ORIGINAL RESEARCH

Submitted: 2.08.2022; Accepted: 24.08.2022; Published online: 20.11.2022

Machine Learning-based Model for Early Prediction of Coronary Artery Disease

Nabeel Ahmad[1]*, Sudeept Singh Yadav[2], Alok Kumar Moharana[3]

[1]School of Allied Science, Dev Bhoomi Uttarakhand University, Dehradun, Uttarakhand, India
[2]Department of Computer Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India
[3]School of Health & Allied Science, ARKA JAIN University, Jharkhand, India

*Corresponding author:
soas.nabeel@dbuu.ac.in

Abstract

The global healthcare industry manages millions of individuals and generates enormous amounts of data. Machine learning-based algorithms are analysing complex medical information and produce superior insights. "Coronary artery disease (CAD)", the most prevalent form of cardiac disease is getting the greatest interest in the development of predictive models due to its large number of modifiable risk factors. This research study aims at comparing five algorithms of supervised machine learning for the CAD prediction. The research utilizes the Cleveland dataset from the UCI repository for training and testing the algorithms. The results of the comparison revealed that KNN is the best algorithm with significant performance measures which can be effective in predicting CAD accurately. Therefore, it can be suggested that these predictive models, which were developed using machine learning (ML) algorithms, can help doctors identify CAD early and may lead to better results that would help to avoid adverse clinical outcomes.

Keywords

Cardiovascular Disease (CVD), Coronary Artery Disease (CAD), Dataset, Prediction, Supervised Machine Learning

Imprint

Nabeel Ahmad, Sudeept Singh Yadav, Alok Kumar Moharana. Machine Learning-based Model for Early Prediction of Coronary Artery Disease. Cardiometry; Issue 24; November 2022; p. 373-378; DOI: 10.18137/cardiometry.2022.24.373378; Available from: http://www.cardiometry.net/issues/no24-november-2022/machine-learning-based-model

1. INTRODUCTION

Worldwide, each year, around 17 million people die from "cardiovascular disease (CVD)", as seen in Figure 1. Among all CVDs, Coronary artery disease which is abbreviated as CAD is one of the most prevalent causes of deaths as illustrated in Figure 2. The American Heart Association recently released figures showing that 13% of fatalities in the USA in 2018 were caused by coronary heart disease [1], [2]. In 2015,

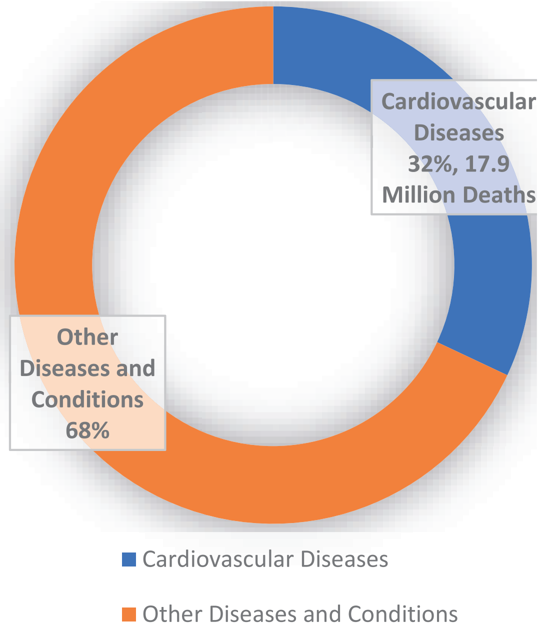

Figure 1: A Graphical Representation of Estimated Percentage of Deaths due to Cardiovascular and Other Diseases (WHO 2019).

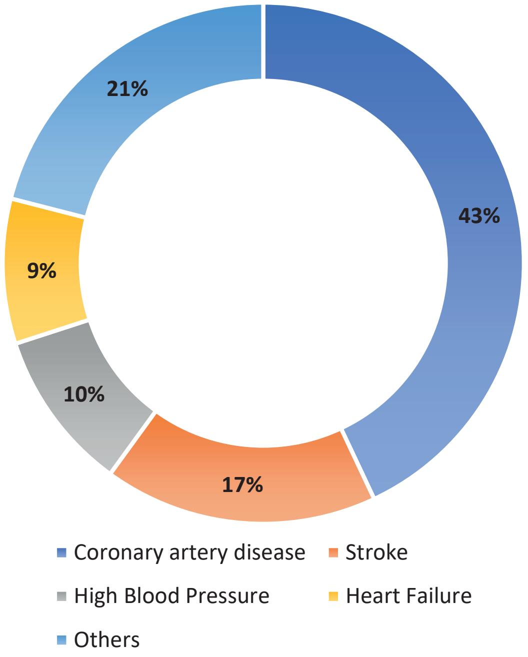

Figure 2: A Graphical Representation of Deaths caused by different Cardiovascular Diseases (CVDs).

CAD accounted for 15.6% of all mortality worldwide, placing it one of the major causes of death. Since this disease is associated with a variety of risk factors having modifiable nature that are associated with behaviour and lifestyle factors. Therefore, the timing of diagnosis and diagnostic performance is very crucial in the therapeutic management of CAD patients [3].

In CAD, plaques build up as a result of the hardening and narrowing of the arteries that provide blood to the muscles of heart [4]. This often happens as a result of the development of plaque inside the arteries, which contributes to a decrease in the quantity of oxygen and blood flow. CAD is more common in males than in women, and symptoms can occur in women 10 years later than in men [5].

As a result of the significant growth in cardiovascular diseases, which puts a significant financial burden on society, health organizations are attempting to develop a method for the accurate and early prediction of CAD utilizing modern statistical approaches, such as data mining. It is worth noting that the healthcare industry is brimming with data [6]. Unfortunately, the necessary data for successful decision-making and the detection of hidden patterns is not extracted. The causes of disease growth, incidence, or spread can be identified by extracting relevant data and uncovering knowledge from enormous volumes of medical data, and clinicians can be supplied with vital information for improved decision-making [7], [8]. As a result, many healthcare providers are looking for realistic solutions for knowledge discovery using Machine learning approaches. These strategies can assist in identifying disease trends and variables.

Machine Learning is used to study the determinants and predict people who are at risk of developing CVD [9], [10]. Machine learning approaches can analyze enormous amounts of data and uncover trends that humans might miss [11], [12]. It often improves efficiency and accuracy in the face of ever-increasing volumes of data being handled. It also enables immediate adaptability without the need for human involvement. Therefore, the present study aims at developing a model for the accurate prediction of CAD.

In order to predict CAD, this study compares multiple supervised algorithms, including “KNN”, “SVM”, “Naive Bayes”, “Decision trees”, and “Random Forest”. There are five sections throughout the entire paper. The importance of doing the study is introduced in the first section. A thorough analysis of the relevant work is provided in the second section. The study methodology is presented in the section 3, which is then trailed by the results in fourth section. The concluding statement is provided in the fifth and final section.

2. LITERATURE REVIEW

Abdar *et al.* developed a novel machine learning algorithm they had created for the accurate identification of CAD to patient data from Iran. Ten conventional ML algorithms were examined, and the top three performers (three varieties of SVM) were then employed in the remaining portions of the investigation. A “genetic algorithm”, “stratified 10-fold cross-validation”, and “particle swarm optimization” were also twice applied. According to the results, N2Genetic-nuSVM obtained an “F-1 value” of 91% and an “accuracy” of 93% when predicting CAD outcomes among the individuals who were participants of the well-established Z-Alizadeh Sani dataset [13].

Joloudari *et al.* suggested an integrated approach based on machine learning. In this study, learning techniques like the “C5.0 decision tree”, “the CHAID decision tree”, “support vector machines (SVM)”, and “random trees (RTs)” are employed. The investigation reveals that the RTs model outperformed other classification models, and the findings of the suggested strategy are encouraging [14].

Orphanou *et al.* used a temporal pattern mining approach to find the most prevalent temporal links among the developed basic “Temporal abstractions (TA)”. They built and evaluated classification methods based on the most prevalent TARs. All generated classifiers are utilised to identify CHD using a longitudinal dataset. The classification algorithm that makes use of the “horizontal support representation” and offers the highest performance was compared to a “Baseline Classifier” that uses the binary representation of the most common TARs. The results demonstrate that, in contrast to other classifiers, the horizontal support classifier performs considerably better than the baseline classifier [15].

Muhammat *et al.* created prediction models based on machine learning for CAD using diagnostic CAD datasets. The dataset was utilized to train algorithms, and the models were assessed using “receiver operating curve (ROC)”, “specificity”, “accuracy”, “sensitivity”, “specificity”, and other performance evaluation approaches. The machine learning model based on the random forest was found to be the best classification model with 92.04%, Regarding the accuracy, the sensitivity support vector machine-based machine learning model was found to be the best classification model with 87.34%, in regard with specificity, machine learning model based on Naive Baye was found to be the best classification model with 92.40%, and the ROC model was found to be the best model with 92.20% [16].

The above studies have developed different frameworks for the effective and accurate prediction of CAD. However, the present study carries out a comparative study to identify the most effective machine learning algorithm for the prediction of CAD.

3. METHODOLOGY

3.1. Design

The objective of the prediction approach is to establish a system that can infer characteristics of the predicted class from only a collection of different inputs. The purpose of machine learning in this study is to develop prediction models based on selected attributes or features or variables. Supervised classification algorithms are used in the present work to accurately predict CAD which will further enable the physician to give the best care feasible as soon as possible. AES encryption is used to protect patient information, which is then stored in databases (Figure 3).

3.2. Dataset Acquisition and Pre-processing

The "UCI Center for Machine Learning and Intelligent Systems (UCI; University of California, CA, USA)" maintains a database from which the dataset utilized in this work is retrieved. A total of four datasets from four various hospitals may be found in the database. While having more entries compared to the other datasets, the Cleveland dataset has fewer missing characteristics. This Cleveland dataset has a total of 14 variables on 303 patients.

In the next step called data preprocessing, rows having unknown values were eliminated resulting in all observations having values. In addition to that binary classification is established by replacing 1-4 in the column of CAD with values 1 for disease and 0

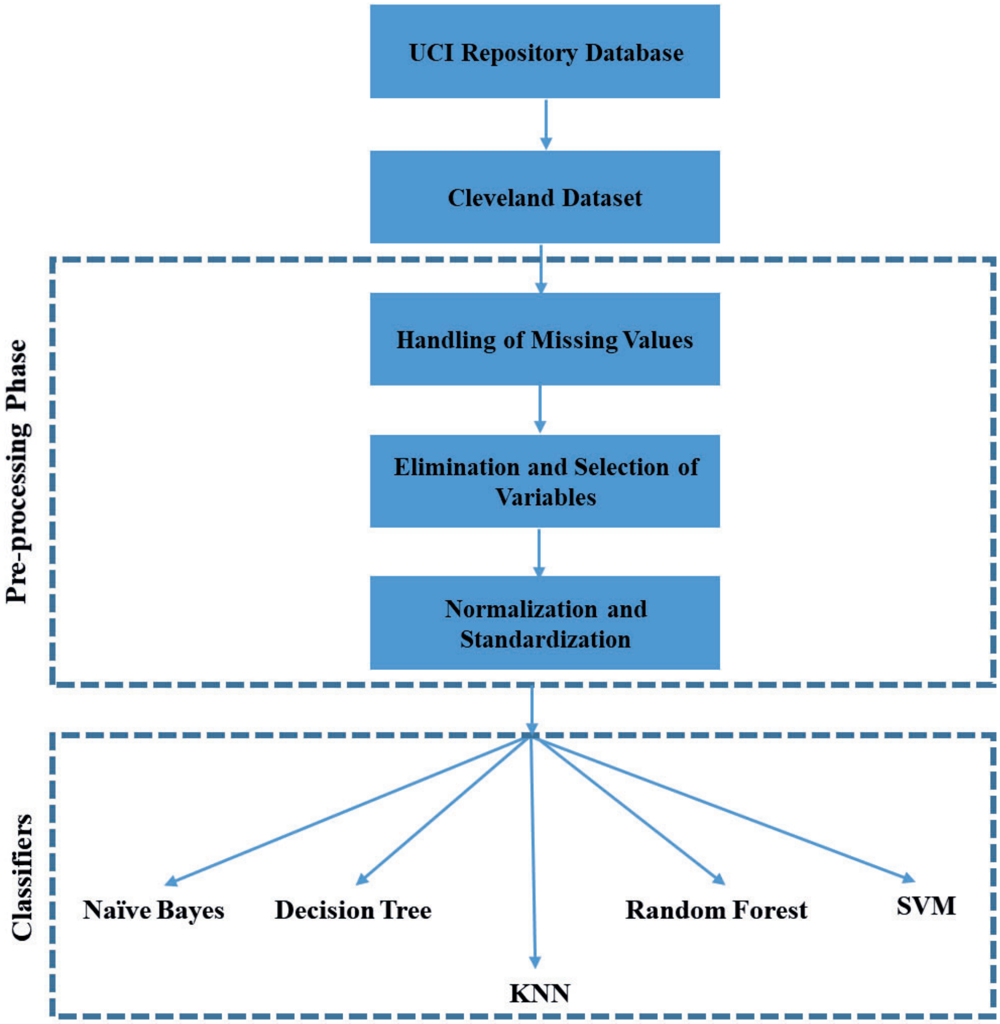

Figure 3: Illustrating the Developed Model for the Prediction of CAD.

for no disease, continuous variables were normalized and these all things are achieved by first converting the factor datatype into the numeric datatype. The Table 1 enlist all the variable, description and their types.

Table 1

Enlisting the names of the Variables, Description, and their types.

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

3.3. Instrumentation

When adopting models based on machine learning, it is widely acknowledged that no particular single approach is greater to the other classifiers. The learning is known to as «supervised» learning as contrasted to «unsupervised» learning, wherein instances are left unlabeled, while every instance in a dataset is delivered to the model in machine learning with labelled data (the associated accurate outputs), as in the "Cleveland dataset". The present research used different five supervised classifiers:

- "Naïve Bayes"
- "Decision tree"
- "KNN"
- "Random Forest"
- "SVM"

3.4. Data Analysis

To help make up for the absence of real-world data, the dataset is split into a «test set» (30% observations) and a «training set» (70% observations) before the analysis is conducted. Care is taken to balance the "class distributions" within the divide. The model has been trained using the «training» dataset, which the model uses to see and pick up new information. The resulting model fits the training dataset and is then evaluated objectively using the «test» dataset. In several instances, we ran manifold experimentations to authenticate model results using different splitting ratios. The performance matrix received using different supervised machine learning algorithms is provided in Tables below.

- Accuracy: the percentage or the proportion of all instances of the dataset that were accurately predicted out of total instances. The accuracy of the all four tested algorithms are given in Table 2.

"Accuracy = (true positives + true negatives)/total"

Table 2

Enlisting the Accuracy of all Tested Machine Learning algorithms.

Classifiers	Naïve Bayes	Decision tree	KNN	Random Forest	SVM
Accuracy %	70	76	85	66	69

- Precision: the proportion of all actual positive instances from all positive predictions. The precision of the all four tested algorithms are given in Table 3.

"Precision = True Positive /
(True Positive + False Positive)"

Table 3

Enlisting the Precision of all Tested Machine Learning algorithms.

Classifiers	Naïve Bayes	Decision tree	KNN	Random Forest	SVM
Precision 0	69	72	85	66	66
Precision 1	72	76	85	66	73

- Recall: Also known as sensitivity is the proportion of all positive instances of the dataset out of all positive predictions. The recall values of the all four tested algorithms are given in Table 4.

"Sensitivity = true positives/
(true positives + false negatives)"

Table 4

Enlisting the Recall of all Tested Machine Learning algorithms.

Classifiers	Naïve Bayes	Decision tree	KNN	Random Forest	SVM
Recall 0	98	80	78	66	79
Recall 1	54	69	69	66	61

- F1 Score: a composite harmonic mean which combines both the recall and precision. The F-1 Score values of the all four tested algorithms are given in Table 5.

“F 1 = 2 × (precision × recall) / (precision + recall)”

Table 4

Enlisting the F-1 Score values of all Tested Machine Learning algorithms.

Classifiers	Naïve Bayes	Decision tree	KNN	Random Forest	SVM
F-1 Score 0	67	72	78	74	72
F-1 Score 1	67	72	78	75	72

4. RESULTS AND DISCUSSION

The actual flow which was used to carry out the research is presented in Figure 3. As shown in Figure 1 the research concentrated on comparing various supervised machine learning algorithms in correctly predicting CAD in patients. The study utilized a usual spit of dataset into 70/30 for training and testing. Furthermore, the dataset is checked with various supervised machine learning algorithms such as “Naïve-Bayes”, “SVM”, “KNN”, “Random Forest”, and “Decision tree”. In addition to that, the confusion matrix which is the foundation to measure all performance parameters was used.

The results demonstrated that KNN has the best performance matrices including accuracy, precision, recall and F-1 Score. Therefore, the results of the our research suggest that KNN is the best Machine learning algorithm which is superior in predicting CAD as a whole in comparison to other four tested algorithms.

5. CONCLUSION

In this study, we showed that using a publicly available dataset, ML algorithms can accurately and reliably identify the existence of CAD. Even though CAD is widespread and may have fatal consequences, an early diagnosis would allow clinicians to tackle modifiable risk factors connected to its development. Using an ML approach allows medical professionals to manage CAD patients proactively with preventive treatments since it can predict the presence of CAD with high recall and accuracy. The fact that ML is simply a predictor of CAD at this early stage should not be disregarded. It is not a diagnostic tool. With more datasets accessible to train the machine learning algorithm on, we anticipate being capable of referring to ML algorithms as “diagnostic tool” in managing the CAD. “Machine learning” employs datasets of patients who have previously been diagnosed, thus as more information are provided to the algorithms used to predict CAD, their predictive ability will increase.

References

1. G. A. Roth et al., “Global, Regional, and National Burden of Cardiovascular Diseases for 10 Causes, 1990 to 2015,” J. Am. Coll. Cardiol., vol. 70, no. 1, pp. 1–25, Jul. 2017, doi: 10.1016/j.jacc.2017.04.052.

2. S. Ali et al., “The burden of cardiovascular diseases in Ethiopia from 1990 to 2017: evidence from the Global Burden of Disease Study,” Int. Health, vol. 13, no. 4, pp. 318–326, Jul. 2021, doi: 10.1093/inthealth/ihaa069.

3. E. A. Leonard and R. J. Marshall, “Cardiovascular Disease in Women,” Prim. Care Clin. Off. Pract., vol. 45, no. 1, pp. 131–141, Mar. 2018, doi: 10.1016/j.pop.2017.10.004.

4. J. A. Rymer and S. V. Rao, “Anemia and coronary artery disease,” Coron. Artery Dis., vol. 29, no. 2, pp. 161–167, Mar. 2018, doi: 10.1097/MCA.0000000000000598.

5. R. Bauersachs, U. Zeymer, J.-B. Brière, C. Marre, K. Bowrin, and M. Huelsebeck, “Burden of Coronary Artery Disease and Peripheral Artery Disease: A Literature Review,” Cardiovasc. Ther., vol. 2019, pp. 1–9, Nov. 2019, doi: 10.1155/2019/8295054.

6. I. Y. Chen, E. Pierson, S. Rose, S. Joshi, K. Ferryman, and M. Ghassemi, “Ethical Machine Learning in Healthcare,” Annu. Rev. Biomed. Data Sci., vol. 4, no. 1, pp. 123–144, Jul. 2021, doi: 10.1146/annurev-biodatasci-092820-114757.

7. P. Mathur, S. Srivastava, X. Xu, and J. L. Mehta, “Artificial Intelligence, Machine Learning, and Cardiovascular Disease,” Clinical Medicine Insights: Cardiology, vol. 14. p. 117954682092740, Jan. 09, 2020. doi: 10.1177/1179546820927404.

8. S. J. Al’Aref et al., “Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging,” Eur. Heart J., vol. 40, no.

24, pp. 1975–1986, Jun. 2019, doi: 10.1093/eurheartj/ehy404.

9. R. Bonetto and V. Latzko, “Machine learning,” in Computing in Communication Networks, Elsevier, pp. 135–167 2020. doi: 10.1016/B978-0-12-820488-7.00021-9.

10. C. Janiesch, P. Zschech, and K. Heinrich, “Machine learning and deep learning,” Electron. Mark., vol. 31, no. 3, pp. 685–695, Sep. 2021, doi: 10.1007/s12525-021-00475-2.

11. P. Lapadula, G. Mecca, D. Santoro, L. Solimando, and E. Veltri, “Greg, ML – Machine Learning for Healthcare at a Scale,” Health Technol. (Berl)., vol. 10. pp. 1485-1495, 2020, doi: 10.1007/s12553-020-00468-9.

12. I. Y. Chen, S. Joshi, M. Ghassemi, and R. Ranganath, “Probabilistic Machine Learning for Healthcare,” Annu. Rev. Biomed. Data Sci., vol. 4, pp. 393-415, 2021, doi: 10.1146/annurev-biodatasci-092820-033938.

13. M. Abdar, W. Książek, U. R. Acharya, R. S. Tan, V. Makarenkov, and P. Pławiak, “A new machine learning technique for an accurate diagnosis of coronary artery disease,” Comput. Methods Programs Biomed., vol. 179, p. 104992, 2019, doi: 10.1016/j.cmpb.2019.104992.

14. J. H. Joloudari et al., “Coronary artery disease diagnosis; ranking the significant features using a random trees model,” Int. J. Environ. Res. Public Health, vol. 17, no. 3, p. 731, Jan. 2020, doi: 10.3390/ijerph17030731.

15. K. Orphanou, A. Dagliati, L. Sacchi, A. Stassopoulou, E. Keravnou, and R. Bellazzi, “Incorporating repeating temporal association rules in Naïve Bayes classifiers for coronary heart disease diagnosis,” J. Biomed. Inform., vol. 81, pp. 74–82, May 2018, doi: 10.1016/j.jbi.2018.03.002.

16. L. J. Muhammad, I. Al-Shourbaji, A. A. Haruna, I. A. Mohammed, A. Ahmad, and M. B. Jibrin, “Machine Learning Predictive Models for Coronary Artery Disease,” SN Comput. Sci., vol. 2, no. 5, p. 350, Sep. 2021, doi: 10.1007/s42979-021-00731-4.