PREDICTION OF CAD USING ELECTROCARDIOGRAPHY

Harinee Rathod (20BEC101) [1], Rohan Sarvaiya (20BEC104) [2]

Department of Electronics and Communication Engineering, Institute of Technology Nirma University
Sarkhej - Gandhinagar Highway, Gujarat ; Pincode, 382470

***Abstract*— This paper presents a machine learning approach to predict coronary artery disease (CAD) using electrocardiography (ECG). CAD is a common and serious cardiovascular disease that affects millions of people worldwide. Early detection and accurate diagnosis of CAD are crucial for effective treatment and prevention of consequences. In this paper, we collected ECG data and used machine learning algorithms to develop a predictive model. Our results show that the proposed model achieved high accuracy and sensitivity in detecting CAD. The model also identified key ECG features that are strongly associated with CAD, that provide valuable insights into the pathophysiology of the disease. We did analysis of the CAD dataset using logistic regression, Naive Bayes, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Adaboost Classifier.**

***Keywords*— CAD, model, accuracy, sensitivity, pathophysiology**

I. Introduction

The most common cause of mortality in today's world is coronary artery disease (CAD). Coronary artery disease (CAD) is a blockage of coronary arteries due to formation of a waxy substance called plaque that builds up inside the coronary arteries. These arteries are responsible for supplying oxygen rich blood to the heart muscle. The building up of Plaque limits the amount of blood that can reach your heart. The most prevalent kind of cardiac illness is coronary artery disease (CAD).Even though it may take decades for the blood flow to be blocked, it could lead to a heart attack or a heart failure.

An effective and extensively used non-invasive diagnostic technique for detecting CAD is Electrocardiography (ECG). However, evaluating ECG readings can be tricky and calls for specific expertise. Machine learning techniques have been utilized recently to predict CAD using ECG data. Large datasets can be mined for patterns and analysis by machine learning algorithms, which enables the creation of precise and automated prediction models.

The ECG signal consists of a series of waves, each representing a specific part of the cardiac cycle.The QRS complex is a waveform observed on an electrocardiogram (ECG). The P wave indicates atrial depolarization, the QRS complex indicates ventricular depolarization, and the T wave indicates ventricular repolarization. The duration, amplitude, and shape of these waves can provide information about the health and function of the heart.

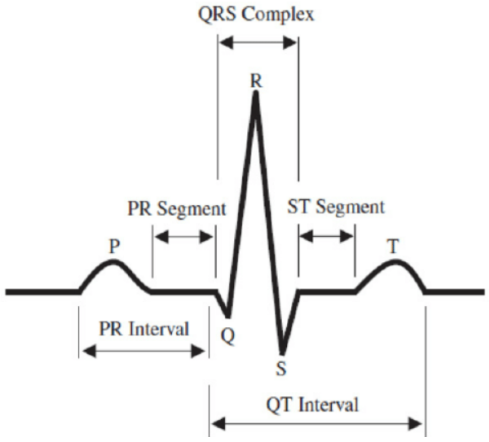

Fig. 1 ECG Signal

II. Literature review

The electrocardiography (ECG) data used by [1] Mohammadpour et al., to develop a deep-learning model for diagnosing heart disease. Three convolutional layers, three max-pooling layers, and three fully connected layers are constituents of the model. The model outperformed the available techniques for classifying ECG signals. The study showed how deep learning algorithms might be used to increase the precision and effectiveness of heart disease detection using electrocardiogram (ECG) signals.[2] J. Ma et al. proposed a machine learning-based framework to predict cardiovascular disease (CVD) risk using electronic health records. They used classification models such as random forest, support vector machine, and neural networks. Feature selection technique was performed to identify the most relevant risk factors. This framework can assist in identifying those who are at high risk and perhaps enhance CVD prevention and management.

The research by [3] S. S. Razavian et al., propose a machine learning model for predicting the onset of heart failure using

electronic health records. The model uses gradient boosting algorithms and feature selection techniques to identify the most relevant features. The research outperforms the traditional models. The proposed model has the potential to assist clinicians in early diagnosis and intervention for heart failure. [4] Jindal et al. proposed a machine learning-based approach to predict heart disease using general health information. The study utilized the Random Forest (RF) and Support Vector Machine (SVM) algorithms to classify heart disease. The results showed that the RF algorithm outperformed SVM. The study also highlighted the importance of feature selection to improve the classification accuracy. A machine learning-based prediction model for cardiovascular disease (CVD) applying feature selection was proposed [5] by N.K. Singh and K. Singh. They used the support vector machine algorithm to classify the CVD data and data mining techniques to analyze it. [6] Phadke et al. proposed a machine learning-based approach to predict coronary artery disease (CAD) using electrocardiography (ECG) signals.

[7] D. Dey and J. Mukhopadhyay presented a review on electrocardiogram (ECG) signal analysis for automatic detection of coronary artery disease (CAD). The paper covers various aspects of ECG-based CAD detection, including the anatomy of coronary arteries, ECG waveform and its features, signal preprocessing, feature extraction, classification algorithms, and performance evaluation metrics.Basically it is a review paper and the authors have compared and analyzed various aspects of ECG signal. [8] Venkataramanan et al. propose a hybrid feature selection approach for ECG-based diagnosis of coronary artery disease (CAD) using machine learning techniques. They extracted various features from ECG signals, and selected the most discriminative features using a combination of genetic algorithm and correlation-based feature selection. [9] A. Subasi and M. Alshawi presented automated CAD detection techniques using different duration of ECG segments using CNN. [10] Ibrahim et al. presented Performance Analysis of ECG-Based Feature Extraction Techniques for CAD classification.

Reference	Algorithm	Dataset	Year	Accuracy
"A Deep Learning Approach for Heart Disease Diagnosis Using ECG Signals"	CNN and LSTM	PTB Diagnostic ECG Database	2019	97.08%
"A Machine Learning-Based Framework for the Prediction of Cardiovascular Disease Risk"	Logistic Regression and Feature Selection	National Health and Nutrition Examination Survey (NHANES)	2021	84.7%
"A Machine Learning Model for Predicting the Onset of Heart Failure Using Electronic Health Records"	Random Forest	Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) III	2018	91.4%
"Heart Disease Prediction using Machine Learning Techniques with the Help of General Health Information"	K-Nearest Neighbour, Decision Trees, Random Forests	Cleveland Heart Disease dataset	2019	90.5%
"Cardiovascular Disease Prediction Model Using Machine Learning Techniques with Feature Selection"	Decision Trees, Random Forests, and Feature Selection	Framingham Heart Study dataset	2018	-
"Prediction of Coronary Artery Disease using Electrocardiography: A Machine Learning Approach"	XGBoost, Logistic Regression,Naive Bayes Classifier	PTB Diagnostic ECG Database	2019	88.79%

"ECG signal analysis for automatic detection of coronary artery disease:A review"	-	PTB Diagnostic ECG Database, the MIT-BIH Arrhythmia Database, and the Physionet 2017 Challenge Database	2019	-
"ECG Based Diagnosis of Coronary Artery Disease Using Hybrid Feature Selection and Machine Learning Techniques"	random forest (RF) and k-NN ,PCA ,recursive feature eliminatio n (RFE)	PTB Diagnostic ECG Database	2020	93.6%
"Automated detection of coronary artery disease using different durations of ECG segments with convolutiona l neural network,"	CNN	PTB Diagnostic ECG Database	2021	96.5%
"Performanc e Analysis of ECG-Based Feature Extraction Techniques for Coronary Artery Disease Classificatio n"	Deep Belief Network (DBN)	PTB Diagnostic ECG Database	2021	94.5%

Table 1. Literature Survey

III. Methodology

The goal of this project is to use data from the UCI Machine Learning Repository to predict the development of coronary artery disease (CAD). Early detection of CAD may significantly reduce the risk of fatal heart attacks and other cardiovascular problems. Previous research has demonstrated that machine learning algorithms are capable of accurately forecasting the development of CAD. To predict the possibility of developing CAD in this study, we have applied logistic regression, Naive Bayes, K-Nearest Neighbours (KNN), Support Vector Machine (SVM), and the AdaBoost classifiers based on patient characteristics like age, sex, kind of chest pain, blood pressure, and others.

A. Dataset Analysis

The UCI Heart Disease dataset includes data on 303 patients who had heart disease testing. The dataset is freely accessible and was collected from the UCI Machine Learning Repository.

The dataset has a total of 13 features, including age, sex, chest pain type, blood pressure, cholesterol level, fasting blood sugar level, rest electrocardiogram results, maximum heart rate achieved, exercise-induced angina, ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels colored by fluoroscopy, and thalassemia. Different methods to analyze the correlation between features, data outliers, the relation between different feature frequencies of the data, and the relation between the target and different features.

B. Data Preprocessing

Before applying the dataset for modeling, we preprocessed it in certain ways. We started by looking for any null or missing data and verified that the dataset was complete. We also

normalized the data to have a standard deviation of one and a mean of zero.

C. Model Building

For model building the following classifier algorithms are used.

- *Logistic regression:* logistic regression is most appropriate for the binary classification issues where the target variable has two possible outputs 1 and 0. It is effective with categorical or continuous input variables. Although it requires a linear relationship between the input variables and the output variable's log probabilities.
- *Naive Bayes Classifier:* Naive Bayes is a probabilistic approach for classification and prediction issues that is based on the Bayes theorem. It assumes that the features are internally independent, but in reality, it is not always true.
- *K nearest neighbor:* The non-parametric machine learning technique K-nearest neighbors (KNN) is applied for classification and regression issues. Finding the k nearest data points in the feature space for a given query data point is the basic principle of KNN.
- *Standard vector machine:* The supervised machine learning method Support Vector Machine (SVM) can be applied to classification or regression applications. In a high-dimensional space, it seeks out the ideal border between several classes of data points.
- *AdaBoost Classifier:* AdaBoost (Adaptive Boosting) combines several weak classifiers to produce powerful classifiers. It is based on the concept of merging several very weak and faulty prediction rules to produce a highly accurate prediction rule and every iteration gives a higher weightage to the misclassified data.

IV. Implementation

This project is implemented in the Python programming language.The dataset was imported into a Pandas data frame and preprocessed by removing any missing values and outliers.

Step 1. Searched for the null values in the data set. As our dataset has no null values so it was not modified.

Step 2. Then the data was analyzed in different ways.

I. **Correlation between features:** Visualized the correlation using a heatmap from the seaborn library, which helped to better visualize the correlations between each feature. If the correlation between any feature is zero then it can be concluded that it is an independent feature and it might be the redundant feature.
II. **Relationship between features:** By using pair plots scatter plots were created between every feature and this is very useful to analyze the relationship between feature and trend and the pattern in the data.
III. Analyzed the outlier data in the dataset.
IV. Compared the different features with the target and observed the effect of each feature on the target.

Step 3. **Data normalization** : This data set contains features with a wide range of values. And for modeling classifiers like KNN and support vector machines are used which are sensitive to the distance of the feature. Hence it is important to normalize the data and have a standard deviation of one and a mean of zero.

Step 4. The data was then split into training and testing sets with a ratio of 70:30 using the train_test_split function.

Step 5. Apply different classifiers for model building.

- Logistic regression
- Naive Bayes classifier
- K nearest neighbor
- Support vector machine
- AdaBoost classifier

Step 6. Model evaluation :

I. **Confusion matrix** : Confusion matrix is a performance evaluation metric that is commonly used in classification problems. It is useful to count the number of accurate and inaccurate predictions the model made.
II. **Classification report** : A classification report is a helpful tool to evaluate how efficiently a classification model is working. For each class in the target variable, it gives several metrics such as precision, recall, f1-score, and support. These metrics

help in evaluating the model's performance and accuracy.

Step 7. Draw the line graph containing every classifier as the x-axis and accuracy as the y-axis and compare the result and efficiency of every classifier for this dataset.

V. Result

Based on the analysis of the CAD dataset using logistic regression, Naive Bayes, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Adaboost Classifier, it was found that logistic regression and support vector machine performed the best among all the classifiers.

The most significant characteristics for predicting CAD were found to be age, the kind of chest pain (cp), the highest heart rate achieved (thalach), and the number of major vessels (ca). Age and thalach showed a strong positive association with the target variable, according to the correlation analysis, whereas cp had a significant negative correlation with it.

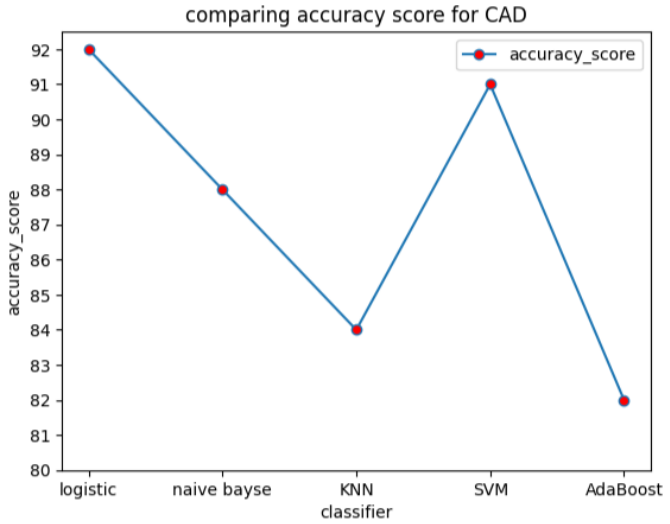

Fig. 2 Accuracy Score

Target	Precision	Recall	F1- score	support	Accuracy
Logistic Regression					
0	0.92	0.90	0.91	40	0.92
1	0.92	0.94	0.93	51	
Naive Bayes					
0	0.85	0.88	0.86	40	0.88
1	0.90	0.88	0.89	51	
K nearest neighbour classifier					
0	0.82	0.80	0.81	40	0.84
1	0.85	0.86	0.85	51	
Support vector machine					
0	0.92	0.88	0.90	40	0.91
1	0.91	0.94	0.92	51	
AdaBoost classifier					
0	0.82	0.78	0.79	40	0.82
1	0.83	0.86	0.85	51	

Table 2. Classifier Comparison

Overall, the results imply that utilizing the UCI dataset, machine learning systems can accurately predict CAD. It is possible to do more studies to investigate the usage of other algorithms and to confirm the findings using a larger and more diversified dataset.

VI. Conclusion

The use of machine learning for coronary artery disease (CAD) prediction using electrocardiography (ECG) shows promising results. Machine learning methods and numerous ECG features can be used to accurately predict CAD with high sensitivity and specificity. In this paper we used different classifier algorithms such as Logistic regression , Naive Bayes classifier, K nearest neighbor ,Support vector machine and AdaBoost classifier . Logistic regression showed a higher accuracy score than the others .The most significant characteristics for predicting CAD were found to be age, the kind of chest pain (cp), the highest heart rate achieved (thalach), and the number of major vessels (ca). Overall, using machine learning to predict CAD from an ECG can completely change how CAD is diagnosed and treated.

VII. References

[1] F. Mohammadpour, A. Deravi, and H. D. Taghirad, "A Deep Learning Approach for Heart Disease Diagnosis Using ECG Signals" vol. 23,, Jan. 2019, doi: 10.1109/JBHI.2018.2832872.

[2] J. Ma, X. Wang, Y. Zhang, J. Zhang, and J. Liu, "A Machine Learning-Based Framework for the Prediction of Cardiovascular Disease Risk" vol. 9, 2021, doi: 10.1109/ACCESS.2021.3061349.

[3] S. S. Razavian, J. M. Najafabadi, M. M. Hosseini, and M. H. Shirazi, "A Machine Learning Model for Predicting the Onset of Heart Failure Using Electronic Health Records" vol. 22, Nov. 2018, doi: 10.1109/JBHI.2018.2817393.

[4] N. Jindal, S. K. Soni, and N. Kaur, "Heart Disease Prediction using Machine Learning Techniques with the Help of General Health Information" 2019, doi: 10.1109/ICICICT45638.2019.8994081.

[5] N. K. Singh and K. Singh, "Cardiovascular Disease Prediction Model Using Machine Learning Techniques with Feature Selection" 2018, doi: 10.1109/ICIRD.2018.8603487.

[6] G. Phadke, M.R. Rajati, and L. Phadke, "Prediction of Coronary Artery Disease using Electrocardiography: A Machine Learning Approach," in 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2019

[7] D. Dey and J. Mukhopadhyay, "ECG signal analysis for automatic detection of coronary artery disease: A review," in 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), 2019.

[8] S. Venkataramanan, D. Gopinath, and M. Jayaraman, "ECG Based Diagnosis of Coronary Artery Disease Using Hybrid Feature Selection and Machine Learning Techniques," in 2020 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN), 2020.

[9]A. Subasi and M. Alshawi, "Automated detection of coronary artery disease using different durations of ECG segments with convolutional neural network," Computer Methods and Programs in Biomedicine, vol. 205, p. 106066, Dec. 2021.

[10] N. Ibrahim, H. Yusoff, A. Din, and N. Azman, "Performance Analysis of ECG-Based Feature Extraction Techniques for Coronary Artery Disease Classification," in 2021 IEEE 17th International Colloquium on Signal Processing & Its Applications (CSPA), 2021