Trends Observed in Patients with and without Coronary Artery Disease (CAD)

Arnav Jain
University of Massachusetts, Amherst
Amherst, United States
arnavjain@umass.edu

Mohit Ganna
University of Massachusetts, Amherst
Amherst, United States
mjainganna@umass.edu

***Abstract*—Coronary Artery Disease (CAD) is characterized by the progressive narrowing of coronary arteries, restricting blood flow to the heart muscle. This multifaceted cardiovascular ailment is a leading cause of morbidity and mortality worldwide. This paper delves into the nuances of utilizing machine learning algorithms to decipher intricate patterns indicative of CAD, emphasizing the potential for non-invasive and proactive diagnostic strategies. The data from the NHANES 2017-March to 2020 Prepandemic survey was consolidated, and various data files from Demographic, Laboratory, Examination, and Questionnaire data sets were also used. The vision behind the project was to be able to detect factors such as family history, diet, high cholesterol, etc., and find weights for each through various machine learning model analyses to understand the quantitative effect of each factor on people with CAD. Ridge Classifier Regression and Random Forrest were used and accuracy, precision, recall, and F1 scores were thereafter used to determine the effectiveness of the models in detecting CAD. From a feature importance perspective, the Ridge classifier proved to be a better model. The dataset consists of a total of 9 features that were used to train the models.**

I. Introduction

Coronary artery disease (CAD) is now more widely acknowledged as a significantly dangerous and potentially life-threatening chronic ailment. The primary cause of heart failure stems from the narrowing and blockage of the coronary arteries.[1] The coronary arteries deliver oxygen-rich blood from the aorta, the body's largest artery, to the heart's four chambers. Picture two traffic lanes that merge into one due to construction. Traffic keeps flowing, just more slowly. With CAD, you might not notice anything is wrong until the plaque triggers a blood clot. The blood clot is like a concrete barrier in the middle of the road. Traffic stops. Similarly, blood can't reach your heart, and this causes a heart attack.[2] Coronary Artery Disease (CAD) can stick around quietly for many years, showing no obvious signs until a heart attack suddenly happens. That's why it's called a "silent killer." It works in the background, causing trouble without making a fuss, making it hard to catch early on. The term "silent killer" sums up how CAD can quietly stay hidden, emphasizing the need to be aware and take action before it causes major harm. [2]

Coronary artery disease is caused by a buildup of plaque—a waxy substance composed of cholesterol, calcium, and fat—in these arteries.[3] Plaque comprises deposits of cholesterol and other substances in the artery. Plaque buildup causes the inside of the arteries to narrow over time, which can partially or totally block the blood flow. This process is called atherosclerosis [4].

The evolution of cardiovascular diagnostics has been marked by a relentless pursuit of precision, particularly in addressing the intricate nuances of CAD. Conventional diagnostic approaches frequently encounter challenges in identifying the subtle manifestations of Coronary Artery Disease (CAD) during its initial phases, where symptoms may elude detection until a critical juncture is imminent. This underscores the pressing need for innovative solutions. This paper explores the transformative impact of machine learning in steering the shift toward diagnostic strategies that are not only more accurate and timely but also less invasive. The ability to predict the factors contributing to a higher probability of contracting Coronary Artery Disease (CAD) represents a pivotal advancement with profound implications for its management, detection, and treatment.

The following image shows how the accumulation of plaque in the interior walls of the artery causes restriction to the blood flow.

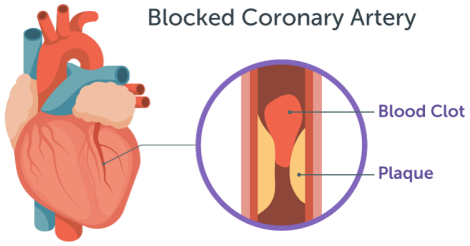

Fig. 1. Coronary Artery Disease

II. Preliminary Analysis

The initial analysis is a fundamental stage in any research undertaking, offering an early exploration and comprehension of the data or topic under examination. We performed the preliminary analysis of all the features in our project to understand the quantitative effect of each feature on the patients with and without CAD changes with changing factors. The study aimed to estimate the quantitative effect of several variables like

LDL Cholesterol levels, age, BMI, Fasting Glucose, HBa1C, smoking, alcohol consumption, etc. on CAD. These variables have already been known to impact CAD thanks to the work of other studies [5].

A. HDL Cholesterol & CAD Relationship

	CAD	NOT CAD
HDL $<$ 60	170	2447
HDL $>=$ 60	48	1097

TABLE I
HDL CHOLESTEROL AND CAD

From the table above we calculate the ODDS RATIO: 1.56 and a RELATIVE RISK: 1.52. An odds ratio of greater than 1 highlights that relative to people with HDL $>=$60, people with HDL cholesterol levels $<$ 60 increase their odds of CAD by a factor of 1.56 times. Similarly a relative risk of greater than 1 people with HDL cholesterol levels $<$ 60 are 1.52 times more likely to have CAD.

B. Smoking & CAD Relationship

	CAD	NOT CAD
SMOKED 100+	136	1465
NOT SMOKED 100+	82	2058

TABLE II
SMOKING 100+ CIGARETTES AND CAD

From the table above we calculate the ODDS RATIO: 2.33 and a RELATIVE RISK: 2.22. An odds ratio of greater than 1 highlights the fact that relative to people who have not smoked 100+ cigarettes, people who have smoked 100+ cigarettes have an increased odds of CAD by a factor of 2.33 times. A relative risk of greater than 1 highlights that people with people who have smoked 100+ cigarettes are 2.22 times more likely to have CAD. Hence, to conclude, smoking has a highly positive association with CAD and thus, should be one of the main features in helping us determine trends for patients with and without CAD.

We performed similar analyses for many other features but decided to include these two in the report since they had the largest effect on CAD as can be observed from the tables above.

III. METHOD

The results of the preliminary analysis were conclusive to help us figure out what input variable to take for the machine learning models. However, we did take help from previous existing literature as well.[5] Table III lists all the variables used as a part of training the models.

Our target variable was an OR combination of MCQ160c and MCQ160D which was if the patient has ever been told by a doctor that they have been diagnosed with CAD or if the patient has ever had angina pectoris as angina pectoris is a very common symptom of CAD.

The first step in the data cleanup was to create a combined dataset of these variables, we downloaded each file and imported them using the pandas library read_sas function. Then using the merge function, we combined this into a single dataset. The next step involved handling the NA and missing values. Then, we filtered out the dataset to remove missing, don't know, or any other irrelevant data. We referred to the NHANES documentation to help us filter this out correctly.[6] Finally, we normalized all the columns so they could be weighted the same and thus be compared more accurately.

Code	Description
RIDAGEYR	Patient Age
URXUCR	Creatinine levels
LBXGH	Glycohemoglobin
LBXGLU	HBA1c levels
SMQ020	Smoked at least 100 cigarettes
LBDHDD	HDL Cholesterol
LBDLDL	LDL Cholesterol
LBXTR	Triglycerides
BMXBMI	Patient BMI

TABLE III
INPUT VARIABLES IN DATASET

After creating a combined dataset of all these variables, we split them into 70% training set, 20% validation set and 10% testing set. Using the same data, we ran many models to find the ideal model that would best predict if the patient had ever been diagnosed with CAD. The ideal way to estimate the quantitative effect would be through a Ridge Classifier Regression which takes input variables and predicts if the person has ever been told that they have had CAD.

A. Ridge Classifier Regression

We used the model of RidgeClassifier from Scikit Learn to run our Ridge Classifier Regression.

Our first step was to find the right α as the hyper-parameter of our model. To do this, we plotted the coefficients for different values of α and observed the point where they started to stabilize. Figure 2 depicts how each coefficient changes depending on the hyper-parameter value. The model was just trained using the training set.

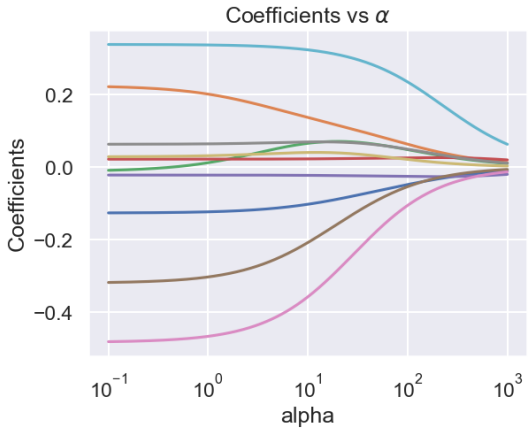

Fig. 2. Showing different coefficient values for different values of α

From this, we see that $\alpha = 10^{1.5}$ seems to be the ideal value at which the coefficients start to stabilize. To confirm

this, we also check the relationship between different metrics of accuracy, recall, and f1-score against different values of α and observe the value that maximizes all. This score was calculated by training on the training set and then running the model on both the training set and the validation set and plotting them both as seen in figures 3, 4, 5.

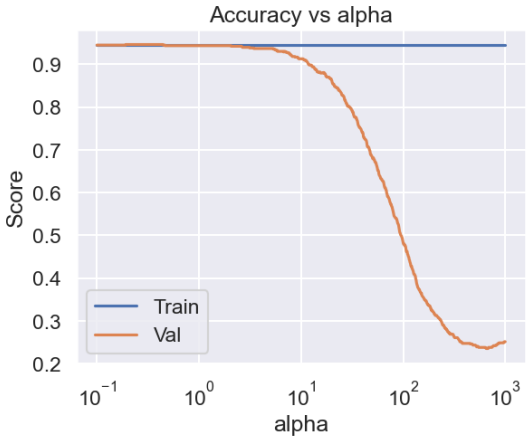

Fig. 3. Showing accuracy values for different values of α

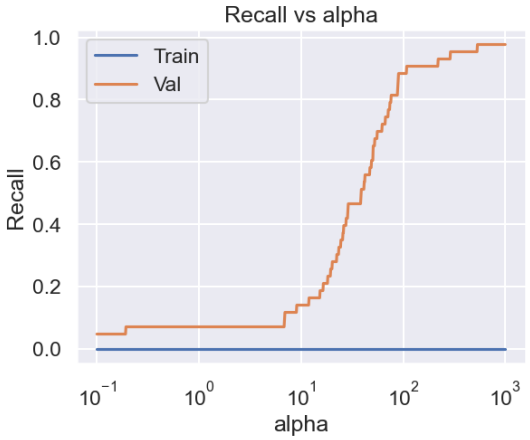

Fig. 4. Showing recall values for different values of α

This supports our ideal $\alpha = 10^{1.5}$ as all four metrics are very high at this point.

B. Random Forest

We used the model of RandomForestClassifier from Scikit Learn to run our Random Forest model.

Our first step was to find the right Num Trees as the hyper-parameter of our model. To do this, we plotted the recall score for different values of Num Trees and observed the point where they started to stabilize. Figure 6 depicts how each coefficient changes depending on the hyperparameter value. The model was just trained using the training set.

From this, we see that Num Trees = 200 seems to be the ideal value as this is where recall is maximized. To confirm this, we also check the relationship between different metrics

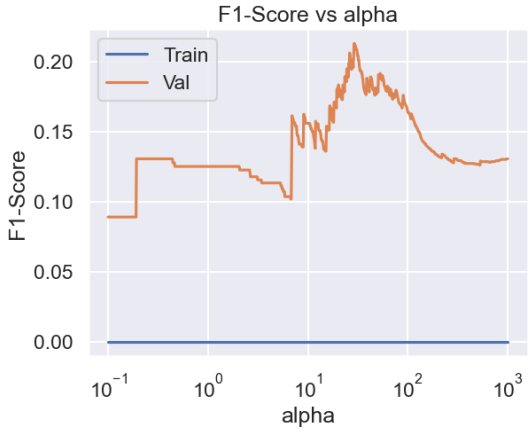

Fig. 5. Showing f1-score values for different values of α

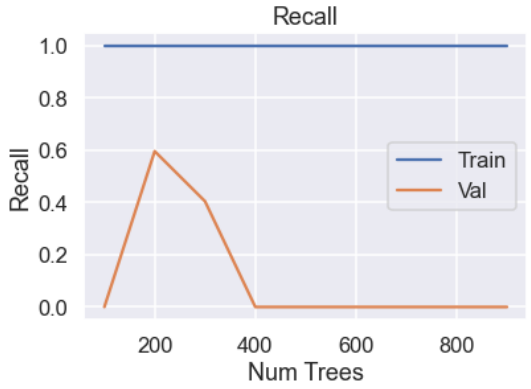

Fig. 6. Showing different recall values for different values of Num Trees

of accuracy and f1-score against different values of Num Trees and observe the value that maximizes all. This score was calculated by training on the training set and then running the model on both the training set and the validation set and plotting them both as seen in figures 7, 8.

This supports our ideal Num Trees = 200 as all four metrics are maximized at this point.

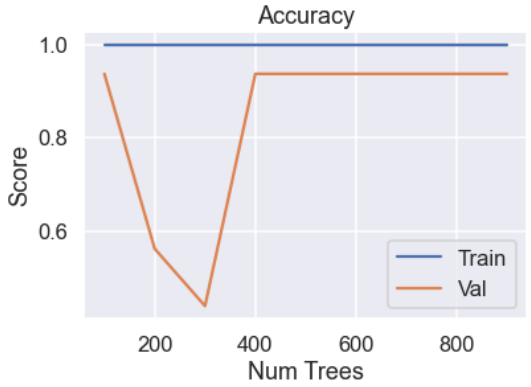

Fig. 7. Showing accuracy values for different values of α

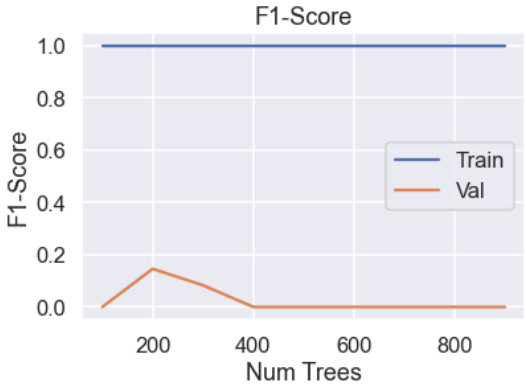

Fig. 8. Showing f1-score values for different values of α

IV. Results

We ran the model using the respective ideal values and then trained it using the training set and then made predictions for the testing set.

We received the metric values for Ridge Regression as seen in IV for the class label of True, that is people who have been diagnosed with CAD.

Metric	Score
Accuracy	0.81
Precision	0.24
Recall	0.63
F1-Score	0.35

TABLE IV
RIDGE REGRESSION METRIC VALUES ON THE TEST SET

We received the metric values for Random Forest as seen in V for the class label of True, that is people who have been diagnosed with CAD.

Metric	Score
Accuracy	0.46
Precision	0.03
Recall	0.33
F1-Score	0.06

TABLE V
RANDOM FOREST METRIC VALUES ON THE TEST SET

V. Limitations and Challenges Faced

Dataset imbalances, with certain classes being underrepresented, present challenges in machine learning. This imbalance can introduce bias and inaccuracy in the models, particularly when diagnosing medical conditions like Coronary Artery Disease (CAD) where diagnosed cases may be significantly fewer than non-diagnosed ones. The prevalence of overfitting in training models often leads to suboptimal performance on validation or test sets. Additionally, the extensive range of variables, reflecting various factors influencing CAD, complicates the selection process, making it challenging to identify the most relevant predictors.

VI. Conclusion and Future Scope

The findings from the odds ratio and relative ratio analyses establish a significant association between low levels of HDL and elevated smoking rates with Coronary Artery Disease (CAD). Addressing the imbalance, we incorporated angina pectoris as a composite target variable; future investigations might benefit from exploring a more extensive array of combinations. While the Ridge Classifier Regression proved superior in the identification of patterns related to patients with and without CAD, the metrics indicate room for improvement, urging further consideration of additional models. An additional recommendation involves incorporating an extended timeline of data, utilizing diverse years for training and separate years for testing to enhance the robustness of the study.

References

[1] Alothman, A.F., Sait, A.R.W. and Alhussain, T.A. 2022. Detecting Coronary Artery Disease from Computed Tomography Images Using a Deep Learning Technique. Diagnostics. 12, 9 (Aug. 2022), 2073. DOI:https://doi.org/10.3390/diagnostics12092073.

[2] Coronary artery disease: https://my.clevelandclinic.org/health/diseases/16898-coronary-artery-disease

[3] Diagnosing coronary artery disease: https://nyulangone.org/conditions/coronary-artery-disease/diagnosis.

[4] Coronary Artery Disease — Cdc.gov: 2021. https://www.cdc.gov/heartdisease/coronary$_a d.htm$.

[5] A Multivariate Model for Prediction of Obstructive Coronary Disease in Patients with Acute Chest Pain: Development and Validation: 2017. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5421469/. Accessed: 2023-12-03.

[6]NHANES questionnaires, datasets, and related documentation: https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/