HIGH SCHOOL EDITION
Journal of Student Research
Volume 12 Issue 4 (2023)

Diagnosis of Coronary Artery Disease Using Adult Data from Blood Tests and Electrocardiograms

Anika Pallapothu

The Harker School, USA

ABSTRACT

Currently, more than 18 million adults are reported to have coronary artery disease (CAD). In 2019, around 360,900 CAD-related deaths were reported in the United States alone. CAD is caused by the deposition of cholesterol in the coronary arteries, followed by inflammation. Coronary arteries form an essential part of the heart, responsible for the circulation of oxygen and nutrients. Blockage in these major vessels due to cholesterol deposition is one of the major causes of heart attacks and strokes. The two major symptoms of CAD are chest pain and shortness of breath. However, most patients can remain asymptomatic even after decades of having CAD. Diagnostic tests for CAD include blood tests, exercise (cardiac) stress tests, and echocardiogram, which are performed only after the presentation of symptoms. In this study, various machine learning (ML) algorithms have been tested on the compiled dataset of UCI heart disease for predicting CAD. A hypothesis testing approach was used that retained only the most significant features. Comparative analysis was done using various evaluation metrics with support vector machines (SVM). With a recall score of 0.9, the SVM outperformed other models with 90% accuracy.The identification of the underlying causes of the disease is possible with further research for better preventative measures.

Introduction

Coronary artery disease (CAD) is a common heart disease predominantly caused by the accumulation of plaques in the coronary arteries. Oxygen is circulated to the heart majorly *via* the coronary arteries. The blood flow to the heart is prevented if there is any blockage in these arteries. According to WHO, CAD is one of the main causes of death worldwide (World Health Organization, 2022). Poor lifestyle, physical lethargy, and age-related factors are among the major risk factors for CAD. In today's digital world, where most data is recorded and digitalized, it is easier for analytical tools to derive meaningful insights from medical data. This reduces the stress on doctors and allows them to treat patients more effectively.

The application of machine learning (ML) and deep learning (DL) has assisted in faster diagnosis of patients in the medical sector. The abundantly available structured and unstructured data can be utilized through different approaches to extract valuable information from it. ML has aided in optimizing businesses by utilizing cases from multiple fields. ML algorithms can acquire vital information from relevant databases and identify patterns that can be incorporated into real-time data to predict and automate processes. ML has seen massive utility in healthcare industries for diagnosing and predicting patients' health issues. Various studies have used ML algorithms in the medical industry to predict heart diseases precisely. Dun et al. (2016) applied ML algorithms to anticipate the prevalence of diseased heart conditions using medical and demographic information. The algorithm examined the patient's data collected over time and then derived patterns for data extraction from it.

ML algorithms are further categorized into two types, Supervised and Unsupervised Learning. Supervised learning is a division of ML that involves learning from pairs of inputs and predetermined outputs (Liu & Wu, 2012). Algorithms generate patterns comparing the given input and generated output, optimize based

on errors, and learn from it. Dangare et al. (2012) utilized distinct additional features, such as obesity and smoking, to create robust predictive models of coronary disease. Decision trees, naïve Bayes, and neural network-based models were applied, which resulted in an accuracy of 99.62 %, 90.74 %, and 100 %, respectively. Feature engineering, a pre-processing step of the ML approach, was applied to remove features that need the assistance of a medical professional. Several algorithms were applied, along with adjusting hyperparameters and ensemble techniques to design the model. A test set achieved 78.3% maximum accuracy following this algorithm. On the contrary, unsupervised learning tends to uncover patterns between the given inputs, not knowing about the output (Thabtah & Hammoud, 2013). Since the present study aims for the prediction of CAD using predetermined outcomes, a supervised learning approach was considered.

Karthiga et al. (2017) developed an ML model for predicting heart conditions using a dataset of 573 records. MATLAB tool was applied with missing values filter to pre-process the data. Thereafter, decision trees and Naïve Bayes algorithms were applied to design prediction models. The decision trees were found to outperform the Naïve Bayes algorithm as observed from the test set accuracy. Gonsalves et al.(2019) used the South African Heart database to analyze different ML models and determine the most effective model. They included 462 records for predicting a diseased heart condition. Naïve Bayes, SVM, and DT were applied as predictive models. A 10-fold cross-validation technique was applied, and Naïve Bayes outperformed other models in diagnosing Coronary Heart Disease (CHD).

The UCI ML heart disease repository (UCI Heart Disease Dataset) was used by Bharti et al. (2021) for a comparative analysis of different methods to design distinct ML models. Normalization of data was done, and the Isolation Forest algorithm was used for feature selection. K-Neighbor ML algorithm rendered an accuracy of 83.29%, while DL algorithm showed 94.2% accuracy. The Cleveland dataset is a constituent of the UCI Heart Disease Dataset, which can be used for comparative analysis using distinct methods to predict model accuracy (Akella & Akella, 2021). Decision trees and logistic regression were used in the initial models. However, these were overfitting and provided significantly less accuracy. Synthetic data was produced using the SMOTE (Synthetic Minority Oversampling Technique) methodology and later deleted to conserve the authenticity of the dataset distribution. Neural networks attained an accuracy of 93%, which was reported to have the highest accuracy among the different ML models applied.

In our study, the Cleveland Heart Disease Dataset was employed. About 303 patient records with 14 distinct features were collected over time using various test approaches. Our objective was to develop predictive models using multiple ML algorithms such as K-nearest neighbor (KNN), Support vector machines (SVM), logistic regression, Random Forests (RFs), decision trees (DTs), adaptive boosting (AdaBoost). We refined our models, validated the results across different folds, and finally tested our model on a standout test set. While developing the model, feature engineering was conducted using hypothesis testing to choose only the significant attributes.

Materials and Methods

Description of Datasets

The datasets used in this research were obtained from UCI Machine Learning (ML) Repository. The Cleveland dataset was utilized in our study, which comprised 303 records with 76 features. Since existing literature has used 14 of these features, the current also includes these features.

Table 1. Dataset description

Name	Type	Description
Age	Continuous	Age of patients (in years)
Sex	Categorical	Gender of a patient (male or female)
Chest pain	Categorical	Chest pain type (typical angina, atypical angina, non-anginal pain, no pain)
Trestbps	Continuous	Resting blood pressure
Chol	Continuous	Serum cholesterol in mg/dL
FBS	Categorical	Fasting blood sugar (Yes or No)
Restecg	Categorical	Resting electrocardiogram (normal, wave abnormality, probable/definite left ventricular hypertrophy)
Thalach	Continuous	Maximum heart rate achieved
Exang	Categorical	Exercise-induced angina (Yes or 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 (normal, fixed defect, reversible defect)
Num	Categorical	Diagnosis of heart disease (angiographic disease status) (0 = absent; 1 to 4 = present)

Data Preprocessing

The first step was determining if the data contained null values or duplicate records. Six records were deleted as their values were unknown. All the categorical attributes were encoded to their respective integer values. The target variable was then changed to a binary distribution. Values higher than “zero” denoted the presence or absence of heart disease and “ones” signifying the opposite.

Exploratory Data Analysis

Univariate and bivariate analyses of some features were conducted to determine the data distribution and examine the correlation of the features to the target variable. A bivariate analysis was also conducted to determine the relationship between various types of chest pain and heart disease.

Feature Selection

The model was developed by hypothesis testing while selecting only significant features. The findings were summarized by generalized linear model approach using the Statsmodel library. A p-value is essential for determining whether or not the null hypothesis needs to be accepted. The null hypothesis is accepted if the p-value is higher than the significant level; otherwise, it is rejected. The p-value from the Statsmodel was used for each variable checked if it was significant.

Model Building

The data was categorized into train and test sets, with 20% data specifically designated for testing. The model's effectiveness was determined by reporting all the evaluation metrics. The data was normalized and employed for a five-fold cross-validation approach for modifying the model hyperparameters.

A logistic regression model was adjusted for the regularization approach and penalty strength hyperparameters. It determined the overall test accuracy to be 83.3% and values of precision, recall, and AUC score as 0.843, 0.833, and 0.929, respectively. Then, a predictive model was developed using the DT algorithm in which adjustments were made using maximum features, maximum depth, complexity parameter, and criterion. The overall accuracy of 83.3% was obtained on test data with values of 0.8371, 0,833, and 0.93 for precision, recall, and AUC scores, respectively. Thereafter RF was used to build DT as the go-to algorithm. The hyperparameters such as maximum depth, minimum samples leaf, minimum samples split, number of estimators, and criterion were considered for adjusting the model. We obtained 88.3% accuracy and 0.883, 0.883, and 0.942 values of precision, recall, and AUC scores, respectively, on the test set. The model was fine-tuned by using SVM with radial basis function as the kernel and taking gamma and regularization as the hyperparameters. Accuracy of 90% was achieved on the test set with SVM, and values of 0.9, 0.9, and 0.95 of precision, recall, and AUC scores were obtained, respectively. KNN algorithm was used by modifying the number of neighbors, weight function, and distance metric hyperparameters. Overall test accuracy of 88.3% and 0.885, 0.883, and 0.938 values of precision, recall, and AUC scores were achieved. Finally, the AdaBoost algorithm was applied, and the number of estimators, learning rate, and maximum depth were used for hyperparameter adjustments. An accuracy of 85% on the test set and 0.856, 0.85, and 0.919 scores of precision, recall, and AUC score were achieved, respectively.

Figure 1 shows patients with asymptomatic chest pain having a greater probability of heart disease than other signs of chest pain. Figure 2 illustrates the age of patients with chances of having heart disease, where elder patients above 55 have a higher probability of having heart disease than younger patients. Care was taken to check the misrepresentation of the data for the patients' distribution with the presence or absence of heart disease. Figure 3 shows the number of normal classes higher than abnormal classes with some margin. Hence, data misrepresentation was ruled out.

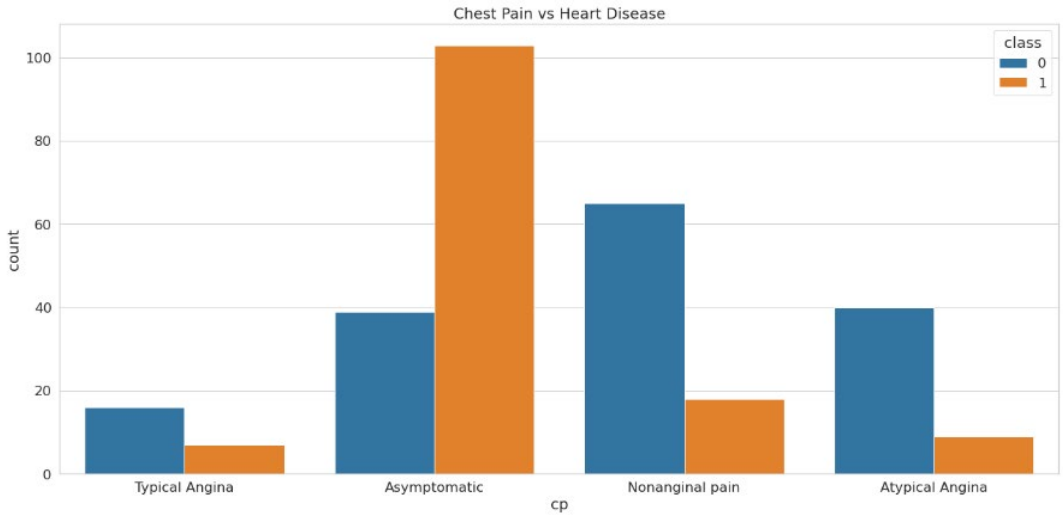

Figure 1. Chest pain vs. heart disease**.**

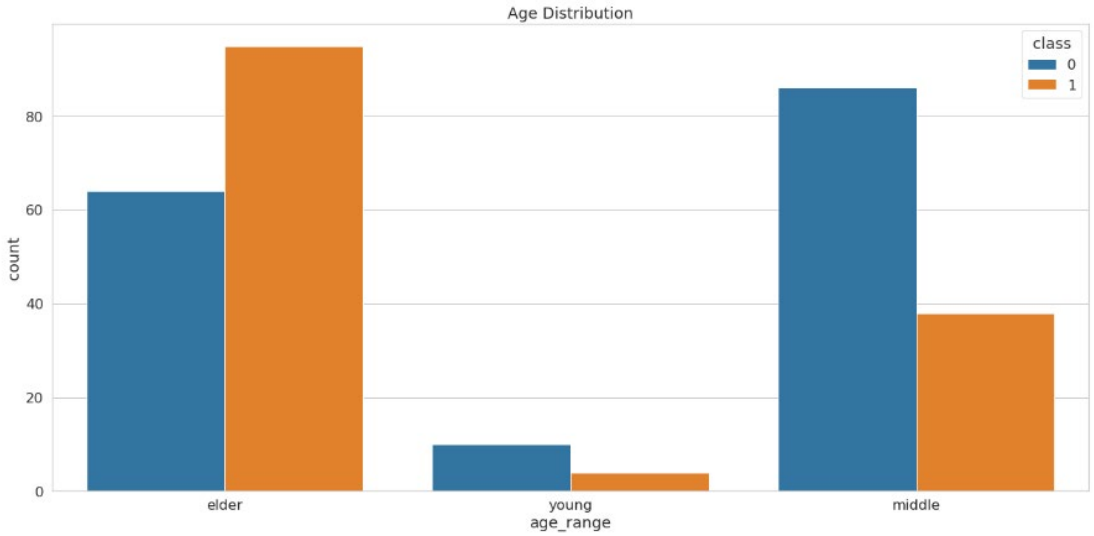

Figure 2. Age vs. heart disease**.**

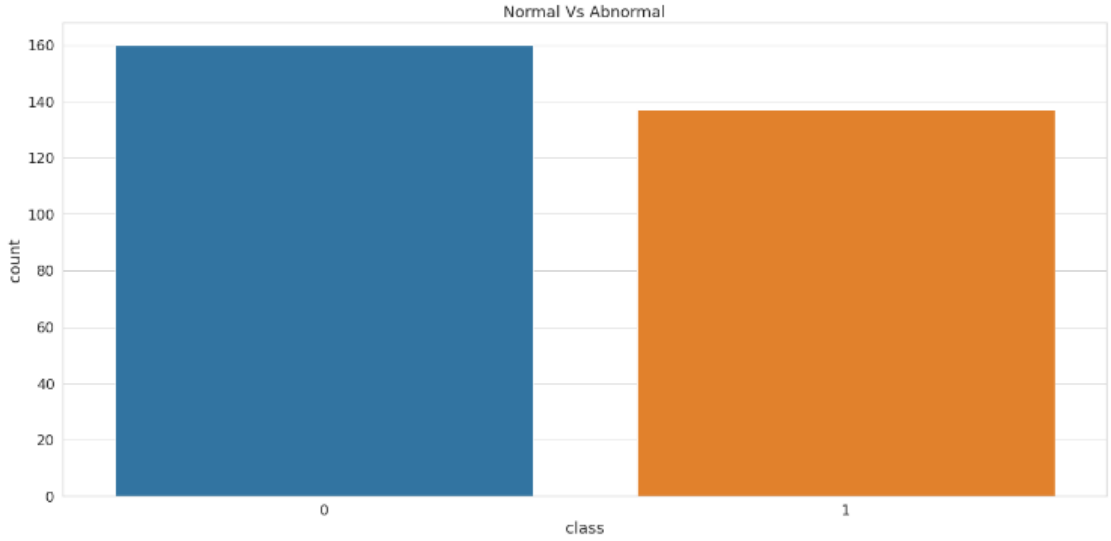

Figure 3. Normal vs. abnormal distribution.

Table 2. Performance scores of ML models

Algorithm	Accuracy	Precision	Recall	AUC score
Logistic Regression	83.3%	0.843	0.833	0.929
DT	83.3%	0.837	0.833	0.93
RF	88.3%	0.883	0.883	0.942
SVM	90%	0.9	0.9	0.95
KNN	88.3%	0.885	0.883	0.938
AdaBoost	85%	0.856	0.85	0.919

Results

A strong predictive model was developed from the UCI heart disease dataset by using 6 different types of ML models. The supervised ML approach was used to build a predictive model. All the experiments were performed using Python programming. Table 2 lists the results of the experiment. It shows that all models perform reasonably well rendering more than 80% accuracy. It was observed that SVM outperformed other models with 90% accuracy with a recall score of 0.9. Our major focus was to reduce false negative outcomes. Since high recall value indicates a report of fewer false negative cases, SVMs were chosen as the go-to model for the prediction of CAD.

Discussion

This study aimed to design a robust machine-learning framework capable of automating the prediction of coronary artery disease. We used the UCI heart disease dataset and restricted it to 14 attributes. All the data were pre-processed by applying six different machine-learning algorithms. The performance of each model was evaluated using distinct scores viz., accuracy, precision, recall, and AUC. Out of all models, support vector machines rendered the best results, with 90% accuracy on the test set and a 0.9 recall score.

Conclusion

The present research can be elaborated by incorporating more features and patient records to design a more robust model. The patient's data can be recorded over time and create a sequential model to predict CAD with better accuracy. Along with prediction, further research can lead to the identification of the underlying cause of the disease so that better preventative measures can be taken.

References

Akella, A., & Akella, S. (2021). Machine learning algorithms for predicting coronary artery disease: efforts toward an open source solution. *Future Science OA*, *7*(6), FSO698. https://doi.org/10.2144/fsoa-2020-0206

Bharti, R., Khamparia, A., Shabaz, M., Dhiman, G., Pande, S., & Singh, P. (2021). Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning. *Computational Intelligence and Neuroscience*, *2021*(Special Issue), 1–11. https://doi.org/10.1155/2021/8387680

Dangare, C., Apte, S., & Student, M. (2012). Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques. *International Journal of Computer Applications*, *47*(10), 975–888. https://doi.org/10.5120/7228-0076

Dun, B., Wang, E., & Majumder, S. (2016). *Heart Disease Diagnosis on Medical Data Using Ensemble Learning*. https://cs229.stanford.edu/proj2017/final-reports/5233515.pdf

Gonsalves, A. H., Thabtah, F., Mohammad, R. M. A., & Singh, G. (2019). Prediction of Coronary Heart Disease using Machine Learning. *Proceedings of the 2019 3rd International Conference on Deep Learning Technologies - ICDLT 2019*, 51–56. https://doi.org/10.1145/3342999.3343015

Karthiga, A., Safish Mary, M., Yogasini, M., & Scholar, M. (2017). Early Prediction of Heart Disease Using Decision Tree Algorithm. *International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)*, *3*(3), 2395–2695. https://www.ijarbest.com/journal/v3i3/969

Liu, Q., & Wu, Y. (2012). Supervised Learning. *Encyclopedia of the Sciences of Learning*, 3243–3245. https://doi.org/10.1007/978-1-4419-1428-6_451

Thabtah, F., & Hammoud, S. (2013). MR-ARM: A Map-Reduce Association Rule Mining Framework. *Parallel Processing Letters*, *23*(03), 1350012. https://doi.org/10.1142/s0129626413500126

World Health Organization. (2022). *Cardiovascular Diseases*. World Health Organization. https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1