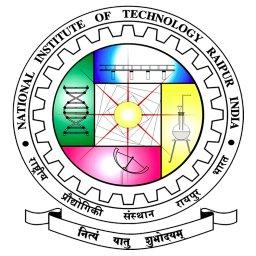
ANALYSIS OF CORONARY ARTERY DISEASE USING VARIOUS MACHINE LEARNING TECHNIQUES

MINOR PROJECT REPORT



**SUBMITTED BY**

Disha Gupta (18118020)

Mukesh Yadav (18118039)

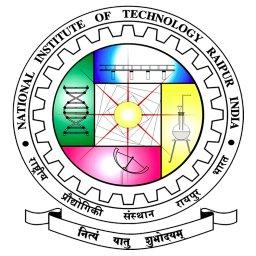
Vinayak Vishwakarma (18118089)

**UNDER THE GUIDANCE OF**

**Dr. Rekh Ram Janghel**

**Department of Information Technology**

**NATIONAL INSTITUTE OF TECHNOLOGY, RAIPUR**



CERTIFICATE

We hereby certify that the present project in the B.Tech Minor Project titled "**Analysis of Coronary Artery Disease Using Various Machine Learning Techniques**," which is submitted as a Minor Project Report & submitted to the Department of Information Technology of National Institute of Technology Raipur (C.G.), is an authentic record of our own work carried out between July 2021 to November 2021 under the supervision of **Dr. Rekh Ram Janghel Information Technology Department.**

We have not submitted the information given in this report for the award of any other degree elsewhere.

*Submitted By:*

**Disha Gupta**

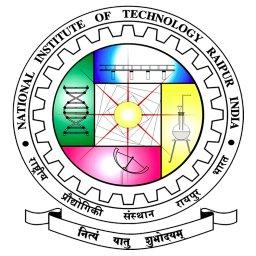
**Mukesh Yadav**

**Vinayak Vishwakarma**

This is to certify that, to the best of our knowledge, the above statement made by the candidates is true.

*Signature of Supervisor*

Date: 01-12-2021  **Dr. Rekh Ram Janghel**



**ACKNOWLEDGEMENT**

The joy, accomplishment, honor, happiness, reward, recognition, and building of our project can't be imagined without the few who, in addition to their usual plans, sacrificed their precious time. A lot of people helped shape and achieve the desired outcome, either directly or indirectly. We are grateful to **Dr. Rakesh Tripathi (HOD), Department of Information Technology, National Institute of Technology, National (C.G.)**, for allowing us to undertake this study. He was a true source of inspiration for me because of his prompt counsel, suggestions for improvement, and supervision.

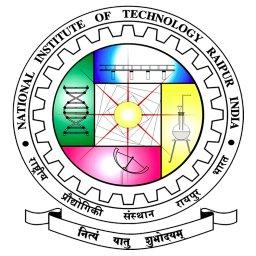
We express our heartfelt gratitude to our guides, **Dr. Rekh Ram Janghel, Assistant Professor Department of Information Technology, National Institute of Technology, Raipur (C.G.)** for their excellent advice, pointers and aid needed to execute the project work time to time. It would have been practically impossible for us to meet the initial level of aim without their guidance and drive.

Last but not least, we are grateful to our parents for constantly motivating us in our studies, as well as to our friends who have actively or passively assisted us in our effort.

Disha Gupta (18118020)

Mukesh Yadav (18118039)

Vinayak Vishwakarma (18118089)



**ABSTRACT**

The biggest cause of death globally is Coronary Artery disease (CAD). It is a complicated cardiac disease with various risk factors and a wide range of indicators. Coronary artery disease (CAD) occurs whenever the heart is not able to get oxygen-rich blood by the coronary arteries due to the buildup of cholesterol plaque on the artery's inner layer. CAD has evolved dramatically during the last decade. The purpose of this study is to create a prototype system employing several ML models and assess their performance in order to determine an appropriate model. Support Vector Machine, KNN, Decision Tree, Random Forest, Logistic Regression, Naive Bayes and XGBoost are seven of the most widely used ML models studied here. The Z-Alizadeh Sani dataset is used in this study to detect CAD.The statistical performance is analysed on the basis of Sensitivity, Accuracy, Specificity, ROC curves. Among these models Random Forest has achieved the best performance and obtained accuracy of 93.44 percent, sensitivity of 76.47 percent, and specificity of 100 percent.

**CONTENTS**

1. Introduction ……………………………………………………………… 8
2. Literature Survey…………………………………………………………..12
3. Methodology ………………………………………………………………17

3.1 Method Flow chart……………………………………………………..17

3.2 Dataset description……………………………………………………..18

3.3 Data normalization……………………………………………………..21

3.4 Data splitting…………………………………………………………...21

3.5 Classification…………………………………………………………...22

3.5.1 Support Vector Machine…………………………………………22

3.5.2 Decision Tree…………………………………………………….24

3.5.3 Random Forest…………………………………………………...25

3.5.4 K-Nearest Neighbour…………………………………………….26

3.5.5 Logistic Regression……………………………………………...27

3.5.6 Naive Bayes……………………………………………………...28

3.5.7 XGBoost………………………………………………………....29

3.6 Evaluation of Performance……………………………………………..29

3.6.1 Confusion Matrix………………………………………………....29

3.6.2 Accuracy………………………………………………………….31

3.6.3 Sensitivity………………………………………………………...31

3.6.4 Specificity………………………………………………………...32

3.6.5 ROC and AUC…………………………………………………....33

1. Result Analysis……………………………………………………………..34

4.1 Performance comparison of Algorithms………………………………..34

4.2 Comparison with previous studies………………………………….......39

1. Discussion…………………………………………………………………..41
2. Conclusion and Future work………………………………………………..42
3. References…………………………………………………………………..43

**List of Figures**

Figure 1. Progression of atherosclerosis and the genesis of CAD……………..9

Figure 2. Flow chart of the proposed method……………………………..17

Figure 3. Support Vector Machine……………………………………………...23

Figure 4. Decision tree structure………………………………………………..25

Figure 5. Random Forest working structure…………………………………….26

Figure 6. SVM score for various kernels………………………………………..34

Figure 7. KNN scores for various K values……………………………………..35

Figure 8. Random forest scores for various estimators value…………………...36

Figure 9. Accuracy comparison bar graph……………………………………...37

Figure 10. ROC-AUC graph of SVM………………………………………………...38

Figure 11. ROC-AUC graph of KNN………………………………………………...38

Figure 12. ROC-AUC graph of Decision Tree…………………………………….....38

Figure 13. ROC-AUC graph of Random Forest………………………………………38

Figure 14. ROC-AUC graph of Logistic regression…………………………………..39

Figure 15. ROC-AUC graph of Naive Bayes………………………………………....39

Figure 16. ROC-AUC graph of XGBoost……………………………………………..39

**List of Abbreviations**

Coronary Artery Disease……………………………………………….CAD

Cardiovascular disease ………………………………………………..CVD

Machine Learning ………………………………………………...……ML

Support Vector Machine…………………………………………...…...SVM

Neural Network………………………………………………………...NN

Naive Bayes………………………………………………………….....NB

Sequential Minimal Optimization ……………………………………..SMO

Principal Component Analysis………………………………………....PCA

Artificial Neural Network……………………………………………....ANN

K-Nearest Neighbor………………………………………………….....KNN

Receiver Operating Curve……………………………………………...ROC

Area Under the Curve…………………………………………………..AUC

**List of Tables**

Table 1. Literature survey table………………………………………………...14

Table 2. Description of Z Alizadeh dataset……………………………………..18

Table 3. Confusion matrix……………………………………………………....30

Table 4. Performance comparison of algorithms………………………………..37

Table 5. Performance comparison with previous studies………………………..40

1. **Introduction**

The set of conditions affecting the blood vessels and the heart are called Cardiovascular diseases(CVDs) [1]. In the 2017 report of the American Heart Association, CVD is common and the top risk factor of mortality all over the world. In 2015, CVDs caused the death of 17.7 million individuals, with 6.7 million fatalities attributed to CAD. The most prevalent cardiovascular defect appears to be CAD; heart disease is a major cause of fatalities all over the world. According to research, one-third of all women, regardless of ethnicity or race, die from CAD. CAD, Alzheimer disease, bronchus, road injury, diarrhoeal diseases, diabetes mellitus, tuberculosis,stroke, lower respiratory infections and lung cancers are among the top ten causes of death, according to the World Health Organization (WHO)[2]. Despite fact that heart disease care has altered dramatically in recent decades, people with stable CAD are nonetheless at risk of a serious cardiovascular event. According to one study[3], in the United States, CAD strikes around 6% of the elderly. As a result, CAD is the most common of the CVDs. Furthermore, by 2030, the number of deaths is predicted to reach more than 22 million. As a result, early detection and medical attention will lower the no. of CAD-related fatalities.

Atherosclerosis - the buildup of fatty deposits and cholesterol in the internal layers of the coronary artery – causes CAD[4]. The evolution of CAD is visually represented in Fig 1.

In the absence of impediment, oxygen-rich blood flows freely via a normal coronary artery. Endothelial dysfunction develops as plaque deposits in the inner

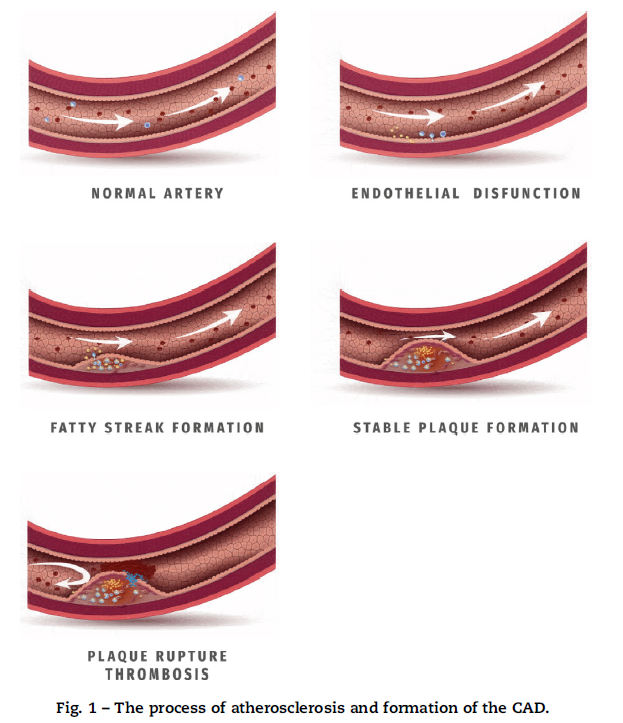


Fig. 1 Progression of atherosclerosis and the genesis of CAD[4].

Layer of the coronary artery. As plaque thickness grows, the lumen of the coronary artery narrows, interrupting blood flow. Fatty streaks and constant plaque alterations in the coronary artery wall are the two types of modifications. The coronary plaque blockage will limit blood flow over time, as a result, the heart muscle receives insufficient oxygen-rich blood. [5] Furthermore, the lipid-rich plaque may become susceptible to rupture and unstable. Acute plaque rupture, which results in the blood clot creation on the plaque's facet that fully obstructs the artery lumen, is the most common reason for acute myocardial infarction (heart attack).

The injured and weakened heart muscle is the root cause of issues such as cardiac failure and arrhythmias. Many persons with CAD are asymptomatic until the illness has progressed to the point where they experience chest discomfort and shortness of breath. As a result, health screening is critical before the disease develops to an irreversible stage.

Traditional invasive coronary angiography has been established as the benchmark of diagnosing individuals with known or suspected CAD[6]. This method, however, is time consuming, intrusive, and costly. Because this treatment normally requires a brief stay in the hospital, its invasiveness may cause some patients discomfort[7]. Furthermore, this method has a low but significant complication rate. The morphological examination of cardiac structures has been made possible thanks to electron-beam computed tomography (EBCT). This is due to EBCT's excellent temporal resolution and its use of prospective electrocardiographic triggering.The EBCT technique, however, wasn’t regarded as the suitable method for detecting the existence of coronary stenosis due to its low spatial resolution. The development of computed tomography (CT) angiography improved the detection and treatment of CAD and also the measurement of heart function in various situations[8].

Machine-learning (ML) techniques are now utilised in a variety of products. These technologies have been employed in medical applications for decades, and they have resulted in numerous advancements on various levels. Specifically, numerous machine-learning algos for diagnosing and understanding illness progression have been proposed. To diagnose CAD , ML models[9] namely Decision trees, Support Vector Machine (SVM), XGBoost, Neural Network (NN), Naive Bayes (NB), Random forest have been used.

In this study, we have proposed seven ML classifiers to classify CAD and analyse the classification results. Prediction of CAD follows the step as- data composition where data was lifted from Z-Alizadeh Sani dataset; then data preprocessing is done for normalising the data; data splitting- complete dataset was divided into testing and training; lastly various models based training data are trained, then in classification step validation of testing data is done by the trained data. Finally, different ML models used above are compared in terms of finding statistical performance, accuracy, specificity and sensitivity.

The rest report is organised as: Section-2 represents the literature survey of various studies performed on the CAD dataset. Section 3 shows the methodology for classification of CAD dataset after that results analysis is done in Section 4. Section 5 shows the discussion and section 6 represents the conclusion and future work of the paper.

1. **Literature survey**

In past years,Many research on the diagnosis of CAD on various datasets employing various ML approaches have been undertaken. In the realm of heart disease, the much more current data that researchers have used is the Z-Alizadeh Sani dataset. To that purpose, we examine the latest work on the Z-Alizadeh Sani dataset[10].

Alizadeh Sani et al. [11] advocated using data mining algorithms focused on ECG characteristics and features to diagnose CAD. In their researches, they used Sequential minimal optimization (SMO) method for feature optimization and then Naïve Bayes algo is applied to classify the CAD. Lastly, using the 10-fold cross-validation strategy performed on the SMO-Naïve Bayes hybrid algo produced accuracy of 88.52% while accuracy achieved by SMO was 86.95% and Naïve Bayes achieved 87.22% accuracy.

In addition, Hosseini et al.[12] developed classification algos Naïve Bayes, SMO, SMO, Neural networks and Bagging with SMO for detection of CAD .Information gain and confidence in CAD have also been utilised to assess beneficial features. As a consequence, the SMO algo achieved 94% using 10-fold cross-validation approach which was highest amongst them.

Zanguei et al.[13] applied CI approaches to diagnose CAD, resulting in the diagnosis of three significant coronary stenosis utilising all characteristics. They investigated the significance of vascular stenosis features using analytical approaches. Eventually, using the SVM with 10-fold cross-validation , the accuracy of the left anterior descending (LAD), left circumflex (LCX), and right coronary arteries coronaries (RCA) was 86.14 percent, 83.17 percent, and 83.50 percent, respectively.

Arabasdi et al. [14] introduced a neural network-genetic hybrid algorithm for CAD detection. In their study, genetic and neural network algos were employed alone and in combination to assess the dataset, with the accuracy of the neural network algo and the neural network-genetic algo using the 10-fold cross-validation technique being 84.62 percent and 93.85 percent, respectively.

Khosravi et al. [15] employed the Naive Bayes, C4.52, and SVM in a feature engineering algorithm for non-invasive CAD diagnosis. Given that their dataset has grown from 303 records to 500 samples. For the Naive Bayes, C4.52, and SVM algorithms, the accuracy attained using the 10-fold cross-validation approach was 86 percent, 89.8 percent, and 96.40 percent, respectively.

Abdar et al. [16] employed a two-level hybrid genetic algo and NuSVM termed N2 Genetic-NuSVM in their investigation. It's being used to improve the SVM parameters and choose the feature in parallel given a two-level genetic algorithm. The accuracy of CAD diagnosis with their suggested technique was 93.087 percent using a 10-fold cross-validation procedure.

**Table 1.** Literature survey table

| **Sr.No** | **Author, [citation number], year** | **Feature extraction method and Classifiers** | **Performance of paper(Results)** | **Use case** |
| --- | --- | --- | --- | --- |
|  | **Accuracy** |  |
| 1. | Alizadeh Sani. al. [11], 2012 | SMO,Naive Bayes | 88.52% | Based on symptoms and ecg features, diagnosis of CAD using data mining techniques. |
| 2. | Hosseini et. al. [12], 2013 | Naïve Bayes,SMO, Neural networks and Bagging with SMO | NN**-** 94.08% | For the diagnosis of CAD, data mining approach . |
| 3. | Zanguei et. al. [13], 2016 | SVM model with 10-fold cross-validation method, features selection | 86.14% | Using computational intelligence methods for CAD detection. |
| 4. | Arabasadi et. al. [14], 2017 | Neural Network-Genetic algo | network-genetic algo- 93.85% | Computer-aided decision making for diagnosis of heart disease utilising hybrid neural network algos |
| 5. | Khosravi et. al. [15], 2018 | Naive Bayes,SVM algos, and C4.5 | SVM- 96.40%, | Non-invasive diagnosis of coronary artery disease in high-risk individuals based on stenosis prediction of individual coronary arteries. |
| 6. | Abdar et al. [16],2019 | two-level hybrid genetic algo | 93.08% | two-level hybrid genetic algo for diagnosis of CAD. |
| 7. | Dipto, I.C., Islam et al.  [17], 2020 | SVM, Logistic Regression and ANN. | ANN**-** 93.3% | Comparison of Different ML  algos for the Prediction of CAD |
| 8. | Javad Hassannataj Joloudari et. al.[9], 2020 | RT, SVM, CHAID, decision tree | RT**-** 91.47% | CAD Diagnosis; Ranking Significant Features Using RTs. |
| 9. | Roohallah Alizadehsani et. al. [18], 2012 | SMO, Naïve  Bayes, C4.5 and AdaBoost | 82% | Diagnosis of CAD Using Data Mining |
| 10. | Savita et. al. [19], 2021 | PCA, Firefly Optimization, Decision tree | 93.3% | Efficient Predictive Modelling for the Classification of CAD |
| 11. | Abdar et al.[20], 2019 | ensemble learning techniques | 94.66% | NE-nu-SVC: A New Nested Ensemble Clinical Decision Support System for CAD Diagnosis |
| 12. | Ali Cüvitoğlu et. al [21] 2018 | PCA, SVM, NB, ANN | ANN-85.55% | Principal Component Analysis & ML methods for Classification of CAD Datasets |

1. **Methodology**

**3.1. Method flow chart**

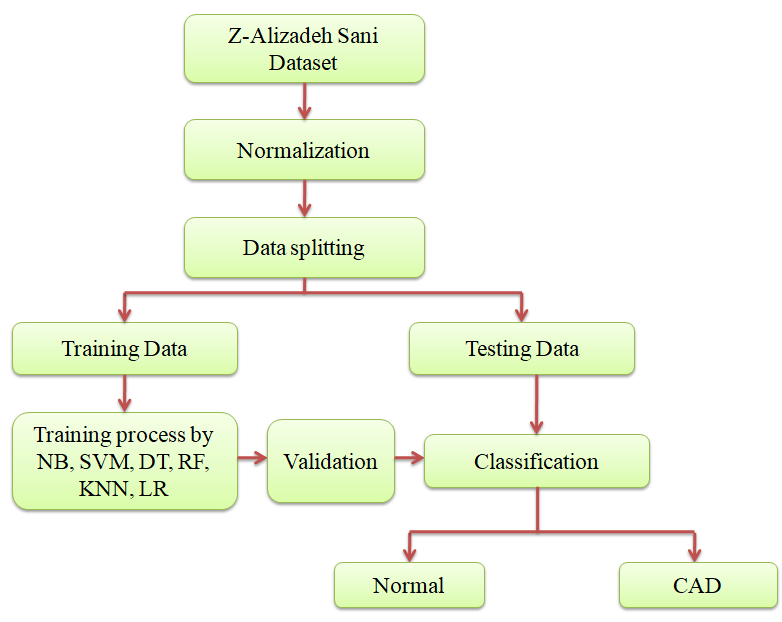
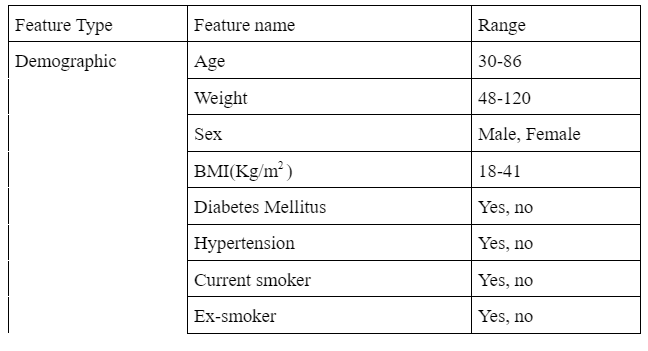
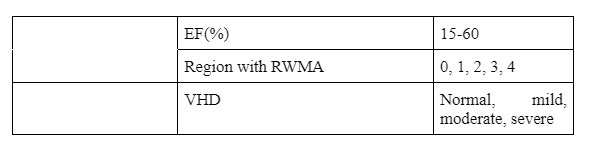
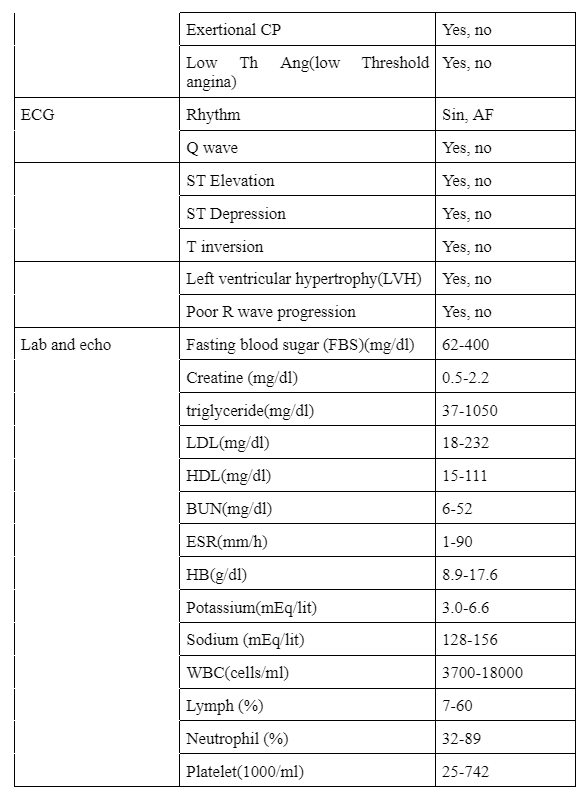
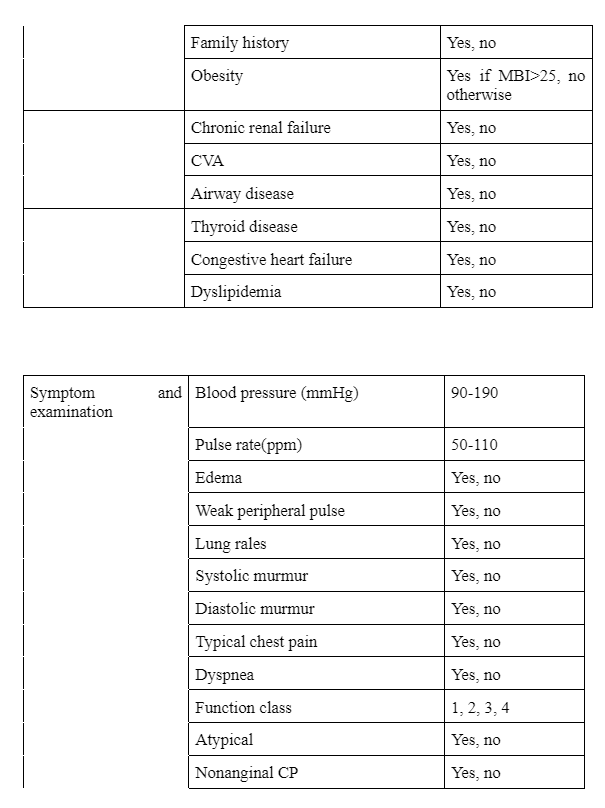


Fig 2. Flow chart of the proposed method

**3.2. Dataset description**

Alizadeh et al. reported a dataset , called the Z-Alizadeh Sani dataset[10], is used in this research paper. This dataset is one of the most famous datasets used in ML for automatic CAD detection. Here there are 303 records of patients which some of them have CAD. If one or more of his/her coronary arteries are stenosis then the patient is categorized as a CAD patient. A coronary artery is described as stenosis if its diameter narrowing is >= 50%. Accordingly, 216 patients had CAD, and the dataset contains 88 patients with the normal situation on the Z-Alizadeh Sani. Each row consists of 55 features which can be used to get idea if the patient has CAD or not.

**Table 2.** Description of Z Alizadeh dataset



**3.3. Data normalization**

In the data analysis process, preprocessing is required after data gathering. The Z-Alizade-Sani dataset was numerical and string[9]. First, the number of records is transformed from nominal to numerical values such as gender, cerebrovascular accident, Airway disease, chronic renal failure, Thyroid disease, dyslipidemia, congestive heart failure, etc. In the preprocessing, we convert all data into numerical form in the range of 0 to 1, which is called as the normalisation. Following the normalisation of numbers, it was important to convert the string to numeric data. Considering the nature of the string data, a value in the interval [0,1] was allocated to them. For instance, the sex feature has female & male values, which can either be 0 or 1.

In general, normalization leads to an increase in the accuracy of the classification models.

**3.4. Data Splitting**

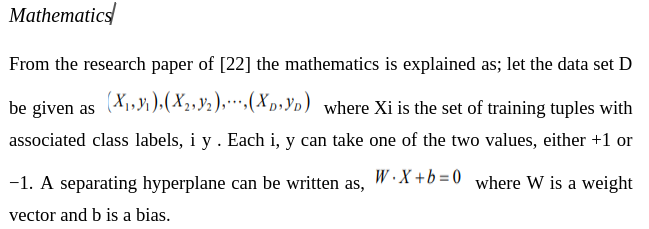
To examine the process, the splitting phase is employed to create the training and testing data. Our entire database is divided into testing and training data, with 20 percent of the data being used for testing purpose and 80% being used for training purpose.

**3.5. Classification**

In classification, ML models are used to analyse split training and testing data. To begin, training data was used to train seven different machine learning models: KNN, Decision Tree, SVM, Random Forest, XGBoost, Logistic Regression, Naive Bayes. Then, using training data with a more accuracy rate of classification , the testing data is validated. The following are seven different algorithms that are described in depth.

**3.5.1. Support Vector Machine**

SVM is a Supervised Learning algorithm which utilises non-straight planning to transform the initial prepared information into a higher dimensional space[22], within which it seeks for the linear optimal separating hyperplane that separates the tuples of class from each other. A hyperplane with an adequate nonlinear mapping to a sufficiently high dimension may always be used to segregate data. The algorithm locates the hyperplane by utilising support vectors and margins.



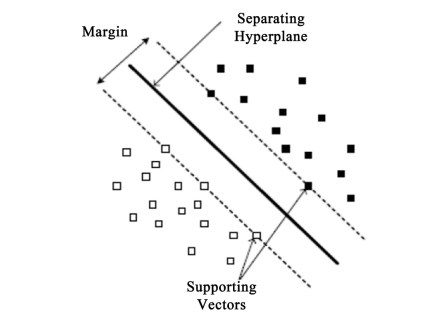
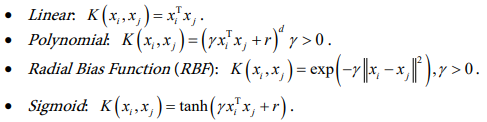


Fig 3. Support Vector Machine

The decision function is applicable to data samples that may be separated linearly. When the data samples are not linearly separable, the resulting function is as follows:

 (1)

where K ( x ,x ) = (φ (x1), φ ( x)) and φ ( x) is the non-linear space from the original space to high dimensional space. All the basic kernel as given below[23]:



**3.5.2. Decision Tree**

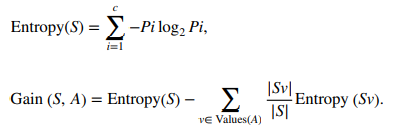
Decision Tree is a technique that is used to work with both category and numerical data. Tree-like structures are created using decision trees. Medical datasets are frequently handled using decision trees. The data in tree-shaped graphs is simple to implement and analyse. The study is based on three nodes in the decision tree model.

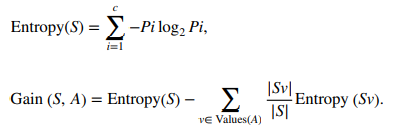
•The root node is the most important node, and it serves as the foundation for all other nodes.

• Interior node: It is in charge of numerous properties.

• Leaf node: This node indicates the outcome of each test.

The data is divided into two or more equivalent sets by this algorithm. The entropy of each characteristic is calculated, and data is divided into predictors with most info gain or the least entropy:

 (2)

 (3)

The results are simple to understand and read [24]. Because it analyses the dataset in a tree-like graph, this technique is more accurate than other algos. However, the data may be reclassified, and for decision-making purposes, only one characteristic is checked at a time..

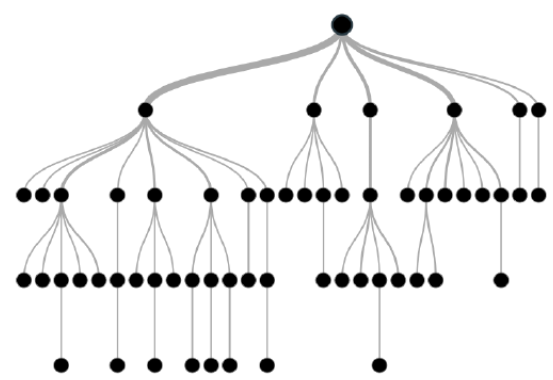


Fig.4. Decision tree structure

**3.5.3. Random Forest**

Random forest Algorithm is a supervised classification method. Forest is created by a few trees in this algorithm. [25] In a random forest, each tree emits a class assumption, and the class with the most votes becomes a models conjecture. More trees in a random forest classifier, the higher the accuracy. The following are the three standard procedures:

* Forest RI (Random information decision);
* Forest RC (Random mix);
* Combination of Forest RI and RC.

It excels at classification tasks and can overcome missing characteristics. Furthermore, because it takes a long time to get predictions because it needs more trees and massive data, therefore results are unreachable.

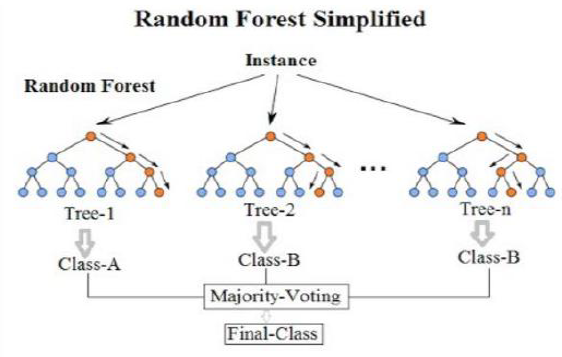


Fig 5. Random Forest working structure

**3.5.4. K‑Nearest Neighbor (K‑NN)**

The K-NN algorithm is a type of classification algorithm. It describes objects that are reliant on their immediate surroundings. It's a type of event-based learning. The distance between a characteristic and its neighbours is calculated using the Euclidean distance [26]. It takes a collection of focuses which are named and then utilises it to stamp other points in the most efficient way possible. The data are grouped together according to their resemblence, and K-NN can be used to fill in the gaps in the data. When the missing qualities are filled in, the data set is subjected to various prediction strategies. Using different combinations of these algorithms, it is possible to improve accuracy. Without a model or different suppositions,it is simple to complete the K-NN algorithm. This algorithm can be used for regression & classification, or searching. Despite the fact that K-NN is the most straightforward algo, uproarious and insignificant highlights influence its accuracy.

## **3.5.5.Logistic Regression**

Logistic regression is a supervised learning classification algo which predict the probability of variable..[27]

The variable is binary in nature having data as 1 or 0 .

Numerically, a logistic regression model predicts P(X=1) as an element of Y. One of the most straightforward ML algos can be utilized for different classification issues, for example, spam discovery, Diabetes prediction etc.

## *Types of Logistic Regression*

In general, logistic regression implies binary logistic regression with binary objective factors, but it can also predict two additional classes of target factors [27].Due to the obvious vast number of classes, logistic regression can be classified as follows.

### Binary or Binomial

### Multinomial

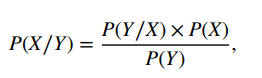
### Ordinal

## *Logistic Regression Assumptions*

* We should be aware of the logistic regression's accompanying assumptions.
* The objective factors in binary logistic regression should always be binary, and the ideal result is addressed by the variable level 1.
* The model should not have any multicollinearity, which means the autonomous factors should be independent of one another.
* Important factors for our model should be remembered.
* For logistic regression, we should choose a large sample size.

**3.5.6. Naïve Bayes’**

The Naive Bayes classifier is a supervised algo.The Bayes hypothesis is used in a simple classification procedure. It expects ascribes to have strong (Naive) autonomy. The Bayes hypothesis is a mathematical formula for calculating probability. The predictors are not linked to one another or identified with one another. Every one of the properties adds to the probability of amplifying it on its own. It is compatible with the Naive Bayes model but does not employ Bayesian techniques. Naive Bayes classifiers are used in a variety of perplexing real-life situations. [28]

****  (4)

The back probability is P(X/Y), the class earlier probability is P(X), the predictor earlier probability is P(Y), and the probability is P(Y/X). Naive Bayes is a simple, straightforward, and effective classification algorithm for non-linear, jumbled data. Regardless, there is a lack of precision because it is based on speculation and class restrictive freedom.

**3.5.7. XGBoost**

The XGBoost is an implementation of gradient boost decision trees which is made for better speed and performance.

*Model Features*

The model's implementation includes the features of the scikit-learn and R implementations, as well as novel enhancements like as regularisation. There are three types of gradient boosting assisted:

1.Gradient Boosting

2.Stochastic Gradient Boosting

3.Regularized Gradient Boosting

The gradient boosting decision tree technique is implemented in the XGBoost library.

*Algorithm Features*

1.Sparse Aware

2.Block Structure

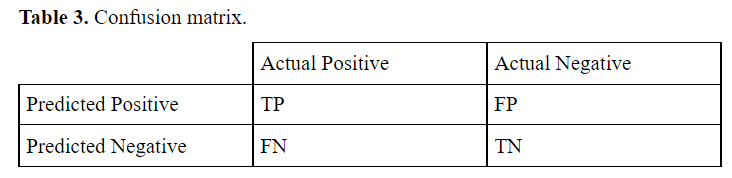
3.Continued Training

**3.6. Evaluation Of Performance**

**3.6.1. Confusion Matrix**

A performance metrix which is used to depict performance of any classifier by calculating assessment parameters[25].

**Table 3.** Confusion matrix.

****

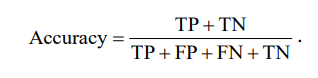
*True Positive* = People predicted as experiencing the infection (or unhealthy) are really experiencing the sickness (unhealthy); all in all, the true positive addresses the quantity of people who are unhealthy and are predicted as unhealthy.

*False Negative* = People who are really experiencing the illness (or unhealthy) are really predicted to be not experiencing the infection (healthy). As such, the false negative addresses the quantity of people who are unhealthy and predicted as healthy. In a perfect world, we would look for the model to have low false negatives as it would end up being dangerous or business undermining.

*True Negative* = People predicted as not experiencing the illness (or sound) are really observed to be not experiencing the sickness (solid); all in all, the true negative addresses the quantity of people who are sound and are predicted as healthy.

*False Positive* = People predicted as experiencing the sickness (or undesirable) are really observed to be not experiencing the illness (sound). At the end of the day, the false positive addresses the quantity of people who are solid and predicted as undesirable.

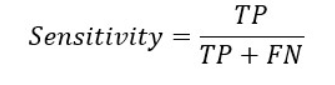
## **3.6.2.Accuracy** Efficiency of any algorithm can be evaluated on the accuracy of CAD forecast which is given by the eqn. Any model’s accuracy is obtained in every fold, and on the end of the preparation process, there will be 10 accuracies for each model. To appreciate the variance, the average is taken, as well as the standard deviation[29].

**** (5)

**3.6.3. Sensitivity**

It is the part of real positive instances which were shown to be positive. Sensitivity a.k.a recall, implies there'll be an additional number of verifiable positive examples that will be misinterpreted as negative (and, thus, could in like manner be named as the false negative). This is often referred to as a false negative rate. The sensitivity and false negative rate would be equal to one. We must learn to consider this using the model used to predict whether or not a person is suffering from the disease. Sensitivity is a measure of the proportion of persons affected by an illness who were expected to be affected by the sickness. As a result, the individual who is troublesome is actually believed to be unfortunate[30].

Sensitivity may be calculated mathematically as follows:

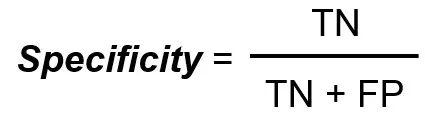
****  (6)

Greater the sensitivity, less the false negative & the greater the value of true positive. The smaller the level of sensitivity, the smaller the true positive & the larger false negative. Models with great sensitivity are required for the healthcare services and financial sectors.

## 

## **3.6.4. Specificity**

It is defined as the part of real negative that were known to be negative. Implies that higherno. of legitimate negatives are there which were misinterpreted to be positives & may be referred by false positives. This degree is often referred to as a false positive rate. The specificity and false positive rate would always be one. What if we try to comprehend this using the model used to predict if a person is exposed to contamination? It is a measure of part of those who are not experiencing the sickness who were predicted to be as sick as those who are experiencing the ailment. All things considered, a strong individual is assumed to be sound[30]. Specificity may be calculated mathematically as follows:

****  (7)

where TN=True Negative, TN= True Negative and FP=False Positive

The greater the specificity, the lower the level of false positives and greater the value of genuine negatives. The lower the specificity number, the greater the false positive value and lower the real negative value .

**3.6.5. ROC and AUC**

We will draw a Receiver operating characteristic curve, known as ROC and an area under the curve, known as AUC[30].Average accuracy of each model will be depicted using the ROC-AUC plot for each.TPR and FPR is the two data that determine the ROC curve (FPR).TPR, a.k.a Sensitivity, Hit Rate,or review, is calculated by:

 (8)

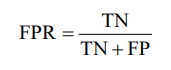
FPR (or Fall-out) is calculated by:

 (9)

We can also get FPR from specificity as:

 (10)

Here specificity can be defined as:

 (11)

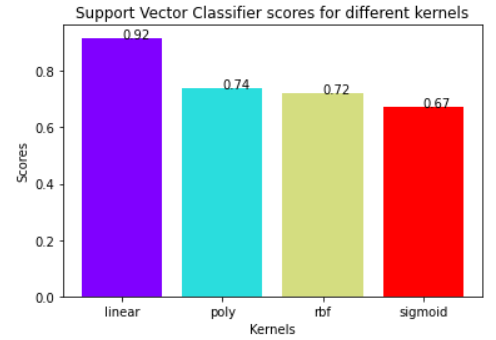
This measure refers to negative points which are wrongly perceived as positive when applied to every negative point data. To put it another way, the greater is FPR, more negative point data will be mistakenly labelled. Two previous measures with various threshold is generated & is plotted on a graph with FPR on x-axis and TPR on y-axis to merge both in single unit, which is to create AUC. The result curve is known as AUROC curve.

1. **Result Analysis**

**4.1. Performance Comparison of Algorithms**

This section displays the classified results from several prediction models. We compared several models using various parameters, which were Accuracy, Sensitivity, and Specificity. The comparison with five models is shown in Table 4.

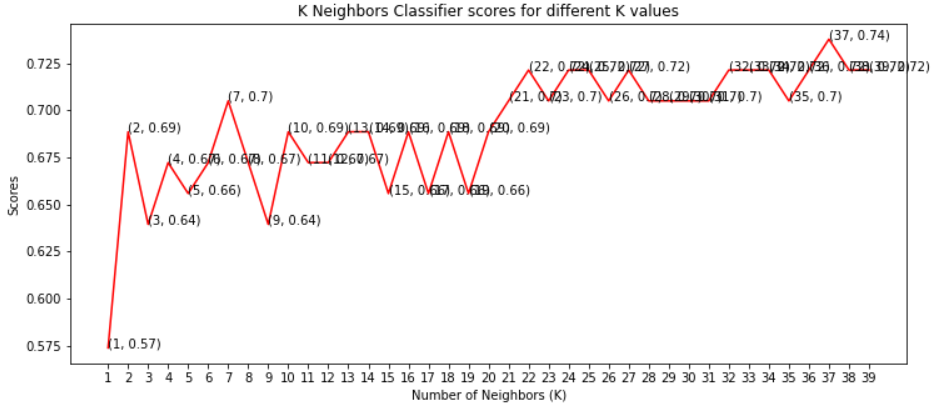
In SVM, the hyperplane is determined by one of multiple kernels. We experimented with four kernels: linear, poly, rbf, and sigmoid.



*Fig. 6 SVM score for various kernels*

The linear kernel fared the finest for this dataset, as seen in the figure above, with a score of 91.8 ~ 92 percent.

In KNN, the number of neighbours can be varied. So we varied the no. of neighbours from 1 - 40 & computed the score of the test in every case. Thereby, in each scenario, produce a line graph of the no. of neighbours and the test score obtained.

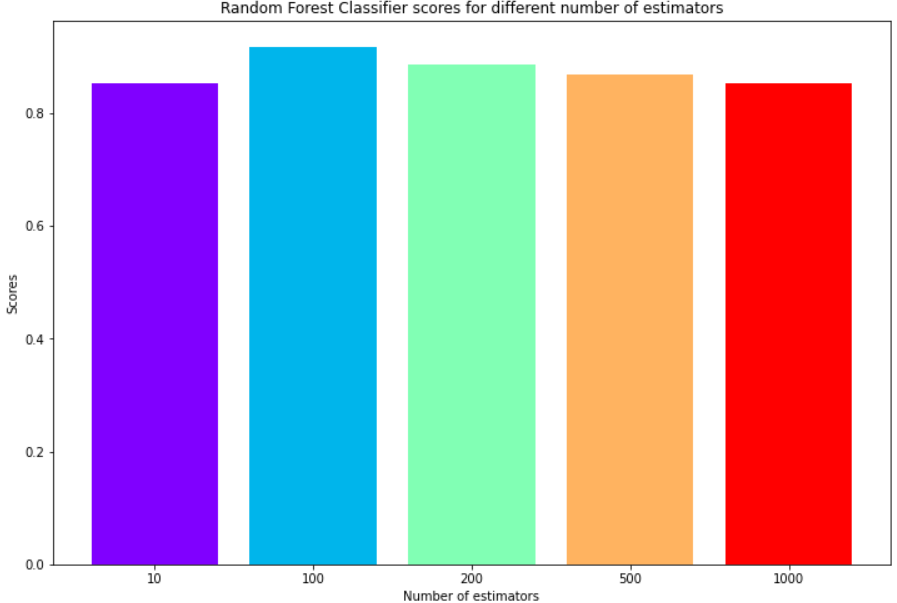


*Fig. 7 KNN scores for various K values*

As you've seen, when the no. of neighbours was set to 37, we received the highest possible score of 73.77~ 74 percent.

Random forest generates a forest of trees, with each tree produced by a randomized feature extraction from the whole set of features. We may change the no. of trees used to forecast the class here. We computed test scores for ten, hundred, two hundred, five hundred, and thousand trees.

Following that, we graph these scores on a bar graph to discover which produced the greatest outcomes.



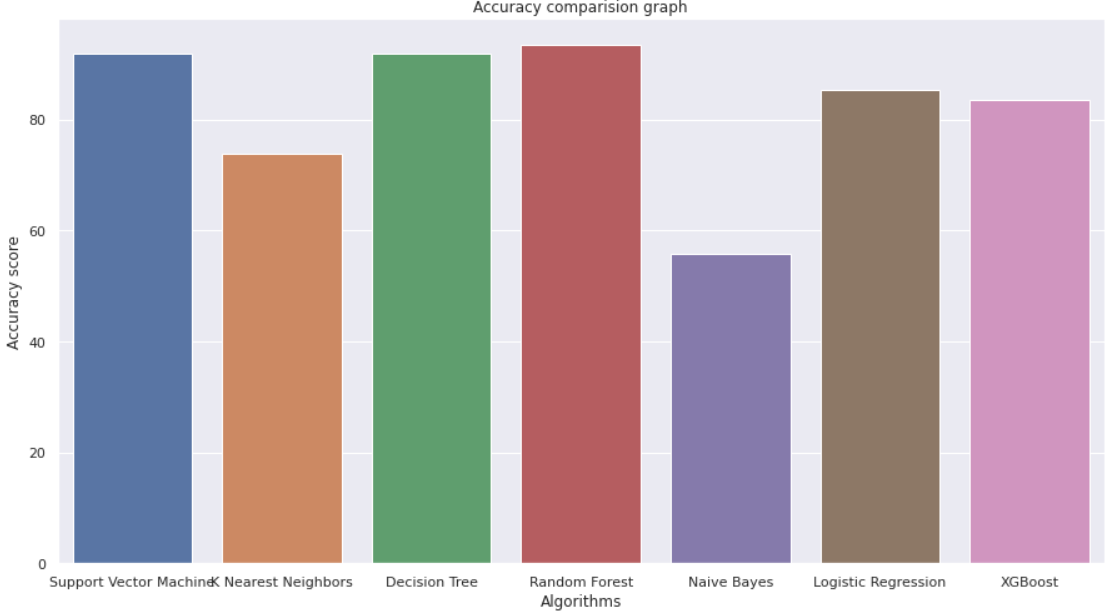
*Fig. 8 Random forest scores for various estimators value*

Looking just at bar graph, we can observe that the highest score was obtained for 100 trees.

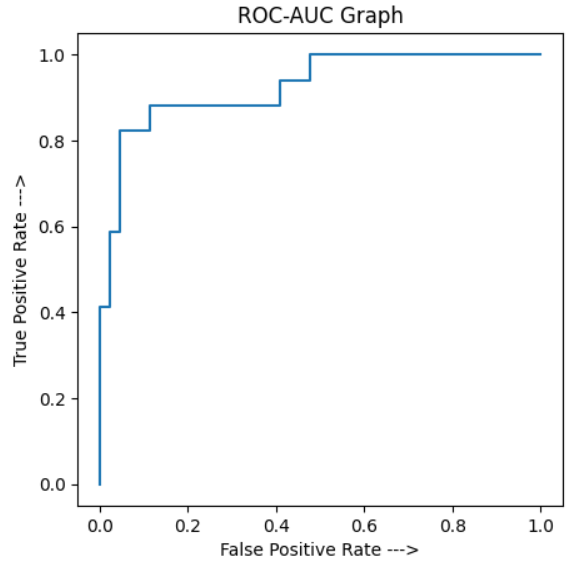
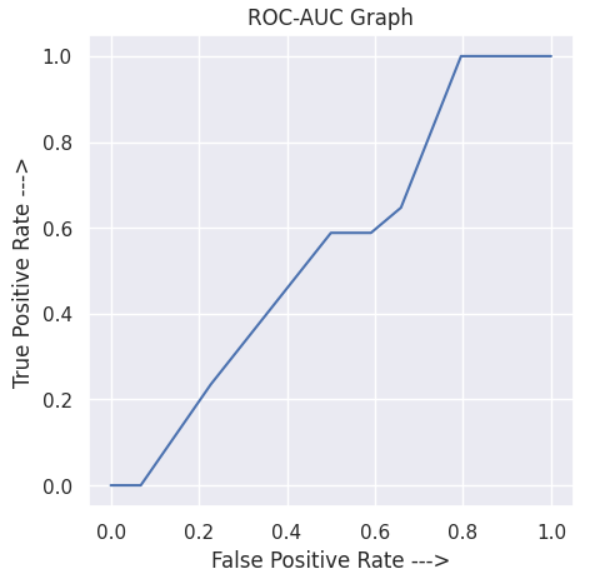
Table 4 compares several ML algorithms for analysing CAD based on accuracy, sensitivity, and specificity.

**Table 4.** Performance comparison of algorithms

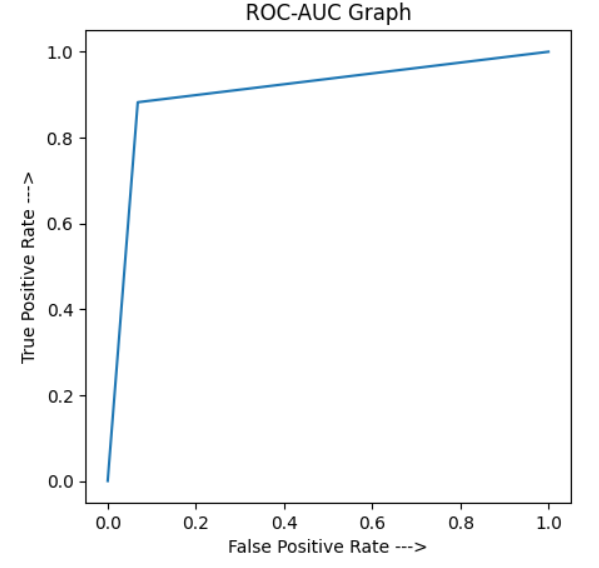
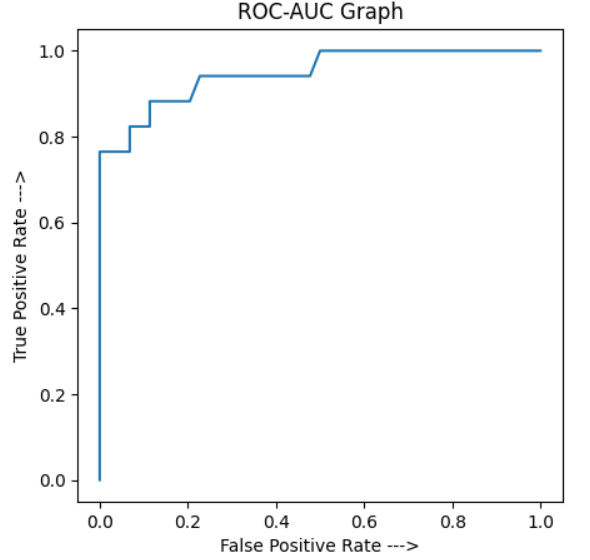
| **Classification algorithm** | **Accuracy(with 10 fold)(%)** | **Accuracy(without 10 fold)(%)** | **Sensitivity(%)** | **Specificity(%)** |
| --- | --- | --- | --- | --- |
| SVM-linear | 85.46 | 91.8 | 82.3 | 95.4 |
| Decision Tree | 76.92 | 91.8 | 88.24 | 93.18 |
| Logistic Regression | 83.85 | 85.25 | 64.71 | 93.18 |
| KNN | 71.28 | 73.77 | 5.88 | 100 |
| Naive Bayes | 52.43 | 55.74 | 88.24 | 95.45 |
| XGBoost | 85.82 | 83.61 | 64.71 | 90.91 |
| **Random Forest** | **85.54** | **93.44** | **76.47** | **100** |

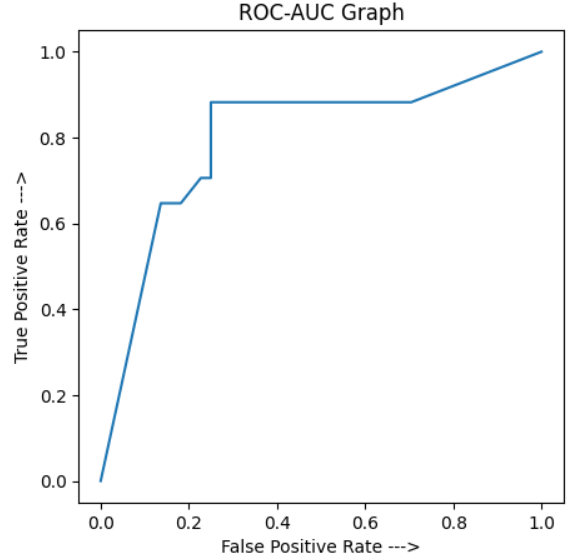
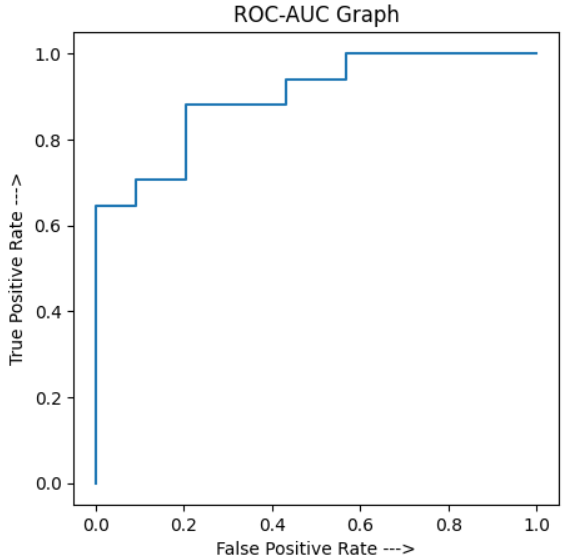


*Fig. 9 Accuracy comparison bar graph*

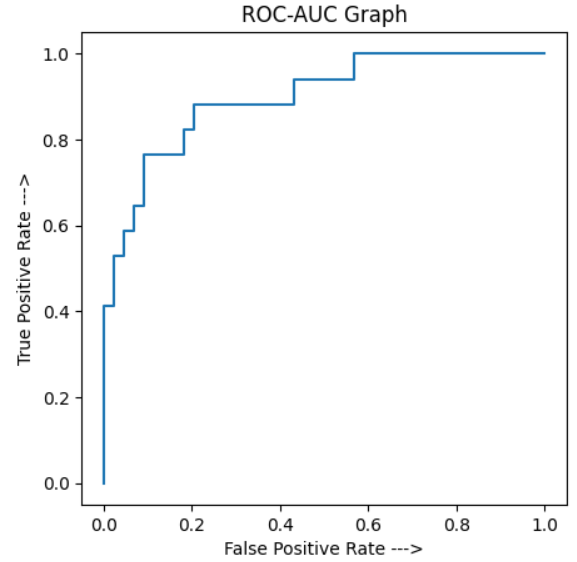
 

*Fig 10. ROC-AUC graph of SVM Fig 11. ROC-AUC graph of KNN*

  *Fig 12. ROC-AUC graph of Decision tree Fig 13. ROC-AUC graph of Random Forest*

*Fig 14. ROC-AUC graph of Decision tree Fig 15. ROC-AUC graph of Random Forest*



*Fig 16. ROC-AUC graph of XGBoost*

As shown in Table 4, the results of training the models on the dataset demonstrate that Random Forest fared better than another five techniques, with the greatest accuracy of 93.44 percent. That's also obvious from the ROC graph plotted for all of the Methods trained on the dataset.

**4.2. Comparison with previous studies**

Several researchers have used before working on the Alizadeh Sani dataset for evaluation of detection and classification of CAD. We compare our proposed method to other methods which used the Alizadeh Sani dataset, in prior studies. Table 5 compares the accuracy, sensitivity, and specificity of several methodologies used in prior research.

**Table 5.** Performance comparison with previous studies

| **Reference** | **Methods** | **Accuracy(%)** | **Sensitivity(%)** | **Specificity(%)** |
| --- | --- | --- | --- | --- |
| [38] | ANN | 85.85 | 72.50 | 91.21 |
| [21] | SVM | 86.14 | 90.96 | 79.37 |
| [26] | Random Tree | 91.47 | - | - |
| [27] | SMO | 82.16 | 90.74 | 60.92 |
| Our study | Proposed | 93.44 | 76.47 | 100 |

1. **Discussion**

In this study, we investigated and examined the several ML algos for predicting CAD, including SVM, Decision Tree, KNN, Random Forest, XGBoost, Logistic Regression and Naive Bayes. Prediction of CAD pursued the step of data preparation, in which details were taken from the Z-Alizadeh Sani dataset; then data preprocessing was done to normalise the data; data splitting was done in which the entire data was split into testing and training; eventually, various models based training data were trained and testing data were validated by the trained data in the step of classification.Finally, several ML models are contrasted above in terms of determining quantitative assessment metrics like sensitivity, accuracy & specificity. We assessed the sensitivity, specificity, and accuracy and discovered that a Random Tree classifier produced the best results with 93.44 percent accuracy, 76.47 percent sensitivity, and 100 percent specificity.

1. **Conclusion and Future works**

Numerous methods were used on the ZAlizadeh Sani dataset in this research work, and the findings were discussed. According to our medical understanding, the traits in this dataset might be indicative of CAD. Furthermore, the characteristics employed in this report may be tested along with little side effects and expenditures. As a result, using the suggested method may detect the CAD condition with cheap cost and great probability. In the future, we hope to examine forecasting the status of each artery separately. Furthermore, it is evident that accurate diagnosis of unhealthy individuals is more crucial than accurate detection of healthy individuals. As a result, an additional objective to achieve is to use budget conscious algos to take this element into account. Eventually, bigger datasets with more characteristics, as well as wider data mining methodologies, might be employed to provide bigger and more important findings.

**References:**

[1]Fujita, H., Sudarshan, V. K., Adam, M., Oh, S. L., Tan, J. H., Hagiwara, Y., ... & Acharya, U. R. (2017, June). Characterization of cardiovascular diseases using wavelet packet decomposition and nonlinear measures of electrocardiogram signal. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems* (pp. 259-266). Springer, Cham.

[2] WHO. Global Health Estimates 2016: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2016. Geneva: World Health Organization; 2018.

[3] Mozaffarian, D., Benjamin, E. J., Go, A. S., Arnett, D. K., Blaha, M. J., Cushman, M., ... & Turner, M. B. (2016). Heart disease and stroke statistics—2016 update: a report from the American Heart Association. *circulation*, *133*(4), e38-e360.

[4]Kumar, M., Pachori, R. B., & Acharya, U. R. (2017). Characterization of CAD using flexible analytic wavelet transform applied on ECG signals. *Biomedical signal processing and control*, *31*, 301-308.

[5]Acharya, U. R., Fujita, H., Lih, O. S., Adam, M., Tan, J. H., & Chua, C. K. (2017). Automated detection of CAD using different durations of ECG segments with convolutional neural network. *Knowledge-Based Systems*, *132*, 62-71.

[6] Becker, C. R. (2002). Assessment of coronary arteries with CT. *Radiologic Clinics*, *40*(4), 773-782.

[7] Sun Z. Coronary Computed Tomography Angiography in CAD: A Systematic Review of Image Quality, Diagnostic Accuracy and Radiation Dose. Nova Science Publishers, Incorporated; 2013.

[8] Sun Z, Choo GH, Ng KH. Coronary CT angiography: current status and continuing challenges. The British journal of radiology 2012;85(1013):495-510.

[9] Hassannataj Joloudari, J., Hassannataj Joloudari, E., Saadatfar, H., GhasemiGol, M., Razavi, S. M., Mosavi, A., ... & Nadai, L. (2020). CAD Diagnosis; Ranking the Significant Features Using Random Trees Model. *arXiv e-prints*, arXiv-2001.

[10] <https://archive.ics.uci.edu/ml/datasets/Z-Alizadeh+Sani>

[11] Alizadehsani, R., Habibi, J., Hosseini, M.J., Boghrati, R., Ghandeharioun, A., Bahadorian, B. and Sani, Z.A. Diagnosis of CAD using data mining techniques based on symptoms and ecg features. European Journal of Scientific Research,. 82(4): p. 542-553, 2012.

[12] Alizadehsani, R., Habibi, J., Hosseini, M.J., Mashayekhi, H., Boghrati, R., Ghandeharioun, A., Bahadorian, B. and Sani, Z.A., A data mining approach for diagnosis of CAD. Computer methods and programs in biomedicine, 111(1): p. 52-61, 2013.

[13] Alizadehsani, R., Zangooei, M.H., Hosseini, M.J., Habibi, J., Khosravi, A., Roshanzamir, M., Khozeimeh, F., Sarrafzadegan, N. and Nahavandi, S. CAD detection using computational intelligence methods. Knowledge-Based Systems, 109: p. 187-197, 2016.

[14] Arabasadi, Z., Alizadehsani, R., Roshanzamir, M., Moosaei, H. and Yarifard, A.A. Computer aided decision making for heart disease detection using hybrid neural network-Genetic algo. Computer methods and programs in biomedicine,. 141, p. 19-26, 2017.

[15] Alizadehsani, R., Hosseini, M.J., Khosravi, A., Khozeimeh, F., Roshanzamir, M., Sarrafzadegan, N. and Nahavandi, S. Non-invasive detection of CAD in high-risk patients based on the stenosis prediction of separate coronary arteries. Computer methods and programs in biomedicine,. 162, p. 119-127, 2018.

[16] Abdar, M., Książek, W., Acharya, U.R., Tan, R.S., Makarenkov, V. and Pławiak, P. A new ML technique for an accurate diagnosis of CAD. Computer methods and programs in biomedicine,. 179: p. 104992, 2019.

[17] Dipto, I.C., Islam, T., Rahman, H.M.M. and Rahman, M.A. (2020) Comparison of Different ML algos for the Prediction of CAD. Journal of Data Analysis and Information Processing , 8, 41-68. <https://doi.org/10.4236/jdaip.2020.82003>

[18] Alizadehsani, R., Habibi, J., Sani, Z. A., Mashayekhi, H., Boghrati, R., Ghandeharioun, A., & Bahadorian, B. (2012). Diagnosis of CAD using data mining based on lab data and echo features. *Journal of Medical and Bioengineering*, *1*(1).

[19] Sharma, G., Rani, G., & Dhaka, V. S. (2021, March). Efficient Predictive Modelling for Classification of CADs Using ML Approach. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1099, No. 1, p. 012068). IOP Publishing.

[20] Abdar, M., Acharya, U. R., Sarrafzadegan, N., & Makarenkov, V. (2019). NE-nu-SVC: a new nested ensemble clinical decision support system for effective diagnosis of CAD. *IEEE Access*, *7*, 167605-167620.

[21] Cüvitoğlu, A., & Işik, Z. (2018, May). Classification of CAD dataset by using principal component analysis and ML approaches. In *2018 5th International Conference on Electrical and Electronic Engineering (ICEEE)* (pp. 340-343). IEEE.

[22] Gudadhe, M., Wankhade, K. and Dongre, S. (2010) Decision Support System for Heart Disease Based on Support Vector Machine and Artificial Neural Network. IEEE International Conference on Computer and Communication Technology, Allahabad, 17 September, 2010, 741-745. https://doi.org/10.1109/ICCCT.2010.5640377

[23]Hsu, C. W., Chang, C. C., & Lin, C. J. (2003). A practical guide to support vector classification.

[24] Mohammad M. Ghiasi , Sohrab Zendehboudi , Ali Asghar Mohsenipour , Decision Tree-Based Diagnosis of CAD: CART Model, Computer Methods and Programs in Biomedicine (2020), doi: <https://doi.org/10.1016/j.cmpb.2020.105400>

[25]Alotaibi, S. S., Almajid, Y. A., Alsahali, S. F., Asalam, N., Alotaibi, M. D., Ullah, I., & Altabee, R. M. (2020, October). Automated prediction of CAD using Random Forest and Naïve Bayes. In *2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS)* (pp. 109-114). IEEE.

[26] Sridhar, C., Acharya, U. R., Fujita, H., & Bairy, G. M. (2016, October). Automated diagnosis of CAD using nonlinear features extracted from ECG signals. In *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 000545-000549). IEEE.

[27] Kurt, I., Ture, M., & Kurum, A. T. (2008). Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting CAD. *Expert systems with applications*, *34*(1), 366-374.

[28]Fatima, M., & Pasha, M. (2017). Survey of ML algos for disease diagnostic. *Journal of Intelligent Learning Systems and Applications*, *9*(01), 1.

[29] Dipto, I.C., Islam, T., Rahman, H.M.M. and Rahman, M.A. (2020) Comparison of Different ML algos for the Prediction of CAD. Journal of Data Analysis and Information Processing, 8, 41-68. <https://doi.org/10.4236/jdaip.2020.82003>

[30] Parikh, R., Mathai, A., Parikh, S., Sekhar, G. C., & Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. *Indian journal of ophthalmology*, *56*(1), 45.