

Machine learning algorithms for predicting coronary artery disease: efforts toward an open source solution

Aravind Akella*,¹  & Sudheer Akella¹ ¹Qualicel Global Inc., Huntington Station, NY 11746, USA*Author for correspondence: aravindbakella@gmail.com

Aim: The development of coronary artery disease (CAD), a highly prevalent disease worldwide, is influenced by several modifiable risk factors. Predictive models built using machine learning (ML) algorithms may assist clinicians in timely detection of CAD and may improve outcomes. **Materials & methods:** In this study, we applied six different ML algorithms to predict the presence of CAD amongst patients listed in 'the Cleveland dataset.' The generated computer code is provided as a working open source solution with the ultimate goal to achieve a viable clinical tool for CAD detection. **Results:** All six ML algorithms achieved accuracies greater than 80%, with the 'neural network' algorithm achieving accuracy greater than 93%. The recall achieved with the 'neural network' model is also the highest of the six models (0.93), indicating that predictive ML models may provide diagnostic value in CAD.

Lay abstract: Coronary artery disease (CAD) is correlated with many preventable risk factors. Early diagnosis of CAD allows for prevention of worsening of CAD and its complications. This study aims to utilize machine learning (ML) algorithms to predict for CAD in patients. Our results indicate that ML algorithms can accurately predict for CAD. Furthermore, by providing our code publicly, we hope to improve the ability for ML algorithms as a diagnostic tool for CAD.

First draft submitted: 11 December 2020; Accepted for publication: 26 February 2021; Published online: 29 March 2021

Keywords: CAD • heart dataset • heart disease • ML algorithms • neural networks

Coronary artery disease (CAD) is the most common type of heart disease, affecting millions worldwide. According to recent statistics from the American Heart Association, coronary heart disease accounted for 13% of deaths in the USA in 2018 [21]. Worldwide in 2015, CAD was found to be one of the most common causes of death, with 15.6% of all deaths resulting from the disease [1]. Because CAD is associated with several modifiable risk factors pertaining to lifestyle and intervention, timing of detection and diagnostic accuracy are especially relevant in the clinical management of patients with CAD.

Over the past several years, approaches that include machine learning (ML) are making significant impact in the detection and diagnosis of diseases [2–7]. In general, the ML approach involves 'training' an algorithm with a control dataset for which the disease status (disease or no disease) is known, and then applying this trained algorithm to a variable dataset in order to predict the disease status in patients for whom it is not yet determined. As larger cohorts of data are introduced, the ML algorithm will be better trained as a predictor for disease status. More accurate disease prediction with ML would empower clinicians with improved detection, diagnosis, classification, risk stratification and ultimately, management of patients, all while potentially minimizing required clinical intervention.

The application of ML concepts to CAD has been significantly hampered by the availability of appropriate clinical datasets. However, one of the components of the 'UCI Heart Disease Dataset,' dubbed as the 'Cleveland Dataset,' is publicly available on the UCI Machine Learning Repository (see [22]). Originally intended to be a teaching aid, the Cleveland dataset has been highly exploited for exploring ML concepts. Available since 1988, the Cleveland dataset has so far received more than one million downloads and is currently ranked as the fifth most popular dataset in the UCI repository. A GoogleScholar search with the terms 'Cleveland dataset,' 'heart

disease' and 'machine learning,' returns little more than 300 records available since 2010, of which most studies analyzed one ML method at a time (for the latest survey review, see [8]). In addition, there is no way to verify claims in any of the publications for the accuracy of the algorithms, as the computer code has not been made publicly available [23]. A PubMed search using the same keywords identified ten original articles and one comprehensive review article [9]. Of the ten original articles, three used the Cleveland dataset, whereas the remaining seven used proprietary datasets. Of the ten, none of the studies have made the computer code publicly available. In our view, all recent studies pertaining to the application of the ML approach to CAD thus far appear exploratory rather than seeking to provide clinical assistance to healthcare practitioners in the treatment of CAD. (Also see 'Related Works' in the Supplementary materials files for a list of recently published articles that encompass ML algorithms with detection and diagnosis of disease.)

We have now undertaken a comparative analysis by applying six different ML algorithms (models) using the UCI Cleveland dataset to predict disease outcomes. In an effort to initiate an open source ML solution for detecting CAD, we have deposited our computer code on GitHub (see [25]), making it available for other researchers to test and improve our work. We also welcome the opportunity to gain access to larger datasets to further our efforts toward an open source solution. As for the six ML algorithms used in our study, we have found that all six (linear regression, regression tree, random forest, support-vector machine, nearest neighbor, neural network and k-nearest neighbor) perform well with an accuracy of greater than ~80%, with the nearest neighbor's accuracy greater than 93%. We consider Accuracy, Recall, F1 score and the Area Under the Curve–Receiver Operating (AUC-ROC) as performance metrics for comparative analysis amongst the ML models.

Numerous risk factor variables contribute to the development of CAD, some of which can be controlled (or modifiable) (e.g., see [10]). These include high blood pressure, high cholesterol, smoking, diabetes, obesity, lack of physical activity, unhealthy diet and stress. The risk factors that cannot be controlled (or nonmodifiable) are age, sex (gender), family history and race or ethnicity. Traditional approaches assess these risk factors to predict future risk (prognosis) of cardiovascular disease [11]. However, a large number of individuals at risk of cardiovascular disease fail to be identified by these approaches, while some individuals not at risk are given preventative treatment unnecessarily (see [12,13]). Several of the ML algorithms have the ability to summarize the impact of individual variables on response variable and are referred to as 'variables of importance' (from the model's perspective), thus aiding to the building of accurate prognostic models [14]. In the present study, we extract the variables of importance in one model in an effort to demonstrate the feasibility of including a risk assessment component in the ML model. Furthermore, we have openly hosted the computer code that was generated in this study on GitHub, so that we may provide the settings of a workflow model that promotes public contribution toward improved risk detection of CAD using ML, as others may contribute by extending the code to larger cohort groups.

The remainder of the manuscript is organized as follows. The Dataset & preprocessing section provides information on the data source, describes the data variables, explains the data preprocessing steps and provides a high-level analysis of the data. The next section provides concise descriptions of the ML models considered in this study and the methodologies of how the model results are evaluated. The results of the modeling efforts are presented and discussed in the Results & discussion section. The concluding section outlines the limitations of the present modeling efforts and provides a workflow model framework for extending our study in future.

Dataset & preprocessing

The dataset used in this study is downloaded from the repository maintained by the UCI (University of California, CA, USA) Center for Machine Learning and Intelligent Systems. The repository contains four datasets from four different hospitals. The Cleveland dataset contains fewer missing attributes than the other datasets and has more records. This dataset has fourteen variables on 303 patients. Table 1 lists all fourteen variables in the dataset, the associated datatype for each and a brief description.

During the data preprocessing steps, six rows of data with unknown values were dropped resulting in a dataset of 297 observations; two dataset columns of factor datatype were converted to numeric datatype; continuous variables (four, excluding age) were normalized; the column 'num' was renamed to 'hd' for clarity; and the values 1 to 4 in the heart disease (hd) column are replaced with value 1 to form a binary classification – disease (value of 1) or no disease (value of 0).

The correlation matrix for the 14 variables in the dataset is shown in Figure 1, depicting the Pearson coefficients, or r , corresponding to the association between two variables, which can range from -1 to +1. A higher absolute value of r indicates a stronger correlation between two variables. A positive r indicates a direct relationship between

Table 1. Dataset[†] variables names, types and descriptions.

Variable	Type	Description
Age	Continuous	Patient age in years
Sex	Categorical	Patient gender (1 = male; 0 = female)
cp	Categorical	Chest pain (1 = typical angina; 2 = atypical angina; 3 = nonanginal pain; 4 = no pain)
trestbps	Continuous	Resting blood pressure (in mmHg) on admission to the hospital
chol	Continuous	Serum cholesterol in mg/dl
fbs	Categorical	Fasting blood sugar higher than 120 mg/dl (1 = true; 0 = false)
restecg	Categorical	Resting electrocardiogram (0 = normal; 1 = ST-T wave abnormality; 2 = probable/definite left ventricular hypertrophy)
thalach	Continuous	Maximum heart rate achieved (during thallium test)
exang	Categorical	Exercise induced angina (1 = yes; 0 = no)
oldpeak	Continuous	ST depression induced by exercise relative to rest
slope	Categorical	Slope of the peak exercise ST segment (1 = up-sloping; 2 = flat; 3 = down-sloping)
ca	Categorical	Number of major vessels (0 to 3) colored by fluoroscopy
thal	Categorical	Thallium heart scan (3 = normal; 6 = fixed defect; 7 = reversible defect)
num	Categorical	Diagnosis of heart disease (angiographic disease status) (0 = absent; 1 to 4 = present)

[†]The Cleveland Heart Disease Dataset hosted on the Machine Learning Repository at the University of California-Irvine uploaded in 1988 containing data on 303 cardiac patients (see text for details).

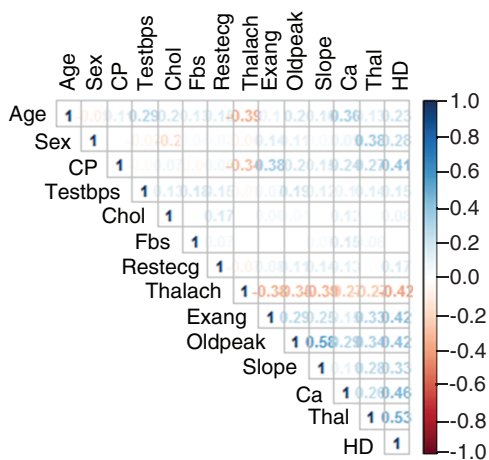


Figure 1. Correlation matrix of the various parameters in the Cleveland Heart Disease Dataset. The color coding scale denotes the degree of Pearson correlation between variables with red being negatively correlated and blue positively correlated.

two variables, whereas a negative r suggests an indirect relationship. In this analysis, we considered an absolute value of 0.5 as threshold, that is, if r is greater than 0.5 or less than -0.5, we assume those two variables are correlated. The above correlation matrix shows that only a few variables correlate with coefficients greater than 0.5, demonstrating poor correlation between the variables. Having a low value of r indicates that all these independent variables and can be included in the machine learning model building.

Additional analysis of the dataset is presented in the Supplementary material.

Machine learning models, model building & model evaluation

In applying machine learning models, it is generally understood that no single algorithm is superior to the others [15]. In machine learning, if every instance in the dataset is given to the model with known labels (the corresponding correct outputs), like in the Cleveland dataset, then the learning is called 'supervised', in contrast to 'unsupervised' learning, in which instances are unlabeled. Below, we present the general idea on how each of the six supervised machine learning algorithms work on the dataset and any assumptions we make in each case. Each algorithm is first trained (or fitted) with a fraction of the dataset, usually known as the 'training set' and then tested on the 'test set' that is put aside as 'unseen data' for evaluating the algorithm. For a detailed description of the models we refer the reader to the excellent treatise by James *et al.* [16].

Logistic regression

In Logistic regression, each independent variable in every instance of the dataset is multiplied by a weight and the cumulative result is passed to a sigmoid function. The sigmoid function maps real values into probability values between 0 and 1. In our present modeling, we left the threshold to the default value, which is 0.5, such that for probability values greater than 0.5, the model predicts the dependent variable to be 1 (the patient has CAD) and for values less than or equal to 0.5, the model predicts the dependent variable to be 0 (the patient does not have CAD).

Decision tree

Decision tree is a tree-like structure that classifies instances by sorting them based on the values of the variables. Each node in a decision tree represents a variable, and each branch represents a value that the node can assume. Instances are classified starting at the root node and sorted based on the values of the variables. The variable that best divides the dataset would be the root node of the tree. Internal nodes (or split nodes) are the decision-making part that make a decision, based on multiple algorithms and to visit subsequent nodes. The split process is terminated when a user-defined criteria is reached at the leaf (for the present modeling, we left it to be the default value, which is 20). The paths from root nodes to the leaf nodes represent classification rules.

Random forest

Random forest is an ensemble model consisting of multiple regression trees like in a forest. Random forest combines several classification trees, trains each one on a slightly different set of the dataset instances, splitting nodes in each tree considering a limited number of the variables. The final predictions of the random forest are made by averaging the predictions of each individual tree, which enhances the prediction accuracy for unseen data. The number of regression trees chosen for present modeling is (ntree) 500.

Support vector machine

In support-vector machine, each data point is plotted in an n-dimensional space with the value of each variable being the value of particular coordinates and classification is performed based on the hyperplane that differentiates the two data classes. Following this, characteristics of new instances can be used to predict the class to which a new instance should belong.

k-Nearest neighbor

k-Nearest neighbor (kNN) is one of the most basic and nonparametric algorithms, it does not make any assumptions about the distribution of the underlying data. The algorithm is based on the principle of Euclidean distance that is the instances within a dataset generally exist in close proximity to other instances that have similar properties. If the instances are tagged with a classification label, then the value of the label of an unclassified instance can be determined by observing the class of its nearest neighbors. For the present modeling, the whole process is repeated three times (repeats = 3) each with a k value of 10 (number = 10) and taking the average of the three iterations.

Artificial neural network

An artificial neuron Network (ANN) is based on the structure and functions of biological neural networks – the network learns (or changes) based on the input and output. The layers in ANN are segregated into three classes: input units, which receive information to be processed; output units, where the results of the processing are found; and units in between known as hidden units. The network is first trained on a dataset of paired data to determine input-output mapping. After training, the weights of the connections between neurons are then fixed and the network is used to determine the classifications of a new set of data. For the present modeling we consider the hidden units to be three (hidden = 3) and threshold to be 0.05 (threshold = 0.05).

To overcome our inability of using real-world data, we split the dataset into a ‘training set’ (70%, i.e., 208 observations) and a ‘test set’ (30%, i.e., 87 observations) making sure to balance the class distributions within the split (see the associated computer code on GitHub, https://github.com/aa54/CAD_1). The ‘training’ dataset is used to train the model; the model sees and learns from this training data. The ‘test’ dataset is then used to provide an unbiased evaluation of a final model fit to the training dataset. In some cases, we ran multiple experiments to validate model results on different splitting ratios.

Table 2. Performance metrics of the ML models applied on the Cleveland Heart Disease Dataset.

Model	Accuracy	Sensitivity	F1 score	AUC	Mean
Generalized linear model	0.8764	0.8000	0.8786	0.883	0.85
Decision tree	0.7978	0.7447	0.7970	0.801	0.78
Random forest	0.8764	0.8261	0.8751	0.880	0.86
Support-vector machine	0.8652	0.7959	0.8662	0.871	0.84
Neural network	0.9303	0.9380	0.8984	0.796	0.88
k-Nearest neighbor ¹	0.8427	0.7872	0.8419	0.847	0.83

Boldface values indicate highest performance group.

The following four metrics [17] were used to evaluate the performance of the predicted models and compare them with one another.

1) Accuracy: the proportion of total dataset instances that were correctly predicted out of the total instances

$$\text{accuracy} = (\text{true positives} + \text{true negatives}) / \text{total}$$

2) Recall (sensitivity): the proportion of the predicted positive dataset instances out of the actual positive instances

$$\text{sensitivity} = \text{true positives} / (\text{true positives} + \text{false negatives})$$

3) F1 score: a composite harmonic mean (average of reciprocals) that combines both precision and recall. For this, we first measure the precision, the ability of the model to identify only the relevant dataset instances

$$\text{precision} = \text{true positives} / (\text{true positives} + \text{false positives})$$

The F1 score is estimated as

$$F1 = 2 \times (\text{precision} \times \text{recall}) / (\text{precision} + \text{recall})$$

4) Area under the curve (AUC): an estimate of the probability that a model ranks a randomly chosen positive instance higher than a randomly chosen negative instance. Additionally, the full area under the curve–receiver operating characteristic (AUC-ROC) curves are plotted to visually compare the models' performance.

Results & discussion

For our machine learning model building, we started with two basic models, logistic regression and decision tree, to predict the presence of CAD. Our intuition was that it would be easy to interpret the results of the basic models and to explain each model to the nonmachine learning audience. After analyzing the results, however, we realized that the decision tree model is prone to over-fitting, and logistic regression did not perform relatively well with this simple dataset. We then considered applying the standard SMOTE methodology [18] to create additional synthetic data. Although we achieved more accurate results with SMOTE (5–6% higher accuracy), we felt that SMOTE is not necessarily a good approach as it creates data points based on the distance algorithm, and these synthetic datapoints may not be 'true' representations of patients. Therefore, we decided not to perform SMOTE on the dataset and proceeded to apply the remaining, more complex ML algorithms.

As shown in Table 2, an accuracy value higher than 0.84 is achieved with all but the decision tree model (just below 0.80). The other performance parameters, recall, F1 score and AUC are also high. A mean value is calculated with all the later three parameters for each of the models (shown in the last column in Table 2) to judge which model performs best as a whole. We excluded accuracy in calculating the mean, as accuracy is often considered to be a misleading indicator in measuring the performance of models with biomedical datasets (e.g., see [3]).

As seen in Table 2, the performance of the neural network model is outstanding with an accuracy of 0.9303 and a recall of 0.9380. In addition, running multiple experiments with differently proportioned training and test sets (changing from 70:30 to either 60:40 or 80:20) with neural network, the performance metrics were unchanged.

A high recall value indicates a lower propensity for false negatives. In disease prediction, a low recall (high frequency of false negatives), would misdiagnose patients with CAD as healthy, which may have devastating

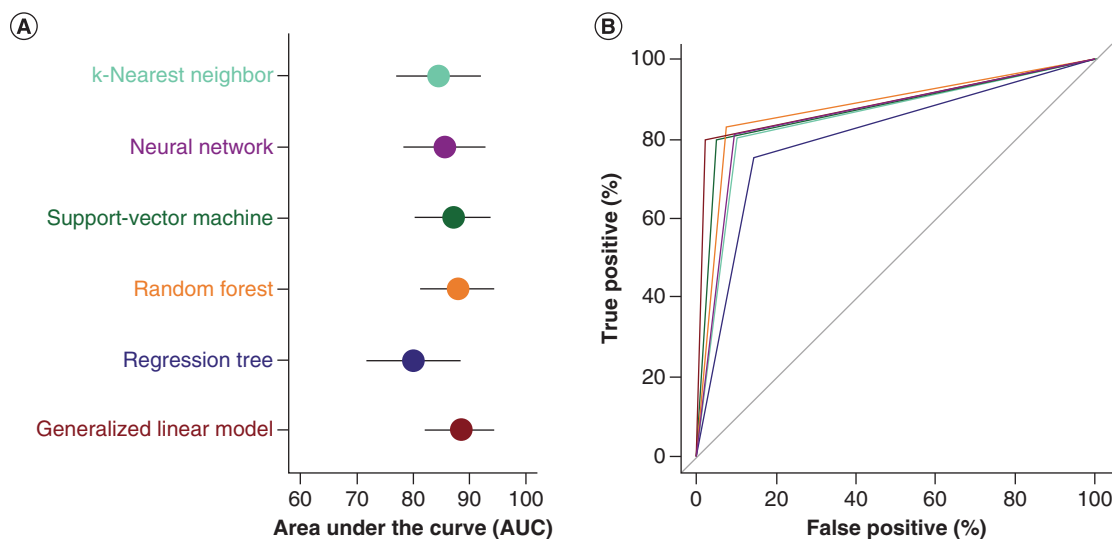


Figure 2. Classifier performance comparison with receiver operating characteristic (ROC) curves obtained in applying machine learning algorithms to the Cleveland Heart Disease Dataset. (A) Area under the ROC curve (AUC) for all the six models with estimated 95% confidence intervals. (B) Full ROC curves for the six models. Note the matching color code.

consequences. For example, a recall rate of 0.9380 (achieved with neural network) implies that this algorithm will correctly identify the presence of CAD in approximately 94 patients out of 100 patients with CAD.

Figure 2 compares the performance of the six ML algorithms using AUC-ROC. AUC-ROC curves allow to visualize the tradeoff between the true positive rate and false positive rate, whereas AUC (see above) is useful to compare multiple algorithms or hyperparameters combinations (these two are obtained by different methods). As seen in the figure, all models except for decision tree perform well, there is very little difference in the AUC-ROC curves. Nevertheless, the ROC curve corresponding to the neural network model has a relatively higher positive slope, indicating that this model gives higher ‘true positive percentage.’ Combining with the results obtained for the performance metrics, it may be concluded that the neural network model is able to predict CAD more effectively than the other models.

The computer code generated in this study, which was developed in R, has been made available on the public repository GitHub and supplementally with this manuscript. As R is a free software tool for statistical modeling with an extensive set of code libraries, collaboration and improvement on the code generated in this study is welcomed and encouraged. The code libraries in R include advanced ML methodologies (voting, bagging, optimization, etc.) with more to be added as R evolves further as an open source tool [19].

We next extracted at the ‘variable of importance’ in the neural network model. Understanding the relative importance of variables may dictate which variables are necessary to be included in the risk prediction of CAD. As shown in Figure 3, ‘restcg’ (resting electrocardiogram [EKG]) and sex (gender) have a relatively high importance, while the three attributes ‘age’, ‘thalch’ (maximum heart rate achieved during the thallium test) and ‘exang’ (exercise induced angina) have the least importance. The top ten variables of importance as assigned by the neural network model, in descending order, are: resting EKG, patient gender, chest pain, slope of the peak exercise ST segment, ST depression induced by exercise relative to rest, fasting blood sugar (higher or lower than 120 mg/dl), thallium heart scan, number of major vessels, resting blood pressure and serum cholesterol.

Heinze *et al.* [20] proposed certain recommendations on the application of variable selection methods to help modeling in life sciences and its worth following these recommendations in future efforts of identifying which of the variables of importance are worth studying in the risk modeling.

Conclusion & future perspective

In this study, we demonstrated that ML algorithms can be applied with high accuracy and recall to detect the presence of CAD using a publicly available dataset. We also demonstrated that the neural network model outperforms other

Figure 3. Variables of importance in the neural networks model. (A) shows the importance as obtained with a built-in code library function and **(B)** is the normalized graph.

restecg: Resting EKG; cp: Chest pain; slope: Slope of the peak exercise ST segment; oldpeak: ST depression induced by exercise relative to rest; fbs: Fasting blood sugar; thal: Thallium heart scan; ca: Number of major vessels; testbps: Resting blood pressure; chol: Serum cholesterol; age: age; thalach: Maximum heart rate achieved; exang: Exercise induced angina.

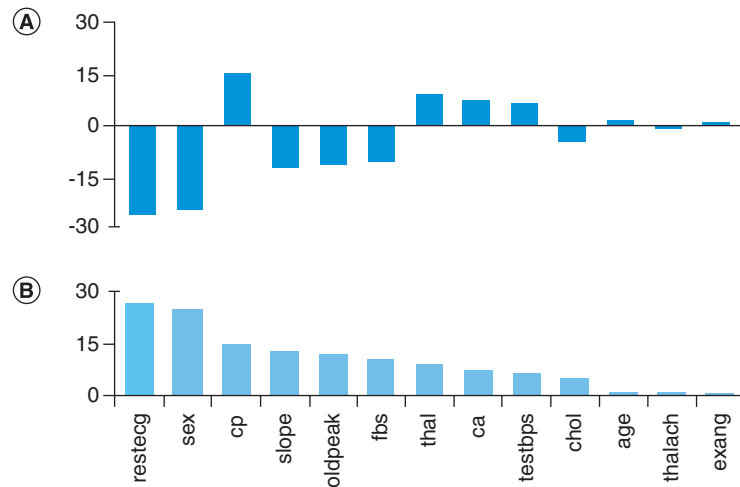
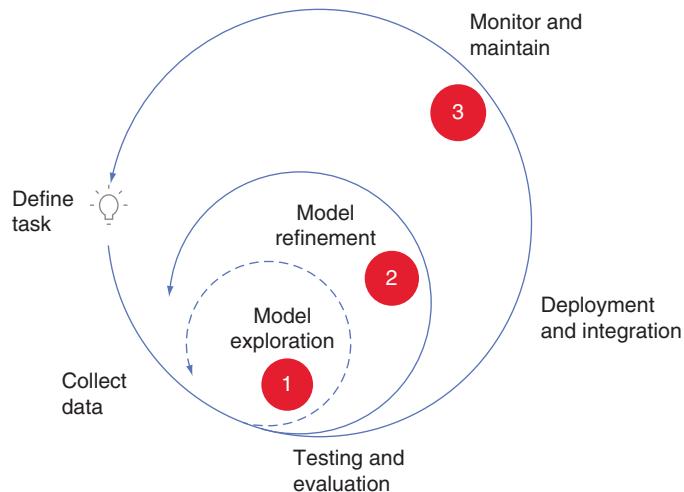


Figure 4. An ideal framework for developing a viable machine learning model. The dashed arrow indicates the iteration involving the machine learning algorithms presented here (also marked 1). Steps 2 and 3 (solid arrows) indicate future explorations. Modified with permission from [26].



ML models to detect CAD. We deposited the associated computer code in the public domain (see [23]) in the hope that we are contributing to an open source community.

Although CAD is both widely prevalent and may lead to fatal consequences, timely detection of CAD would empower clinicians to treat modifiable risk factors associated with the progression of CAD. Using an ML approach provides the ability to predict the presence of CAD with high accuracy and recall, and thus allows practitioners to practice preventative medicine in patients with CAD in a timelier manner. However, at such initial stages, it should be noted that ML serves solely as a predictor of CAD rather than a diagnostic tool. We hope that as more datasets are available for training the algorithm, we may be able to label ML algorithms as diagnostic steps in CAD management. Because machine learning utilizes datasets of patients who have already been diagnosed, the predictive ability of an ML algorithm for CAD would improve as more data are supplied to the algorithm. We visualize that a viable ML solution for predicting coronary artery disease (CAD) evolves in three steps: (1) model exploration; (2) model refinement; and (3) monitoring and maintenance. A proposed framework for arriving at a practical ML solution for the detection of CAD is depicted in Figure 4. We believe that with our current effort, we have achieved the first step (marked 1 in Figure 4). We are hopeful that this study can help form the basis for further testing/validating of our algorithms with multiple and larger datasets. We look forward to gain access to larger datasets for further validation and refinement (depicted as step 2 in Figure 4), with the eventual goal of providing an open source solution (step 3 in Figure 4) to aid healthcare practitioners in the detection and treatment of CAD.

Summary points

- Coronary artery disease (CAD) is the most common type of heart disease, with many modifiable risk factors.
- Early detection of CAD provides for management that would help prevent undesirable clinical outcomes. A comparative analysis of six machine learning algorithms was conducted, utilizing the UC Irvine Cleveland dataset to predict disease outcomes.
- The dataset was preprocessed, and six ML algorithms were trained and applied to the data.
- ML algorithms were evaluated based on accuracy, recall, F1 score and area under the curve. The six ML models performed well, with accuracies found to be greater than 0.79.
- The computer code generated in this study, which was developed in R, has been made available on the public repository GitHub and supplementally with this manuscript. ML algorithms can be applied with high accuracy and recall to detect the presence of CAD using a publicly available dataset.
- We envision that moving forward, a viable ML solution for predicting CAD evolves in three steps: (1) model exploration; (2) model refinement; and (3) monitoring and maintenance.

Supplementary data

To view the supplementary data that accompany this paper please visit the journal website at: www.future-science.com/doi/suppl/10.2144/fsoa-2020-0206

Author contributions

A Akella conceived the study; A Akella and S Akella planned the coding and analysis; A Akella wrote the script; A Akella and S Akella analyzed the results and prepared the manuscript.

Acknowledgments

The authors thank V Kaushik for help with the R programming and M Cassar for comments on the manuscript.

Financial & competing interests disclosure

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

No writing assistance was utilized in the production of this manuscript

Ethical conduct of research

The authors state that they have obtained appropriate institutional review board approval or have followed the principles outlined in the Declaration of Helsinki for all human or animal experimental investigations.

Open access

This work is licensed under the Creative Commons Attribution 4.0 License. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

References

1. GBD 2015 Mortality Causes of Death Collaborators. Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet* 388(10053), 1459–1544 (2016).
2. Sajda P. Machine learning for detection and diagnosis of disease. *Annu. Rev. Biomed. Eng.* 8, 537–565 (2006).
3. Foster KR, Koprowski R, Skufca JD. Machine learning, medical diagnosis, and biomedical engineering research-commentary. *Biomed. Eng. Online* 13(1), 94 (2014).
4. Deo RC. Machine learning in medicine. *Circulation* 132(20), 1920–1930 (2015).
5. Obermeyer Z, Ezekiel JE. Predicting the future – big data, machine learning, and clinical medicine. *N. Engl. J. Med.* 375(13), 1216–1219 (2016).
6. Tsay D, Patterson C. From machine learning to artificial intelligence applications in cardiac care: real-world examples in improving imaging and patient access. *Circulation* 138(22), 2569–2575 (2018).
7. Cuocolo R, Perillo T, De Rosa E, Ugga L, Petretta M. Current applications of big data and machine learning in cardiology. *J. Geriatr. Cardiol.* 16(8), 601–607 (2019).

8. Khan Y, Qamar U, Yousaf N, Khan A. Machine learning techniques for heart disease datasets. *Proceedings of the 2019 11th International Conference on Machine Learning and Computing – ICMLC '19* (2019). doi: 10.1145/3318299.3318343.
9. Alizadehsani R, Abdar M, Roshanzamir M *et al.* Machine learning-based coronary artery disease diagnosis: a comprehensive review. *Comput. Biol. Med.* 111, 103346 (2019).
10. Hajar R. Risk factors for coronary artery disease: historical perspectives. *Heart Views* 18(3), 109–114 (2017).
11. Goff DC *et al.* 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation* 129(25 Suppl. 2), S49–73 (2014).
12. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS ONE* 12(4), e0174944 (2017).
13. Kakadiaris IA, Vrigkas M, Yen AA, Kuznetsova T, Budoff M, Naghavi M. Machine learning outperforms ACC/AHA CVD risk calculator in MESA. *J. Am. Heart Assoc.* 7(22), e009476 (2018).
14. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: Applying machine learning to address analytic challenges. *Eur. Heart J.* 38(23), 1805–1814 (2017).
15. Kotsiantis SB. Supervised machine learning: a review of classification techniques. *Informatica* 160, 3–24 (2007).
16. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. Springer, New York, 112, 3–7 (2013).
17. Fawcett T. An introduction to ROC analysis. *Pattern Recognition Lett.* 27(8), 861–874 (2006).
18. Chawla NV, Bowyer KW, Hall LOH, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J. Artif. Int. Res.* 16, 321–357 (2002).
19. Lantz B. *Machine learning with R*. Packt Publishing (2013).
20. Heinze G, Wallisch C, Dunkler D. Variable selection – a review and recommendations for the practicing statistician. *Boim. J.* 60(3), 431449 (2018).
21. Heart Disease and Stroke Statistics Update Fact Sheet At-a-Glance. (2021). https://www.heart.org/-/media/phd-files-2/science-news/2/2021-heart-and-stroke-stat-update/2021_heart_disease_and_stroke_statistics_update_fact_sheet_at_a_glance.pdf?la=en
22. Heart Disease Data Set. <https://archive.ics.uci.edu/ml/datasets/Heart+Disease>
23. It should be noted that we noticed a second copy of the ‘Cleveland Dataset’ is available on the UCI repository with a different name as ‘Statlog (Heart) Dataset’ (see [24]) and several publications combine these two datasets for model building, or build on one dataset and validate on the other, which may lead to misleading predictions.
24. Statlog (Heart) Data Set. [http://archive.ics.uci.edu/ml/datasets/statlog+\(heart\)](http://archive.ics.uci.edu/ml/datasets/statlog+(heart))
25. R code for the UCI Cleveland dataset. <https://github.com/aa54/CAD-1>
26. Organizing machine learning projects: project management guidelines. <https://www.jeremyjordan.me/ml-projects-guide/>