructive pulmonary disease. In: Proceedings of the 2018 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC); 13–15 December 2018; Madurai, India. pp. 1–4.
30. Mehta S, Shukla D. Optimization of C5.0 classifier using Bayesian theory. In: Proceedings of the 2015 International Conference on Computer, Communication and Control (IC4); 10–12 September 2015; Indore, India. pp. 1–6.
31. Coronary artery disease analysis & prediction. Available online: https://www.kaggle.com/code/homelysmile/coronary-artery-disease-analysis-prediction/data?select=DataClean-fullage.csv (accessed on 15 September 2022).
32. Wang M, Gao K, Wang L, Miu X. A novel hyperspectral classification method based on C5.0 decision tree of multiple combined classifiers. In: Proceedings of the 2012 Fourth International Conference on Computational and Information Sciences; 17–19 August 2012; Chongqing, China. pp. 373–376.
33. Jincheng Y, Ping J, Guangyu C, et al. Application of C5.0 algorithm in failure prediction of smart meters. In: Proceedings of the 2016 13th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP); 16–18 December 2016; Chengdu, China. pp. 328–333.
34. Pashaei E, Ozen M, Aydin N. Improving medical diagnosis reliability using Boosted C5.0 decision tree empowered by Particle Swarm Optimization. In: Proceedings of the 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 25–29 August 2015; Milan, Italy. pp. 7230–7233.
35. Dalal S, Onyema EM, Kumar P, et al. A hybrid machine learning model for timely prediction of breast cancer. *International Journal of Modeling, Simulation, and Scientific Computing* 2023. doi: 10.1142/S1793962323410234
36. Edeh MO, Dalal S, Dhaou IB, et al. Artificial intelligence-based ensemble learning model for prediction of hepatitis C disease. *Frontiers in Public Health* 2022; 10: 892371. doi: 10.3389/fpubh.2022.892371
37. Onyema EM, Shukla PK, Dalal S, et al. Enhancement of patient facial recognition through deep learning algorithm: ConvNet. *Journal of Healthcare Engineering* 2021; 2021: 5196000. doi: 10.1155/2021/5196000

38. Ramesh TR, Lilhore UK, Poongodi M, et al. Predictive analysis of heart diseases with machine learning approaches. *Malaysian Journal of Computer Science* 2022; 2022: 132–148. doi: 10.22452/mjcs.sp2022no1.10
39. Chauhan AS, Lilhore UK, Gupta AK, et al. Comparative analysis of supervised machine and deep learning algorithms for kyphosis disease detection. *Applied Sciences* 2023; 13(8): 5012. doi: 10.3390/app13085012
40. Asif D, Bibi M, Arif MS, Mukheimer A. Enhancing heart disease prediction through ensemble learning techniques with hyperparameter optimization. *Algorithms* 2023; 16(6): 308. doi: 10.3390/a16060308