

Prediction of Coronary Artery Disease using Machine Learning – A Comparative study of Algorithms

Ramanathan G. ¹ & Jagadeesha S. N. ²

¹ Research Scholar, Institute of Computer Science and Information Science, Srinivas University, Mangalore, Karnataka, India,

ORCID: 0000-0001-9621-3294; Email: ramanathanmca2006@gmail.com

² Research Professor, Institute of Computer Science and Information Science, Srinivas University, Mangalore, Karnataka, India,

ORCID: 0000-0002-5185-2233; Email: jagadeesha2012@gmail.com

Area/Section: Health Management.

Type of the Paper: Comparative Study.

Type of Review: Peer Reviewed as per [|C|O|P|E|](#) guidance.

Indexed in: OpenAIRE.

DOI: <https://doi.org/10.5281/zenodo.10451894>

Google Scholar Citation: [IJHSP](#)

How to Cite this Paper:

Ramanathan, G. & Jagadeesha, S. N. (2023). Prediction of Coronary Artery Disease using Machine Learning – A Comparative study of Algorithms. *International Journal of Health Sciences and Pharmacy (IJHSP)*, 7(2), 180-209. DOI: <https://doi.org/10.5281/zenodo.10451894>

International Journal of Health Sciences and Pharmacy (IJHSP)

A Refereed International Journal of Srinivas University, India.

Crossref DOI: <https://doi.org/10.47992/IJHSP.2581.6411.0116>

Received on: 18/07/2023

Published on: 30/12/2023

© With Author.



This work is licensed under a [Creative Commons Attribution-Non-Commercial 4.0 International License](#) subject to proper citation to the publication source of the work.

Disclaimer: The scholarly papers as reviewed and published by Srinivas Publications (S.P.), India are the views and opinions of their respective authors and are not the views or opinions of the SP. The SP disclaims of any harm or loss caused due to the published content to any party.

Prediction of Coronary Artery Disease using Machine Learning – A Comparative study of Algorithms

Ramanathan G. ¹ & Jagadeesha S. N. ²

¹ Research Scholar, Institute of Computer Science and Information Science, Srinivas University, Mangalore, Karnataka, India,

ORCID: 0000-0001-9621-3294; Email: ramanathanmca2006@gmail.com

² Research Professor, Institute of Computer Science and Information Science, Srinivas University, Mangalore, Karnataka, India,

ORCID: 0000-0002-5185-2233; Email: jagadeesha2012@gmail.com

ABSTRACT

Purpose: *Heart illness is one of the major killers of humans worldwide. Heart illness and the possibility of experiencing a heart attack have both increased in recent years. Medical professionals face significant difficulties when attempting to forecast heart disease. One of the medical field's virtuosi is early prediction, and this is particularly true in cardiology. The early prediction model-building studies illuminated the most up-to-date methods for locating variations in medical imaging. The study of computer-assisted diagnosis is a dynamic and quickly developing field. Since wrong medical diagnoses can lead to dangerous treatments, a lot of work has been done recently to enhance computer programs that help doctors make diagnoses. Computer-assisted diagnosis relies heavily on machine learning. The basic aspect of pattern recognition is the capability to learn from precedents. Pattern identification and artificial intelligence have a lot of promise to improve the accuracy with which biomedical professionals perceive and diagnose illness. They also help make decisions more objectively. Machine learning is a promising method for developing elegant and automatic algorithms for the study of high-dimensional and multimodal bio-medical data. Two heart disease-related datasets were considered for the purpose of this research. The study implements several machine learning algorithms and compares their prediction accuracy and a handful of other performance metrics to determine which one is the most effective.*

Objective: *The primary goal of the research is to evaluate the performance of several machine learning algorithms using different evaluation criteria such as f1 score, roc, and auc values. The aim is to discover the most effective machine learning algorithm for the datasets obtained for the study.*

Design/Methodology/Approach: *The research utilizes datasets from Kaggle heart information. Python, Skilearn, Pandas, and Jupyter Notebook have been used to build various machine learning prediction models and the outcomes have been compared.*

Findings/Results: *Both datasets comprise of different parameters, therefore pre-processing had to be customized. Applying machine learning algorithms to the training dataset and comparing the trained models to the testing dataset yielded varied results for each dataset. Model performance was measured by accuracy and AUC. Both datasets gave good results with boosting algorithms, however the Cleveland dataset did better with decision trees.*

Originality/Value: *The research included an examination of two Kaggle heart databases. It has been seen how data is distributed, how various features depend on each other, and how all the features influence the target feature of heart disease prediction. Models have been constructed and trained using different machine learning methods, each with its own set of hyper-tuning parameters. To learn which machine learning model is most effective for a given collection of data, the study has looked into both the prediction results using the trained models and the performance parameters of the individual models. Through this study, we now know more about how different machine learning methods work. To determine the most effective*

algorithm, it is necessary to conduct additional research of the datasets using Deep Learning techniques.

Paper Type: Comparative Study

Keywords: Cardiovascular diseases, Diagnosis of Coronary heart disease, Artificial Intelligence, Machine Learning Algorithm.

1. INTRODUCTION :

The human heart is both the most important and complicated organ in the body. One hundred thousand times a day, thirty million times a year, and two and a half billion times over the period of a typical human existence are the average number of times the heart beats. The human heart is barely bigger than a fist, but it pumps 7,000 gallons of blood every day, 2.5 million every year, and 200 million in a lifetime. It has four chambers and four valves that work together to precisely control the filling, ejecting, and general pump function owing to the interaction of electrical and mechanical fields. The electrical and mechanical fields interact to regulate the filling of the chambers, and the coordinated opening and shutting of these valves regulate their correct ejection. Heart problems, such as valvular stenosis, valvular regurgitation, ventricular arrhythmias, and heart failure, can have critical physiological effects [1].

Diseases of the heart are grouped together by the generic phrase "heart disease," which includes conditions as varied as heart failure, coronary artery disease, arrhythmias (abnormal heart rhythms), and valve disease. These conditions can lead to a variety of symptoms and complications, ranging from mild discomfort to life-threatening emergencies.

Heart disease is a major cause of mortality in the current times. High blood pressure is a risk factor for cardiovascular disease, and risk factors include unhealthy lifestyle choices like smoking, drinking alcohol, and having a diet high in fat. The World Health Organization says that more than 10 million persons die every year from cardiovascular disease. Living a healthy lifestyle and getting checked regularly is the only reliable way to prevent cardiovascular disease.

The greatest challenge in today's healthcare system is providing timely, accurate diagnoses and the best possible treatment to patients. Heart disease has become the leading cause of death worldwide, yet it is also one of the easiest to manage. Diagnosis at the earliest possible time is the single most important factor in successful disease treatment.

It can be difficult to identify and diagnose coronary heart disease in its early stages, so computer-aided techniques have been developed. Machine learning, a type of Artificial Intelligence that processes and analyses clinical data in order to make diagnoses for medical conditions, is becoming increasingly popular among computer-aided detection techniques in medical institutions. Artificial intelligence (AI), machine learning (ML), and deep learning (DL) are set to revolutionize nearly every aspect of human life, including the treatment of cardiovascular disease. As modern computers are capable of millions of computations per second, more complex ML systems are now possible, bringing AI closer to human intelligence [2].

Current methods for determining the extent of heart illness include electrocardiograms (ECGs or EKGs), echocardiograms, exercise stress tests, chest X-rays, cardiac catheterization, coronary angiograms, coronary artery calcium scans, and cardiac magnetic resonance imaging (MRI) [3]. By analysing, processing, and interpreting patient data, ML technology facilitates medical diagnoses. Clinical data-based heart disease analysis has been studied using a wide diversity of machine learning and artificial intelligence techniques, including decision trees, artificial neural networks, support vector machines, fuzzy neural networks, ensemble machine learning, binary particle swarm optimization, random forest classifier, evolution classifier based on principal components analysis, Bayesian algorithms, and neuro fuzzy classifiers [2].

This research aims to compare the accuracy and other performance metrics of several supervised ML algorithms used to predict the occurrence of heart disease in an individual. To determine the best-performing model, the research compares the algorithms' results on two different data sets.

2. OBJECTIVES :

The research is being conducted to comprehend the datasets and their constituent parameters. In addition, the relationship between the different parameters plays a significant role in predicting the objective variable. Understanding how each algorithm operates and the accuracy of its predictions is

more essential in determining the optimal algorithm for the problem under study. The primary objectives of the investigation are as listed below.

- (1) Identifying and understanding the different characteristics of the collected data.
- (2) Determining the relationship that exist among the different features in the dataset and the effect of each of the parameters on the target variable.
- (3) Pre-processing the dataset to handle missing values, imbalance in the data distribution.
- (4) Applying the different machine learning algorithms on the pre-processed datasets and understanding how well the trained models work depending on the tuning parameters for each model.
- (5) Evaluating the performance of the various trained models depending on the prediction accuracy and other metrics like f1 score, roc, and auc values and identifying the best-performing machine learning model for the datasets considered for the study.

3. CORONARY ARTERY DISEASES AND THEIR RISK FACTORS :

The common type of heart disease is coronary artery disease (CAD). Ischemic heart disease, also termed as arterial heart disease, is another name for this condition. Angina or a heart attack may be the first symptom of CAD in some individuals. Plaque accumulation within the artery walls that carry blood to the heart (called coronary arteries) is the root cause of coronary artery disease. Cholesterol builds up over time and forms plaque. Plaque builds up inside the vessels and gradually makes them narrower. This is called atherosclerosis in the medical world [3].

There are different symptoms of heart disease. The following are common symptoms: [8]

- Angina or chest pain
- Arrhythmia or irregular heart beat
- Heart failure
- Sudden cardiac arrest
- Pain in the arms, jaws, neck and back
- Fatigue
- Difficulty in breathing
- Anxiety or nervousness, light-headedness
- Heart palpitations
- Nausea

The following are the different types of CAD [4].

- Obstructive CAD – Obstructive CAD refers to the progressive narrowing or closing of coronary arteries which are responsible for supplying blood to the heart. This blockage is due to plaque build-up [5].
- Non – Obstructive CAD – Despite its rarity, non-obstructive CAD is just as deadly as its obstructive counterpart. This develops when cardiac muscle presses on the heart's vessels, or when the heart's arteries are otherwise compromised. Numerous complications can emerge, including injury to the arterial inner lining, improper constriction, dysfunction in the minor branches, and pressure from the overlying cardiac muscle [6].
- Spontaneous CAD – A spontaneous coronary artery dissection is a medical word for a tear or separation in the coronary artery walls. A tear can occur in any of the three segments that make up the coronary artery. Blood starts leaking from the crevices. The confined blood exerts inward pressure on the artery, causing it to expand inward. This causes a slowing or stopping of blood supply to the heart. Damage or death to cardiac tissue can result from reduction of blood flow to the heart [7].

The likelihood of getting CAD can be reduced or avoided altogether by being aware of the risk factors associated with it. Age also raises the danger of developing CAD. Men are at a higher risk for the illness starting at age 45, while women are at a higher risk starting at age 55, depending on age alone as a risk factor. Having a family history of CAD also increases the chance for developing the condition [8]. The following are a few of the risk factors:

- high blood cholesterol levels
- obesity

- inactivity
- insulin resistance /diabetes mellitus/ hyperglycemia
- tobacco smoking
- emotional stress
- high blood pressure
- excessive alcohol consumption
- unhealthy eating habits
- history of preeclampsia during pregnancy
- obstructive sleep apnea

4. LITERATURE REVIEW :

Extensive study into the capability of machine learning and deep learning for CAD prediction has yielded numerous journal papers detailing the results of the study. The following are brief summaries of some of the published study articles.

Table 1: Scholarly Literature on ML for prediction of CAD published between 2018 - 2022

S. No.	Area & Focus of the Research	The outcome of the Research	Reference
1	Classification and Prediction of CAD using ML and DL Techniques	In this review article, researchers evaluate and contrast different methods for classifying and predicting CVD. Prediction methods for CVD using data mining, ML, and DL techniques are discussed.	Swathy, M et al., (2022). [11]
2	Prediction of CVD using Machine Learning Classification	The group has developed a system for identifying cardiac disease using machine learning. Methods like FCMIM, LASSO, LLBFS, MRMR, and LASSO are employed to solve feature selection issues. The system selects hyperparameters with LOSO cross-validation.	Saboor, A et al., (2020). [12]
3	Clinical Applications of ML Algorithms	If adequate resources and time are spent explaining how ML algorithms work, they have the capability to improve clinical knowledge and patient care.	Watson, D. S et al., (2019). [13]
4	Heart Disease Prediction using Random Forest Algorithm	An evaluation of 303 samples and 14 different characteristics has been performed. Data processing and classification are performed by random forest. Percentages are used to illustrate the accuracy, sensitivity, and specificity of a dataset. The receiver operating characteristics show that random forest can precisely predict coronary disease with a 93.3% success rate.	Pal, M et al., (2021). [14]
5	Heart Disease Prediction using Particle Swarm Algorithm	The study presents PM-LU algorithm, a hybrid of the Lion Algorithm (LA) and the Particle Swarm Optimization (PSO) algorithm. The aim of this paper is to improve the accuracy of forecasts. Finally, existing methods are compared to the suggested work, and	Cherian, R. P et al., (2020). [15]

		its superiority in some performance criteria is demonstrated.	
6	Explainable Machine Learning in Cardiology	The article explores the application of explainable machine learning in the field of cardiology. There are many restrictions that can be applied to methods that prioritize explainability. The paper concludes with a general rule for employing black-box models with justifications.	Petch, J et al., (2021). [16]
7	Challenges in Translation of AI to clinical practice	The primary challenges and restrictions of AI in healthcare, and the steps required to bring these technologies out of the lab and into actual practice are discussed.	Kelly, C. J et al., (2019). [17]
8	AI in Cardiology	Data-driven decision-making is essential for cardiologists, and the field has access to more quantifiable data than most others. Using AI, medical professionals will be able to analyze more data, leading to better patient care.	Johnson, K. W et al., (2018). [18]
9	Clinical Decision Support system using SPM for CVD	The study improves sequential pattern mining and association rules to create a clinical decision support system for predicting cardiovascular illness. The highest disease prediction accuracy (92.101%) is achieved by the SPM with ARM algorithm.	Harini, C et al., (2021). [19]
10	Prediction of Stress and Anxiety using ML Algorithms	Anxiety, depression, and stress levels were calculated using five different machine learning methods. Significant reasons for Anxiety, Depression, and Stress were identified.	Priya, A et al., (2020). [20]

Swathy, M et al. [11] discuss that obesity, cholesterol, high blood pressure, and cigarette use are major causes of CAD and mortality in the population. Artificial Intelligence and Data Mining have a research scope with its huge techniques that would help predict CVD priority and find behavioural patterns in big volumes of data. The authors compare data mining, classification, machine learning, and deep learning models used to predict CVD.

Saboor, A et al. [12] propose a system based on classification methods including Support vector machine, Artificial neural network, Logistic regression, Naive bays, K-nearest neighbour, and Decision tree. Feature selection methods such as Relief, Least absolute shrinkage selection operator, Local learning, and Minimal redundancy maximal relevance are used to remove irrelevant and redundant features. Results of the experiment suggest that the proposed feature selection technique is practical for creating an advanced intelligent system to spot heart disease. The suggested method for diagnosing was more accurate than the methods that had been used before. With the suggested method, heart problems are easy to spot in healthcare.

Watson, D. S et al., [13] argue that rethinking explanation will maximize the clinical benefits of machine learning algorithms. Popular machine learning algorithms are black boxes that issue conclusions without rationale. Patients, doctors, and data scientists must create new approaches to extract model-centric and patient-centric explanations for global and local comprehension.

Pal, M et al., [14] offer a strategy for foretelling the different categories of heart disease. The random forest algorithm has been utilized to make accurate predictions about the occurrence of heart illness.

The sensitivity number achieved in the experiments was 90.6%. The value of the specificity is 82.7, and the prediction accuracy is 86.9.

Cherian, R. P et al., [15] suggest a new model for predicting cardiovascular disease by utilising methods such as Feature Extraction, Recording, Attribute Minimization, and Classification. Feature extraction begins with the collection of basic statistical and higher-order statistical characteristics. The "curse of dimensionality" can be alleviated using Component Analysis PCA, which is then applied during the record and attribute reduction process. The dimensionally reduced features are then fed into a Neural Network (NN) model, which performs the prediction.

Petch, J et al., [16] state there are 12 cardiology-approved machine learning algorithms. In health care, where many decisions are life-or-death, uninterpretable predictive algorithms can damage faith. Explainable ML seeks to explain how a model works and why it makes specific predictions. The study explains the merits and limits of explainable ML to cardiologists and cardiovascular researchers.

Kelly, C. J et al., [17] state that AI affects care quality, healthcare professional variability, efficiency and productivity, and patient outcomes. Independent datasets indicative of future target populations should be curated to compare algorithms while checking for bias and fitting to unexpected variables. AI tool developers must be aware of unforeseen repercussions and construct algorithms with the global community in mind.

Johnson, K. W et al., [18] in their study discuss cardiology-related predictive modelling principles including feature selection and dichotomization. They explore supervised learning methods and cardiology applications. Deep learning and related technologies, generally dubbed unsupervised learning, present contextual examples in general medicine and cardiovascular care, and show how these methods could be applied to enable precision cardiology and enhance patient outcomes.

Harini, C et al., [19] propose a new and improved sequential pattern mining algorithm (Two phase) that utilizes the association pattern mining (APM) technique. The model has used discriminant analysis (DA) to determine whether the symptom-based data clusters were statistically significant. The model then combines sequential pattern mining with association pattern mining to create a clinical decision support system. Next, comparisons are made using assessment tools.

Priya, A et al. [20] state that an adequate learning algorithm is needed to accurately diagnose anxiety, sadness, and stress. Five machine learning algorithms predicted anxiety, depression, and stress on five severity levels. Because they are accurate and well-suited to forecasting psychological issues. The f1 score measure helps identify the best accuracy model as the Random Forest classifier.

5. MACHINE LEARNING FOR THE PREDICTION OF HEART DISEASES :

Machine learning is used to train computers to perform better with data. Learning from data is machine learning's core purpose. How to create self-learning robots has been the subject of a lot of research. Numerous researchers and computer scientists employ various strategies to solve this puzzle. Machine learning is used by many sectors, from the medical to the military, to mine large datasets for useful insights [9].

An enhanced technique of detecting CAD could have significant clinical implications. We use machine learning methods to integrate genetic, epigenetic, and phenotype data, depending on the hypothesis that the systemic effects of CAD risk factors are the outcome of a complex interplay of genetic and environmental variables [10].

a. Machine learning algorithms used in the study:

In these ever-changing times, many varieties of ML algorithms have been developed to aid in the resolution of complex issues encountered in the real world. Automated and self-improving, ML algorithms are constantly evolving to meet new challenges. Knowing the various machine learning algorithms and how they are categorised is essential. The major categories of ML algorithms are listed below:

- Supervised learning
- Unsupervised learning
- Semi – supervised learning
- Reinforcement learning
- Ensemble Learning
- Neural Networks
- Instance based learning

The following figure (Figure 1) shows the different algorithms that fall under individual categories of ML listed above:

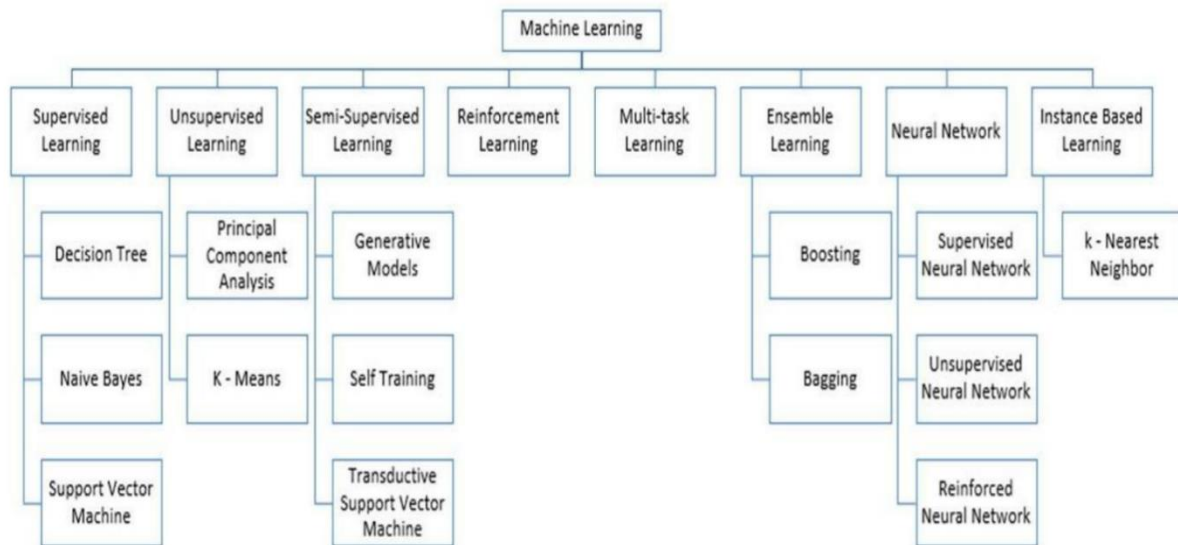


Fig. 1: Categories of Machine Learning Algorithms [9]

This section describes the different ML algorithms applied in the study briefly. As part of the study, the algorithms listed below have been used to build prediction models.

- ✓ Logistic Regression
- ✓ Decision Tree
- ✓ Random Forest
- ✓ Adaptive Boosting
- ✓ Gradient Boosting
- ✓ Extreme Gradient Boosting
- ✓ Light Gradient Boosting
- ✓ Support Vector Machine
- ✓ K – Nearest Neighbour

Logistic Regression Algorithm

In the realm of Machine Learning, Logistic Regression is employed to resolve problems of categorization. It is a form of probabilistic analysis used to make predictions. To foretell the probability of a categorical dependent variable, the classification algorithm Logistic Regression is employed. The purpose of Logistic Regression analysis is to establish a connection between observable features and the probability of a target result which is a binary value of 0 or 1. In contrast to a Linear Regression model, which uses a simple linear cost function, a Logistic Regression model makes use of a more complex cost function, the Sigmoid function or logistic function. The Sigmoid function is represented as below:

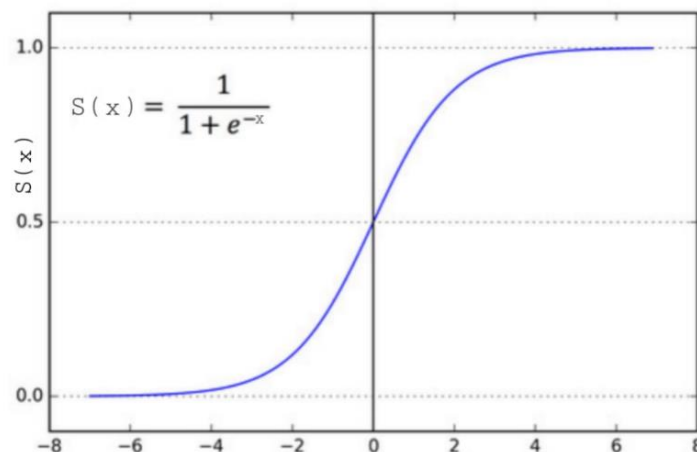


Fig. 2: The Sigmoid Function for Logistic Regression [21]

Decision Tree Algorithm

A decision tree is a diagram that shows the various paths that can be taken to reach a conclusion. In the instance of supervised learning algorithms, tree-based algorithms are currently the most common option. Both classification and regression complications are amenable to decision tree methods. Nonetheless, you'll find that their primary function is for solving classification problems. It is structured like a tree with a root node, branches, internal nodes, and leaf nodes. In cases where the dependent factors are continuous, a Regression tree can be constructed. When the dependent variables are categorical, a Classification tree is used. The same greedy top-down strategy is used in both. The decision-tree method is shown here in its simplest form.

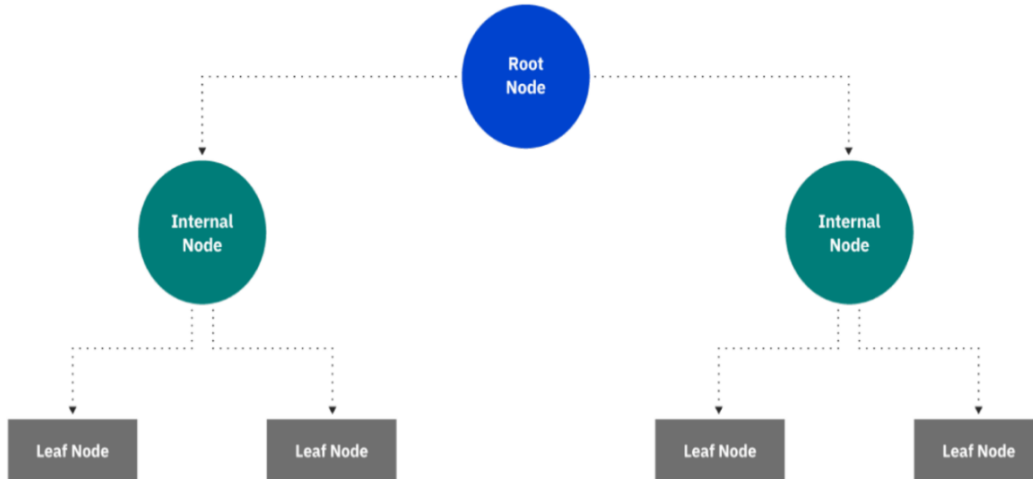


Fig. 3: Representation of Decision Tree Algorithm [22]

Random Forest Algorithm

Machine learning ensemble models aggregate the conclusions drawn by several different models to achieve improved outcomes. Bagging is a method that takes a generalized outcome by combining the results of several models (such as all decision trees). Random Forest is a bagging-style ensemble method aimed at machine learning. In the occurrence of a classification issue, it constructs several decision trees using various samples and then uses the results of the consensus to decide. Different samples are drawn at random from the whole collection. Rather than using the same collection of features to determine the best split at each node of the decision tree, random forest picks them at random. In conclusion, Random Forest constructs numerous trees from data points and features chosen at random (Forest). This is an illustration of the Random Forest Algorithm.

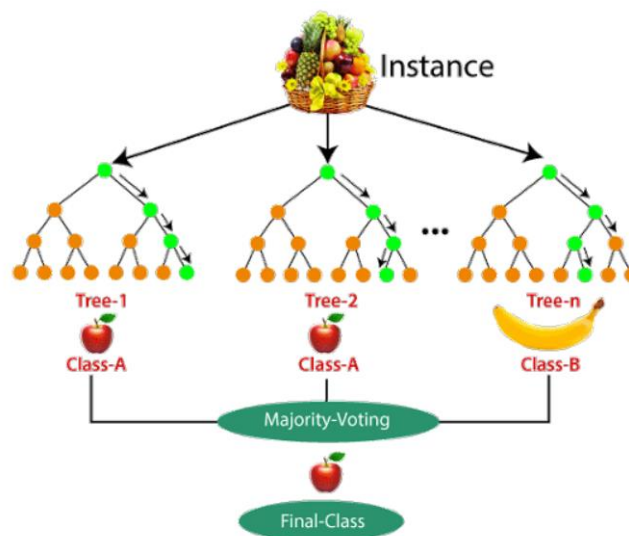


Fig. 4: Illustration of Random Forest Algorithm [23]

Adaptive Boosting Algorithm

Boosting is a method implemented in machine learning ensemble models. Machine learning practitioners use boosting to progress the precision of their predictive analyses of data. In direction to lessen the number of mistakes made during training, boost learning merges several weak learners into one robust one. In boosting, a subset of the available data is chosen at random, a model is fitted to it, and the models are trained successively, with each subsequent model attempting to compensate for the shortcomings of the previous one. Over time, the weak rules generated by each classifier are merged into a single, robust rule for making predictions. One of the easiest boosting algorithms is adaptive boosting, also known as AdaBoost. Decision trees are frequently used in modelling. As each successive model is built, its mistakes are fixed. To improve its predictions, AdaBoost gives weights to the wrongly predicted data.

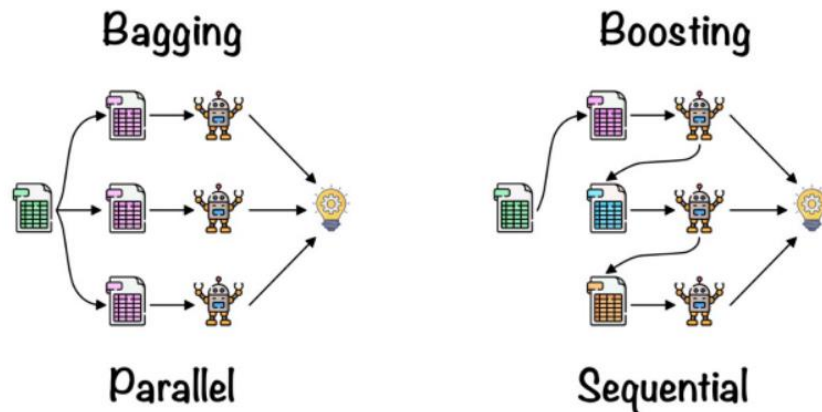


Fig. 5: Difference between Bagging and Boosting Ensemble Techniques [23]

Gradient Boosting Algorithm

Another ensemble machine learning method that can solve both regression and classification issues is gradient boosting, also known as GBM. GBM employs the boosting method, which combines several weak learners into a single robust one. When using regression trees as a learner, the errors from one tree are used to inform the construction of the next tree in the sequence. Gradient boosting, in contrast to AdaBoost, trains on the residual errors of the prior predictor rather than changing the weights of individual data points. The gradient descent algorithm is combined with the boosting technique to form the gradient boosting technique.

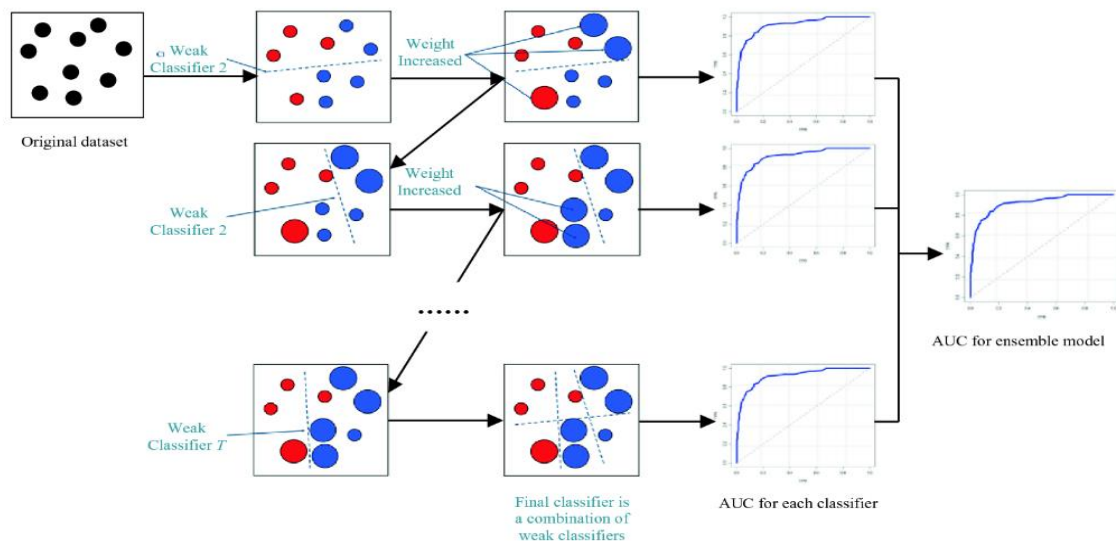


Fig. 6: Working of Gradient Boosting Algorithm [24]

Extreme Gradient Boosting Algorithm

Extreme Gradient Boosting (XGBoost) is a supervised ensemble learning method. The XGBoost machine learning library is a scalable, distributed gradient-boosted decision tree (GBDT) library. It is

the premier machine learning library for regression, classification, and ranking problems and offers parallel tree boosting. It can manage various data types and can be tailored to specific requirements. Instead of the real class designations, residuals are used to build the trees. Consequently, despite the emphasis on classification problems, the base estimators in these algorithms are regression trees, not classification trees. This is since residuals are continuous, not discrete. The algorithm permits control over the maximum size of the trees in order to minimise the danger of data overfitting. Similar to Random Forest and Adaptive Boosting, the algorithm constructs a large number of trees. Ultimately, the ultimate forecast is derived from all of the trees. The value of each tree is proportional to the rate of learning. This allows for a more gradual and consistent improvement at each phase of the process.



Fig. 7: Features of Extreme Gradient Boosting Technique [28]

Light Gradient Boosting Algorithm

Models built with Gradient Boosting Decision Tree (GBDT) or GBDT-based XGBoost fall short on efficiency and scalability as data feature sizes grow. For this particular behaviour, the primary reason is that each feature must analyse all the different data instances to estimate all the possible split points, which is extremely time-consuming and laborious. The Light GBM or Light Gradient Boosting Model is used to overcome this issue. It employs two techniques: GOSS (Gradient-Based on Side Sampling) and EFB (Exclusive Feature Bundling). GOSS will actually exclude the vast majority of statistics facts with small gradients and use only the outstanding figures to evaluate the aggregate information gain. With the EFB, mutually exclusive features are placed alongside nothing, but it will rarely accept a non-zero value to decrease the number of features. Other boosting algorithms divide the tree depth-wise or level-wise as opposed to leaf-wise, whereas Light GBM divides the tree leaf-wise. In other terms, while other algorithms grow trees horizontally, Light GBM grows trees vertically. Light GBM is sensitive to overfitting and can therefore overfit small data with relative ease.

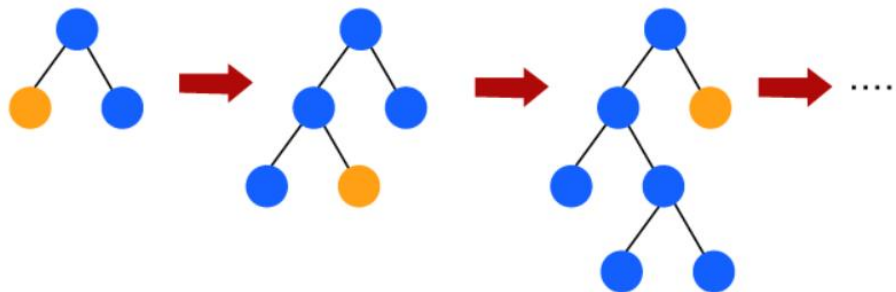


Fig. 8: Light Gradient Boosting Model (Leaf-wise split) [29]

K Nearest Neighbour Algorithm

The k-nearest neighbours algorithm, commonly referred to as KNN or k-NN, is a supervised learning, non-parametric classifier that employs proximity to make predictions or classifications regarding the grouping of a single data point. Although it can be used for both regression and classification problems, it is typically employed as a classification algorithm, based on the premise that similar points are typically found in close proximity. In the healthcare industry, KNN is used to predict the likelihood of

heart attacks and prostate cancer. The algorithm computes the most probable gene expressions. The KNN algorithm typically suffers from the constraint of dimensionality, meaning it performs poorly with high-dimensional data inputs. Due to the "curse of dimensionality", KNN is also more susceptible to overfitting. The KNN algorithm locates the nearest neighbours of a agreed data set. The algorithm then examines the labels of k adjacent points to generate a classification prediction. Here, k is a constraint that can be modified to construct the KNN model.

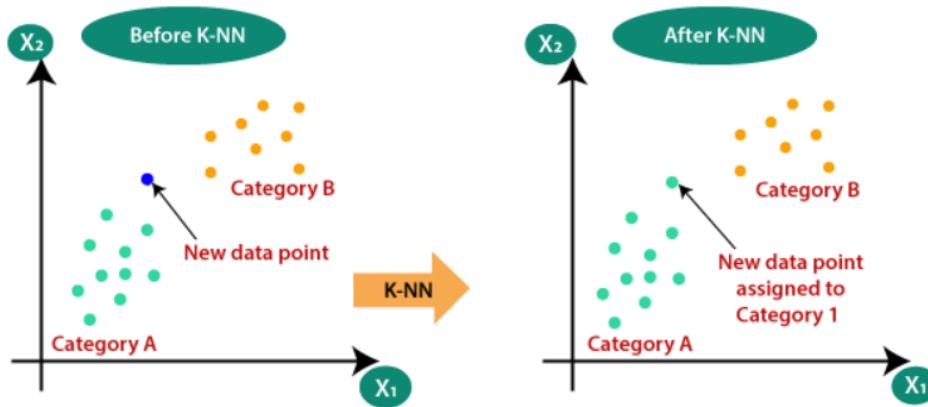


Fig. 9: K Nearest Neighbor Algorithm [30]

Support Vector Machine

The Support Vector Machine is a potent supervised algorithm that performs best on small but complex datasets. Support Vector Machine, abbreviated as SVM, can be used for both regression and classification tasks, but works best for classification problems. Similar to logistic regression, the algorithm attempts to recognize a hyperplane that best separates the two classes. Both algorithms seek to categorize the optimal hyperplane, but logistic regression takes a probabilistic approach while support vector machine relies on statistical methods. SVM accomplishes this by determining the utmost boundary between the hyperplanes, which corresponds to the greatest distance between the two classes. SVM is insensitive to outliers and can be used to categorize images. SVM performs best with linear data. In SVM, hyper parameter modification is problematic to visualise.

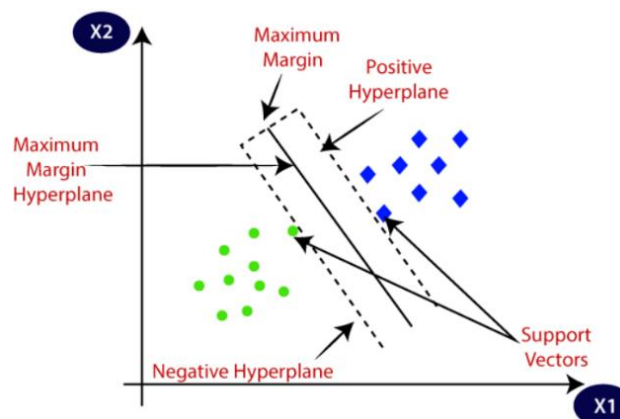


Fig. 10: Support Vector Machine Hyperplane [31]

b. Steps involved in prediction

Using the machine learning method, the following steps are taken to identify heart disease.

- Analysis of the dataset
- Cleaning of dataset
- Pre-processing of data
- Creating the training dataset
- Algorithm implementation and model creation
- Training of model

- Evaluation of model using training dataset
- Determining the best prediction model

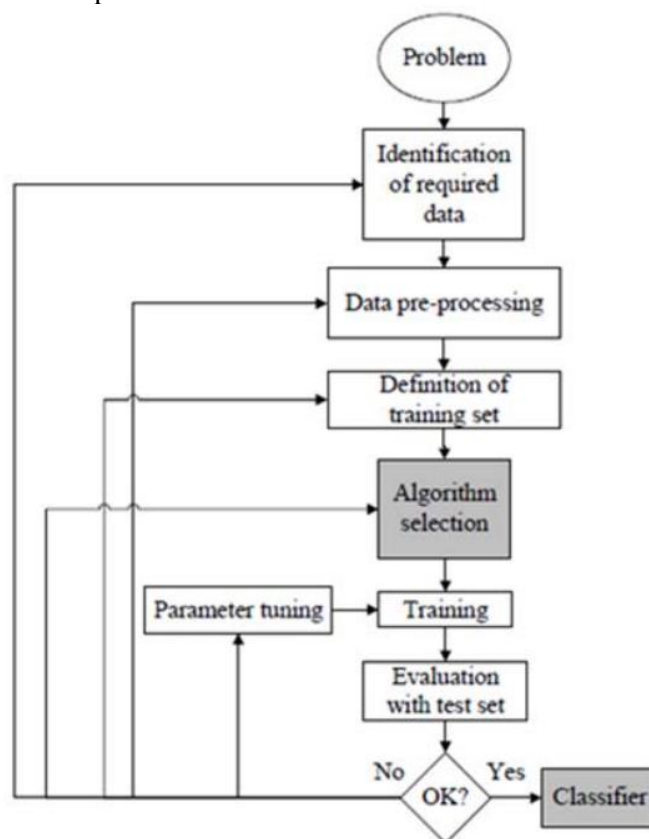


Fig. 11: Workflow of a Machine Learning prediction model [9]

6. DETAILS OF THE DATASETS USED IN THE STUDY :

As part of the research, two datasets have been used for analysis and prediction of CAD. The datasets have been attained from Kaggle. The following tables (Table 2 and Table 3) list the details of the datasets:

Table 2: Parameter Description of the Cleveland Dataset (1025 records) obtained from Kaggle for prediction [32]

S. No.	Parameter Name	Parameter Description
1	Age	Age of the person in years
2	Sex	Sex of the person (1 = male, 0 = female)
3	ChestPain	Chest Pain experienced by the person 1 = Typical angina 2 = Atypical angina 3 = Non-anginal pain 4 = Asymptomatic
4	RestingBPS	Resting blood pressure of the person (mm Hg)
5	Cholesterol	Cholesterol measurement of the person (mg/dl)
6	FastingBS	Fasting blood sugar of the person (>120 mg/dl) (1 = true, 0 = false)
7	RestECG	Resting electrocardiographic measurement 0 = Normal 1 = Having ST-T wave abnormality 2 = Probable or definite left ventricular hypertrophy by Estes' criteria
8	MaxHRate	Person's maximum heart rate achieved

9	ExerAng	Exercise induced angina (1 = yes, 0 = no)
10	STDepress	ST depression induces by exercise relative to rest
11	slope	Slope of the peak exercise ST segment 1 = upsloping 2 = flat 3 = downsloping
12	ca	Number of major vessels (0 – 3)
13	thal	A blood disorder called thalassemia characterised by less haemoglobin 0 = Null 1 = Normal blood flow 2 = Fixed Defect (No blood flow in some parts of the heart) 3 = Reversible Defect (A blood flow is observed but it is not normal)
14	target	Presence of heart disease (0 = no, 1 = yes)

Table 3: Parameter Description of the Framingham Dataset (4240 records) obtained from Kaggle for prediction [33]

S. No.	Parameter Name	Parameter Description
1	male	Sex of the person (1 = male, 0 = female)
2	age	Age of the person in years
3	education	Education of the person (Values between 1 – 2 based on the level of education)
4	currentSmoker	Whether the person is a current smoker 0 = No 1 = Yes
5	cigsPerDay	Average number of cigarettes the person smokes per day
6	BPMeds	Whether the person is on blood pressure medication 0 = No 1 = Yes
7	prevalentStroke	Whether the person previously had stroke 0 = No 1 = Yes
8	prevalentHyp	Whether the person was hypertensive 0 = No 1 = Yes
9	diabetes	Whether the person has diabetes 0 = No 1 = Yes
10	totChol	Total Cholesterol level of the person
11	sysBP	Systolic blood pressure of the person
12	diaBP	Diastolic blood pressure of the person
13	BMI	Body Mass Index of the person
14	heartRate	Heart rate of the person
15	glucose	Glucose level of the person
16	TenYearCHD	10 year risk of CHD of the person 0 = No 1 = Yes

The preceding tables (Tables 2 and 3) summarise the parameters involved in the datasets used for analysis in this learning. In the description, the values of the parameters with ordinal properties have been specified. The remaining parameters all have continuous values.

6.1 Analysis And Pre-Processing Of The Cleveland Dataset

Analysis of Cleveland Dataset:

The Cleveland Dataset comprises of 14 biomedical parameters for 1025 patients. There are no null values as part of the dataset.

Plotting a histogram of each parameter gives a better understanding of the dataset. The number of patients are represented along the Y axis in all the histograms shown in figure (Figure 12).

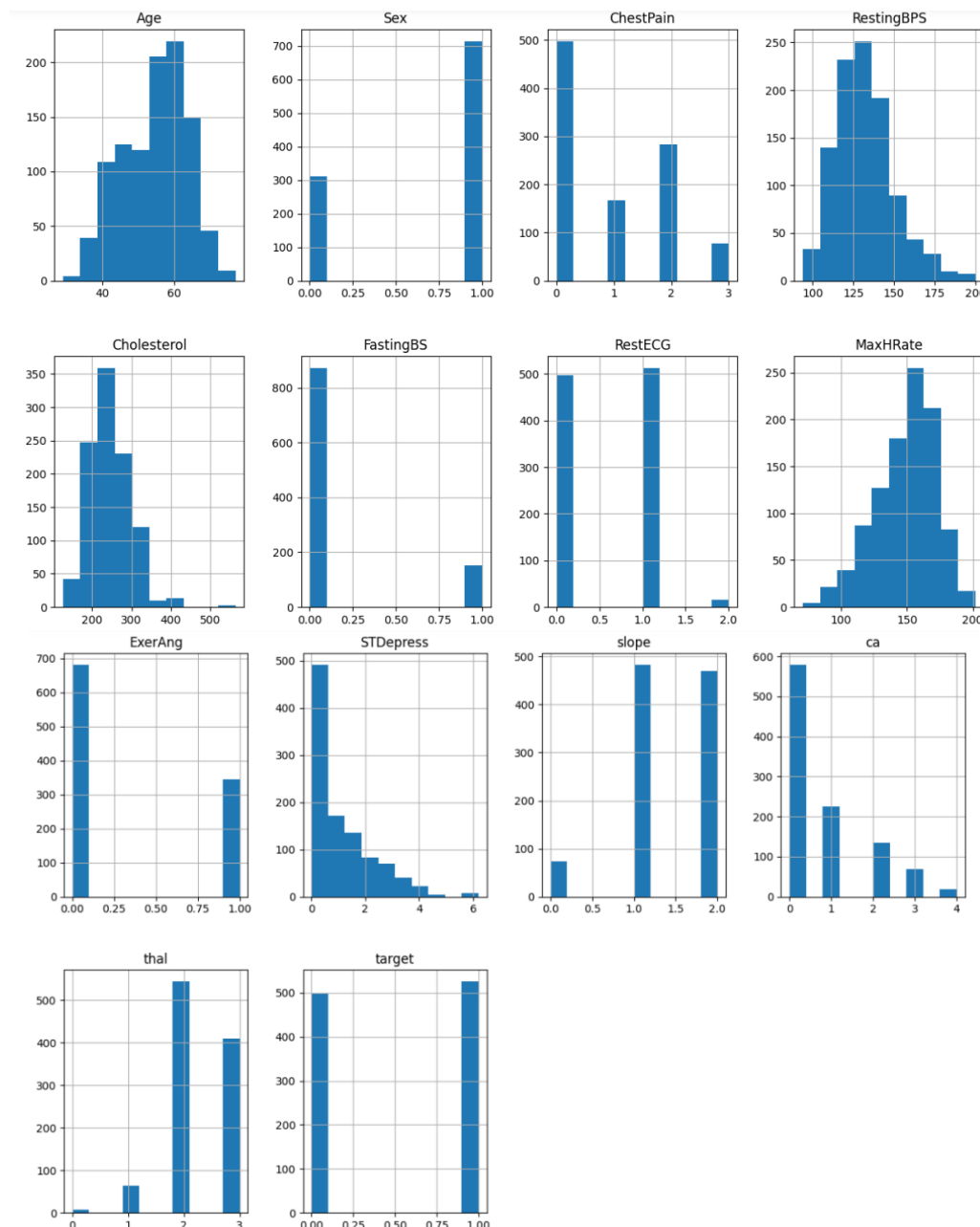


Fig. 12: Histogram plot for each parameter (Cleveland Dataset)

Ordinal characteristics and the objective variable "target" have been stacked in a bar chart for easier analysis. To better understand the distribution, the following ordinal characteristics have been illustrated.

- Sex – 72% of the female population and 42% of the male population have heart related disease.
- Fasting Blood Sugar – Out of the 872 patients having fasting blood sugar < 120, 455 patients have heart related disease. 153 patients have fasting blood sugar > 120 and 71 patients out of them have the disease.

- Presence of Exercise induced angina – Though 680 patients did not have angina or chest pain during exercise, 455 patients out of them have the disease. Only 71 patients out of 345 patients who have chest discomfort during exercise have the disease.

The above details have been illustrated in the stacked bar chart below.

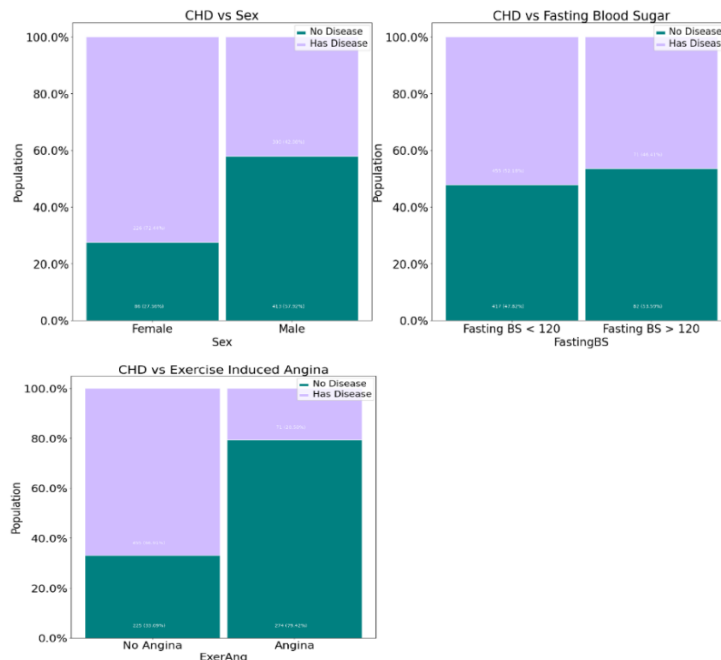


Fig. 13: Stacked bar chart of ordinal features and target variable “target” (Cleveland Dataset)

A heat map is a color-coded, two-dimensional depiction of data. Correlation A heat map is a two-dimensional representation of a statistical measure of dependence (correlation) between categories of data indicated by different colours. The degree of similarity is shown by the colour gradient's intensity. The linear connection among two variables can be measured by calculating their correlation. A correlation heatmap had been employed to figure out how closely related certain parameters are to the target parameter.

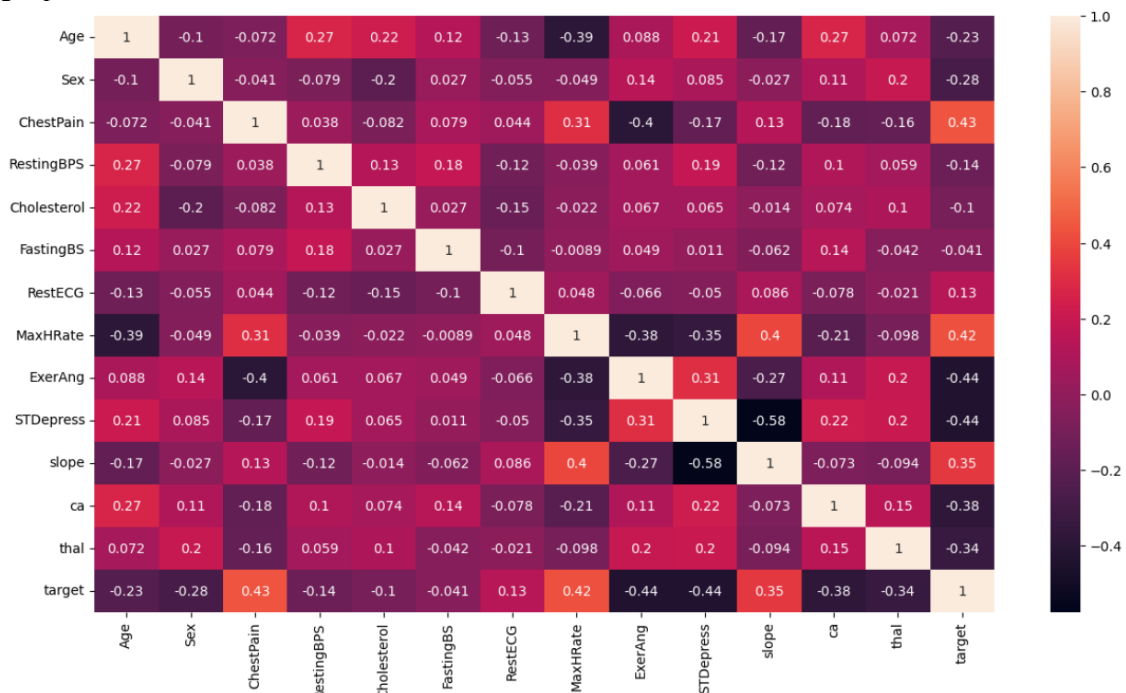


Fig. 14: Correlation Heatmap (Cleveland Dataset)

The "target" parameter is shown to be influenced by all the other factors in the correlation heatmap. Therefore, it is necessary to take into account all of the factors while making CHD prediction using this dataset.

Detecting Outliers

Data points that deviate wildly from the norm are said to be outliers. They are outliers that distort the data and occur when people make mistakes entering information. Since outliers are either extremely low or extremely high values in a dataset, their existence can often affect the findings of statistical analysis on the dataset. This could result in less accurate and valuable predictions being made. Outliers can be detected using different techniques namely:

- Standard Deviation
- Z-score
- Interquartile Range (IQR)
- Percentile

As part of the analysis of Cleveland Dataset, the outliers in the dataset have been detected using Z-score technique.

Z-Score is an effective method for finding and deleting outliers, although it is limited to certain data sets. However, this method is applicable only to perfectly or nearly normally distributed data.

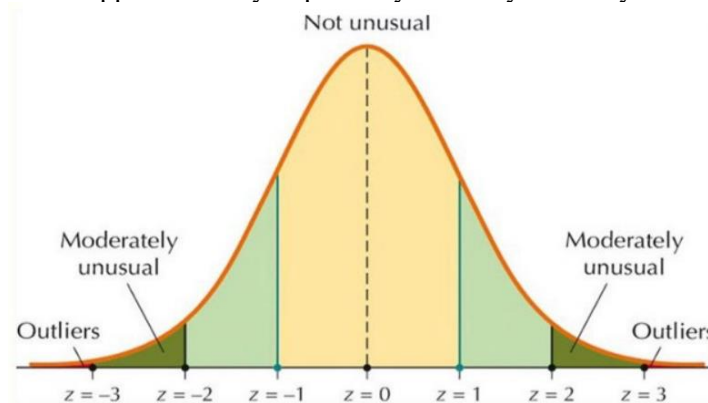


Fig. 15: Detecting Outliers using Z-Score [25]

The histogram plots of the various parameters of the Cleveland dataset shows that the following 5 continuous parameters can be considered for outlier detection using Z-score technique:

- Age
- RestingBPS
- Cholesterol
- MaxHRate
- STDepress

The z-score for a given value x in the dataset, assuming a normal distribution with mean μ and standard deviation σ , is given by:

$$z = (x - \mu) / \sigma$$

Looking at the above equation, the following points can be derived:

- When the value of $x = \mu$, the z-score value is 0
- When the value of $x = \mu \pm 1$, $\mu \pm 2$, or $\mu \pm 3$, the z-score value is ± 1 , ± 2 , or ± 3 , respectively.

Under this transformation, any data points below the *lower limit* of $\mu - 3\sigma$ are transformed into z-scores less than -3. Similarly, on the z-score scale, all locations with coordinates greater than the *upper limit*, $\mu + 3\sigma$, correspond to a z-score greater than 3. Hence all the facts points lying outside these limits become outliers. There are two ways of dealing the outliers namely *trimming* and *capping*.

Trimming - In this method, the outliers are isolated by applying a filter condition that works on any data distribution.

Capping - This method is useful when there are a lot of outliers and it would not be possible to get rid of a lot of data. Then capping comes into play, since it will not get rid of them. Instead, it brings back those data points within the range specified conferring to the z-score value.

While analysing the Cleveland dataset for outliers for the continuous parameters specified above, using the z-score technique, the amount of data points found to be outliers have been shown in the Figure 16. The outliers have been capped at the 99th percentile.

Outlier caps for Age:
 -95p: 68.0 / 43 values exceed that
 -3sd: 81.6 / 0 values exceed that
 -99p: 71.0 / 9 values exceed that

Outlier caps for RestingBPS:
 -95p: 163.2 / 52 values exceed that
 -3sd: 184.1 / 7 values exceed that
 -99p: 180.0 / 7 values exceed that

Outlier caps for Cholesterol:
 -95p: 330.0 / 45 values exceed that
 -3sd: 400.7 / 13 values exceed that
 -99p: 407.0 / 9 values exceed that

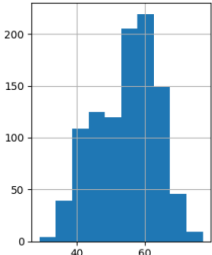
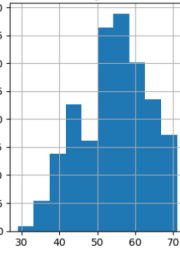
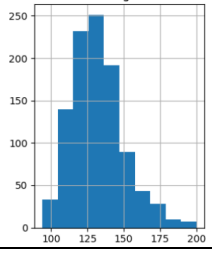
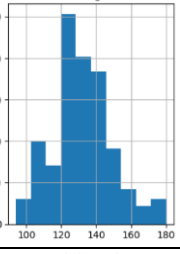
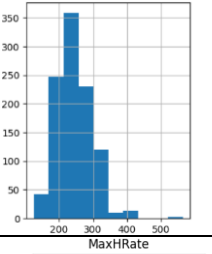
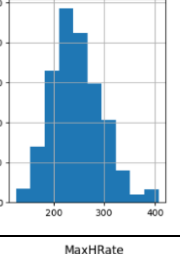
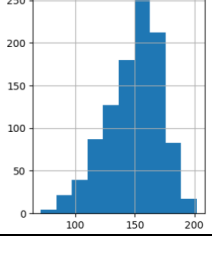
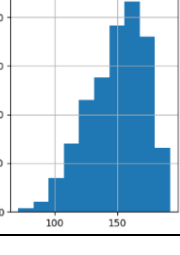
Outlier caps for MaxHRate:
 -95p: 182.0 / 35 values exceed that
 -3sd: 218.1 / 4 values exceed that
 -99p: 192.0 / 10 values exceed that

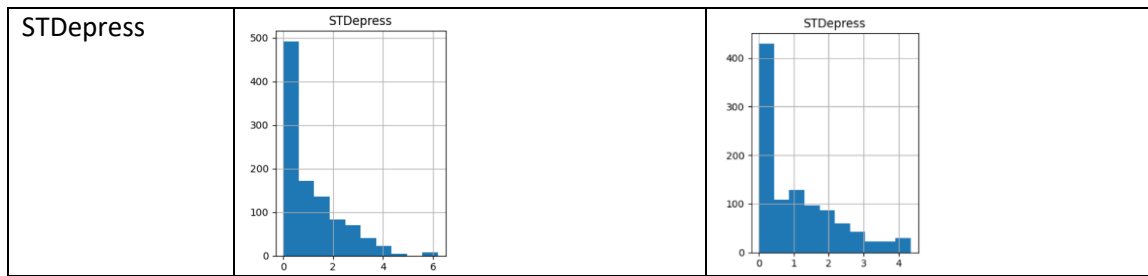
Outlier caps for STDepress:
 -95p: 3.4 / 51 values exceed that
 -3sd: 4.6 / 7 values exceed that
 -99p: 4.4 / 11 values exceed that

Fig. 16: Outliers in the Cleveland Dataset for the continuous parameters

The following table (Table 4) shows the histogram plots of the above-specified continuous features before and after outlier detection and capping.

Table 4: Histogram of the parameters with continuous values before and after outlier capping [Cleveland dataset]

Parameter	Histogram before capping	Histogram after capping
Age		
RestingBPS		
Cholesterol		
MaxHRate		



Data Transformation

Feature Transformation is a part of handling data before it is used. Feature transformation is a technique that must be used no matter what kind of model is being used, whether it is for classification, regression, or unsupervised learning. Feature transformation is a type of mathematical transformation in which a mathematical formula is used to change the standards of a certain feature so that they can be used in further research. This makes the model work better.

A few Machine Learning models, such as Linear and Logistic regression, presume that the factors have a normal distribution. Variables in real datasets are more likely to have a skewed distribution. By changing these skewed factors in certain ways, the skewed distribution can be changed to a normal distribution. This makes the models work better.

There are 3 different types of transformations available as part of scikit-learn:

- Function Transformation
- Power Transformation
- Quantile Transformation

The parameters with continuous values in the Cleveland data have a skewed distribution. The Power Transformation method is used to transform these parameter values. There are two types of Power Transformation technique: the Box-Cox transform and the Yeo-Johnson transform. To make the data distribution less skewed, this study uses the Box-Cox power transformation on parameters with continuous values. Figure 17 shows a histogram that was generated for the values after they were transformed.

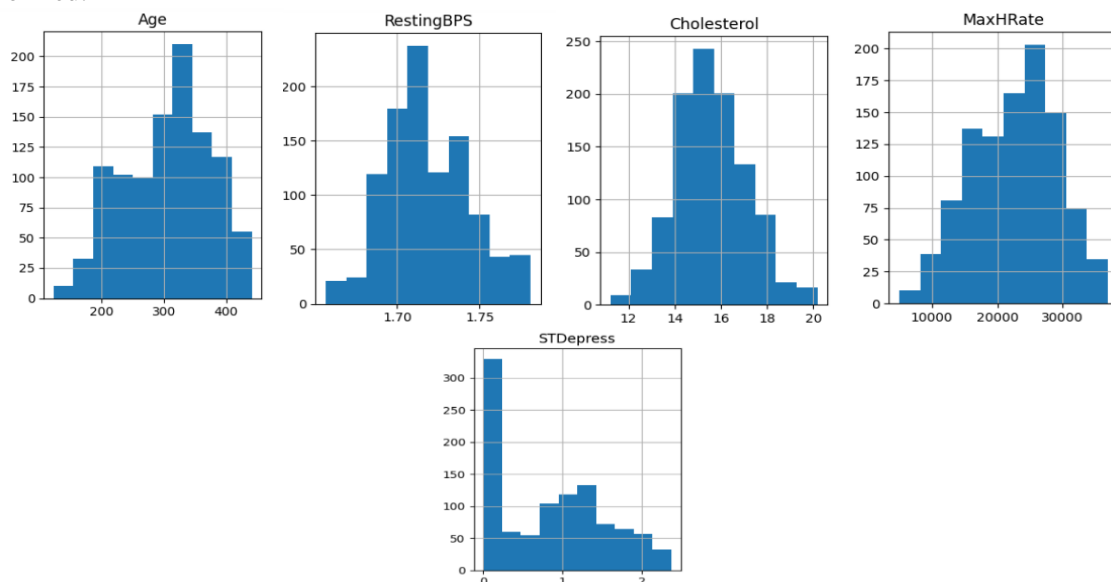


Fig. 17: Histogram of parameters with continuous values after Box-Cox Transformation

6.2 Analysis and Pre-Processing of the Framingham Dataset:

Analysis of Framingham Dataset

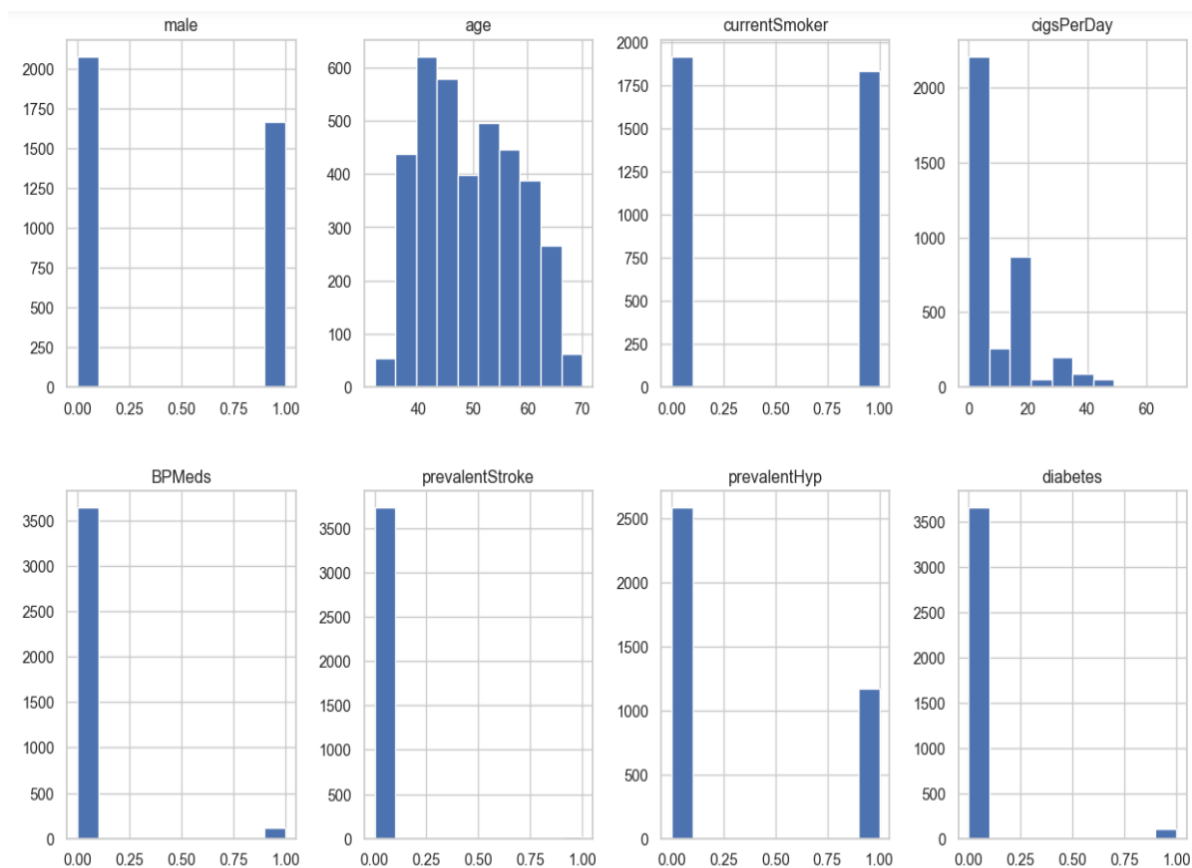
The Framingham Dataset includes 16 variables for 4240 patients. Here's a full analysis of the dataset and a description of how the parameters relate to each other.

The dataset has null values and the total missing data amount to about 12.74% of the overall data. The following table (Table 5) shows the % of missing data for individual parameters.

Table 5: Percentage of Missing values for individual parameters

Parameter	Total Missing values	Percentage
glucose	388	9.150943
BPMeds	53	1.250000
totChol	50	1.179245
cigsPerDay	29	0.683962
BMI	19	0.448113
heartRate	1	0.023585

Since the total entries with missing data represent only 12% of the entire dataset, these entries have been eliminated. The education feature has been dropped since this has no relevance to the study. A histogram has been plotted for each parameter to understand the distribution of the facts in Figure 18. The histogram shows a clear distribution of both the ordinal and continuous parameters in the dataset.



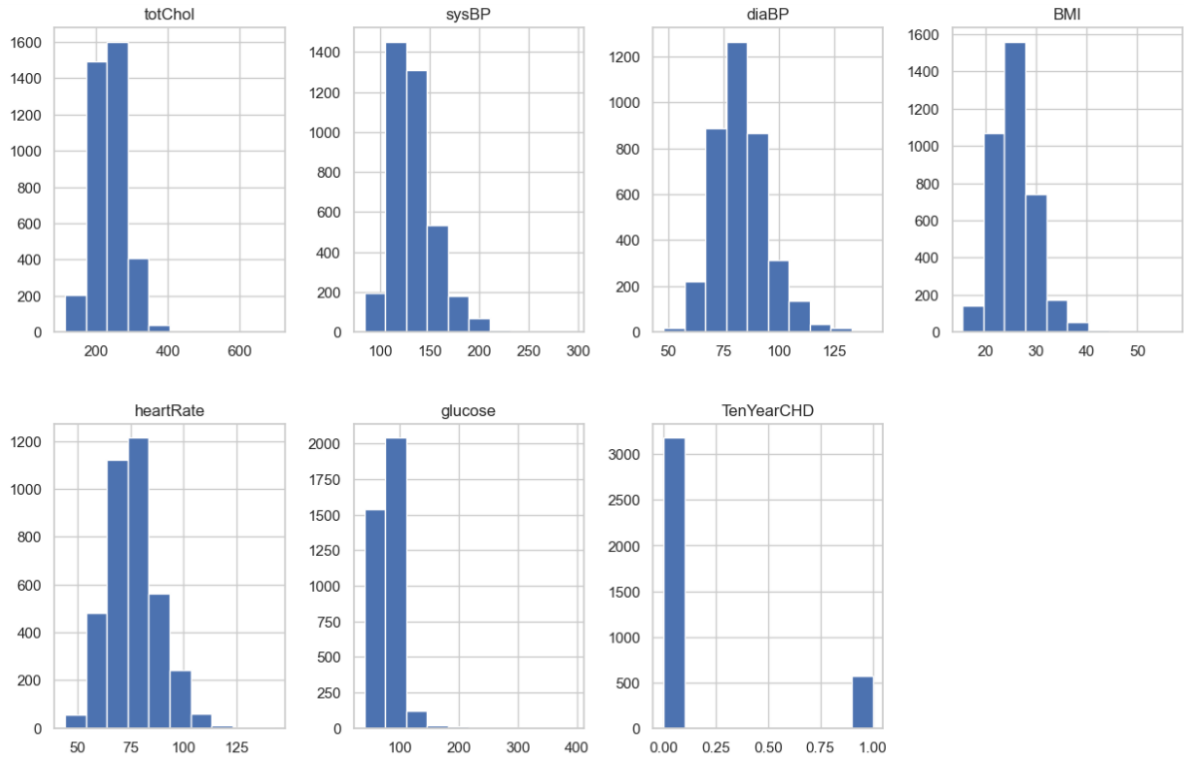
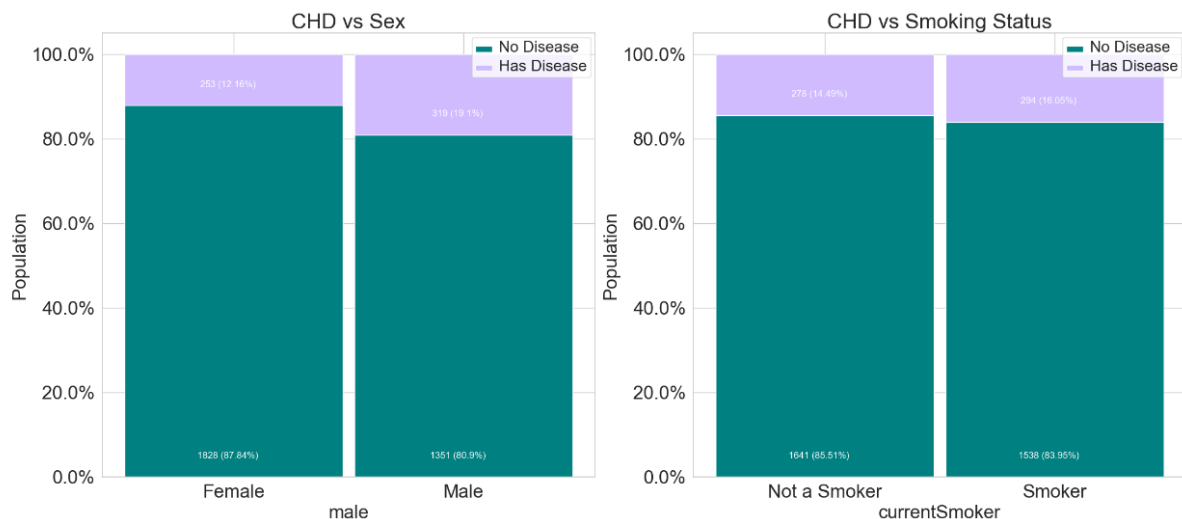


Fig. 18: Histogram plot for each parameter (Framingham Dataset)

To comprehend the association between ordinal features and the objective variable "TenYearCHD", a stacked bar diagram has been generated. The following ordinal characteristics have been depicted to enhance comprehension of the distribution.

- Sex (male)
- Smoking status (currentSmoker)
- Diabetes (diabetes)
- BP Medication (BPMeds)
- Hypertension (prevalentHyp)



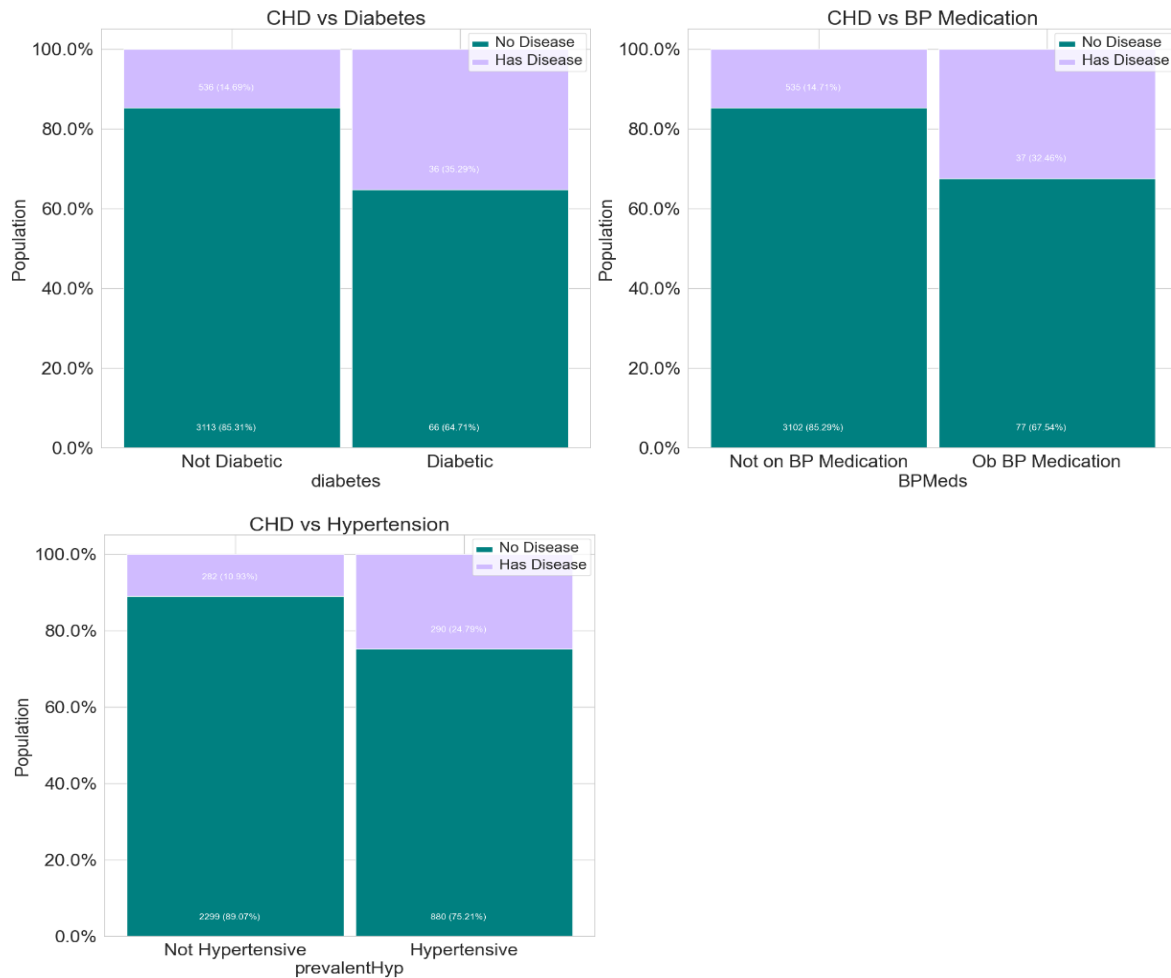


Fig. 19: Stacked bar chart of ordinal features and target variable “TenYearCHD” (Framingham Dataset)

The correlation between the multiple parameters and the target parameter has been determined using a correlation heatmap.

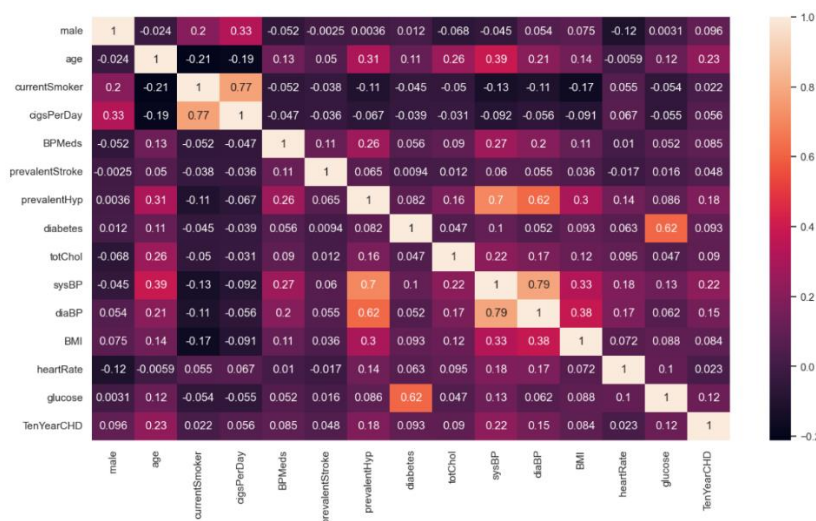


Fig. 20: Correlation Heatmap (Framingham Dataset)

The data set is skewed. Patients with hypertension and diabetes are at increased risk for coronary heart disease. There is little distinction between Smokers and non-Smokers.

According to the heatmap, no feature has a correlation with the target variable "TenYearCHD" of greater than 0.50.

There is a strong relationship between the following variables:

1. Blood glucose and Diabetes
2. Systolic BP and Diastolic BP
3. Smoking and Number of Cigarettes per day

All the aforementioned features and correlations are insufficient to construct a precise machine learning model for predicting the target variable "TenYearCHD."

Consequently, it is essential to conduct a feature selection in order to regulate the optimal features for constructing the prediction model.

Feature Selection

In machine learning, the objective of feature selection technique is to find the best group of features that can be used to build the best models for prediction. The following are the various types of feature selection techniques:

- Filter methods
- Wrapper methods
- Embedded methods
- Hybrid methods
- Dimensionality Reduction methods

In this study of the Framingham dataset, the Boruta algorithm was used to select features. The Boruta algorithm is a framework around the Random Forest algorithm. Boruta employs a technique known as "all-relevant feature selection," which means it takes into consideration all features that are occasionally significant to the result variable. This algorithm identifies all features that are either extremely essential or relatively important to the selection variable. This makes it an excellent choice for biomedical applications in which a person wishes to determine which human genes (features) are associated with a particular medical condition (target variable).

The following is the working of Boruta algorithm:

- The technique adds randomness to the supplied data set by making shadow features, which are copies of all the features that have been shuffled.
- Using the augmented data set, it trains a random forest classifier and assigns a feature importance measure (Mean Decrease Accuracy by default) to rank the significance of each feature, with a higher value indicating greater significance.
- At each iteration, it compares the feature's Z score to the maximum Z score of its shadow features to see if the true feature is more essential than the best of the shadow features, and then discards the ones that aren't.
- When all features have been validated or rejected, or when the maximum number of random forest iterations has been reached, the method terminates.

On using the Boruta algorithm on the Framingham dataset, the following 3 features were listed as the best ones for better prediction using the dataset.

- age
- sysBP
- BMI

But as part of the analysis, the top 7 features have been considered to build the prediction algorithm. The following are the top 7 features considered.

- age
- sysBP
- BMI
- diaBP
- totChol
- heartRate
- glucose

The dataset will be made more representative of the population by employing the Synthetic Minority Oversampling Technique (SMOTE).

Balancing the dataset

The term "imbalanced data" is used to describe data sets in which the number of observations for the target class is not distributed evenly between the two possible class labels. There is a great deal of bias in the Framingham dataset. There are at least six negative cases for every positive one. As a result, our Classifier model will provide mostly negative predictions.

There are many techniques to deal with imbalanced dataset. A few are listed below:

- Choosing proper evaluation metric
- Resampling (Oversampling or Undersampling)
- Synthetic Minority Oversampling Technique (SMOTE)
- Balanced Bagging Classifier
- Threshold moving

The Framingham dataset will be made more representative of the population by employing the Synthetic Minority Oversampling Technique (SMOTE).

SMOTE is a technique of oversampling the underrepresented group. This method does not merely add duplicate records of minority class as it does not contribute any novel data to the model. New instances are synthesised in SMOTE by mining the existing database. To generate a random synthetic instance in feature space, SMOTE first investigates instances from the minority class, then utilises nearest neighbour to pick a random nearest neighbour.

Below is a plot (Figure:17) showing the total count of the target parameter "TenYearCHD" before and after applying the SMOTE approach to achieve data balance. The count of the target parameter "TenYearCHD" is as below:

- Before SMOTE: 0 – 3179, 1 – 572 (Total – 3751)
- After SMOTE: 0 – 3178, 1 – 2543 (Total – 5721)

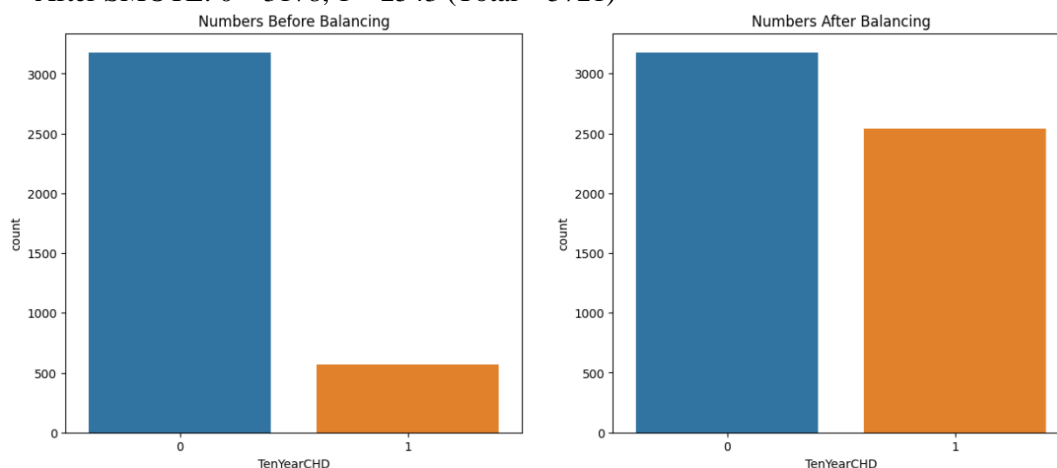


Fig. 21: Count of "TenYearCHD" before and after SMOTE data balancing

7. ALGORITHM IMPLEMENTATION :

As part of this research, the following machine learning algorithms were used to analyse both datasets.

- Logistic Regression
- Decision Tree
- Random Forest
- Support Vector Machine
- K Nearest Neighbours
- Adaptive Boost
- Gradient Boost
- Extreme Gradient Boost
- Light Gradient Boost

The details of the software used in the analysis are as follows:

- Python 3.9
- Pandas 1.5
- Jupyter Notebook
- Scikit-learn

- Seaborn
- Numpy 1.23
- Matplotlib 3.6

Training and Testing Dataset

To determine how efficiently a machine learning model can predict the outcome of unknown data, cross-validation is an evaluation approach used in machine learning. It is a widespread option because it's simple to grasp, can handle a little data set, and provides a fair assessment.

K-Fold Cross Validation is the most effective cross-validation method in machine learning since it is intuitive and straightforward to see and study. This is straightforward, using only a standard resampling method and no actual data substitution.

Each set (fold) would undergo training and testing exactly once throughout the entire procedure. Training a model with all obtainable data at once might lead to overfitting and affect accuracy and performance. Using k-fold cross-validation, a more robust, generic model can be constructed.

This K-Fold cross-validation technique has been used to in this research to split the datasets into training and testing datasets. The following Figure 18 represents the working architecture of the K-Fold cross-validation.

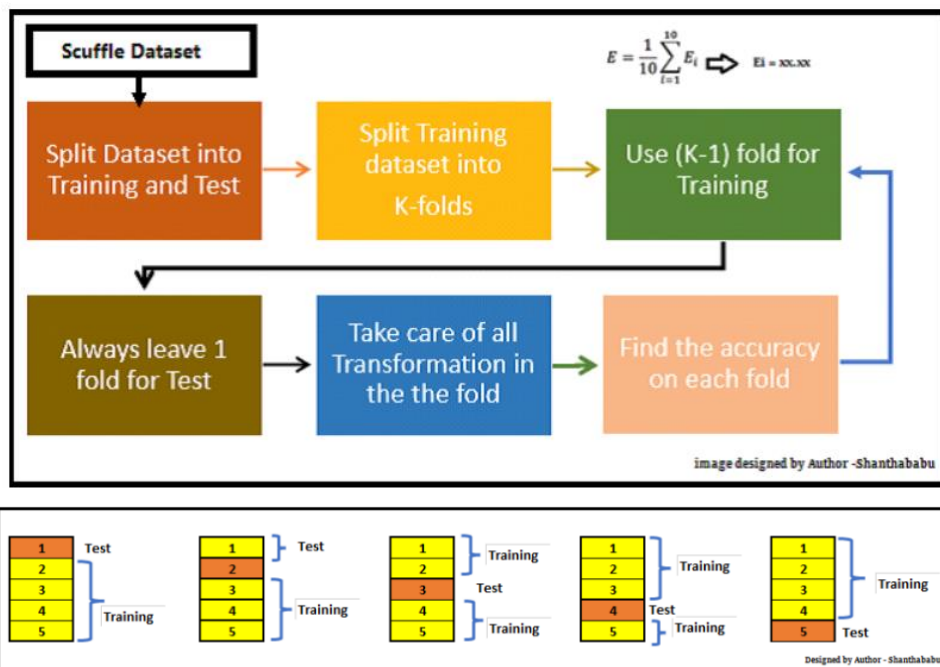


Fig. 22: Train and Test dataset using K-Fold Cross validation [26]

Hyperparameter tuning of the models

Every machine learning model selection is a significant undertaking, and it is entirely dependent on choosing the appropriate set of hyperparameters, which are required to train a model. It always refers to the parameters of the chosen model, which cannot be learned from the data, and must be provided before the model enters the training phase. Ultimately, the efficacy of the machine learning model enhances with a more acceptable selection of hyperparameter tuning techniques.

In a straightforward linear regression model, "Train-Test Split Ratio (80-20)" is an example of a hyperparameter.

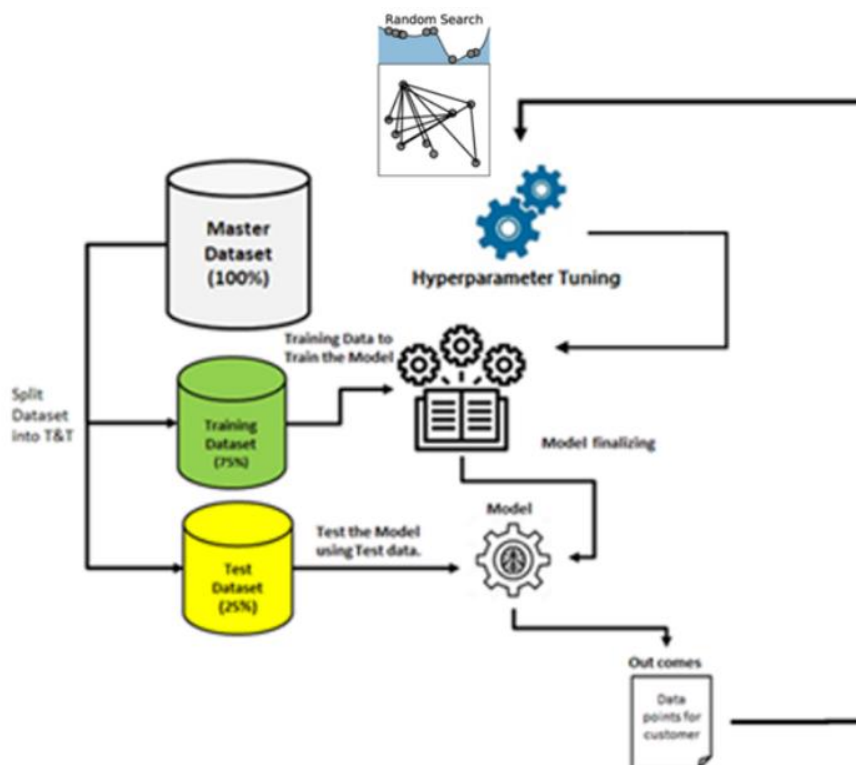


Fig. 23: Role of Hyperparameters in improving the performance of a model [27]

There is a set of hyperparameters for any given algorithm, and it is essential to regulate the optimal combination of hyperparameters and strategically adjust them to achieve the best results. This method will provide a foundation for Hyperparameter Space, and this combination will yield the most optimal outcomes. Finding this combination is difficult. The entire space of hyperparameters must be explored. Here, each combination of the chosen hyperparameter values is mentioned to as the "MODEL" and is evaluated immediately. GridSearchCV and RandomSearchCV are two generic methods for efficiently searching in the hyperparameter space. Here CV represents Cross-Validation.

Hyperparameters influence the following factors in the models:

- Linear Model – the degree of polynomial features to be used
- Decision Tree – Maximum depth allowed
- Random Forest – the minimum number of samples required at a leaf node
- Boosting Model – Learning Rate, number of layers
- K-Near Neighbour – Number of the neighbours to be used

As part of the research, for each of the nine models used for the Cleveland dataset and the Framingham dataset, specific hyperparameters were trained using the training set and evaluated using the testing set for specific performance parameters. The analysis results are discussed in the subsequent section.

8. RESULTS AND DISCUSSION :

The efficacy of each model was evaluated based on its Accuracy, f1 score, precision, and receiver – operator characteristics curve area under the curve (AUC). The Tables 6 and 7 depict the values achieved using the various algorithms during the study on the Cleveland dataset and Framingham dataset. The entire simulation study is conducted on HP Laptop with following Hardware and Software Specifications:

Hardware Specification:

Device name: LAPTOP-K9QUO1GT
 Processor: 11th Gen Intel(R) Core (TM) i3-1115G4 @ 3.00GHz
 Installed RAM: 8.00 GB (7.65 GB usable)
 System type: 64-bit operating system, x64-based processor

Software Specification:

Windows Version: Windows 10 Home Single Language

OS build: 19045.3086
Experience Windows Feature Experience Pack 1000.19041.1000.0

Table 6: Performance values of the Cleveland Dataset

Model	Accuracy	f1 score	Precision	AUC
Logistic Regression	0.81	0.81	0.76	0.872
Decision Tree	0.98	0.98	1.0	1.0
Random Forest	0.97	0.97	0.94	0.995
AdaBoost	0.96	0.96	0.94	0.969
Gradient Boost	0.97	0.97	0.94	1.0
Extreme Gradient Boost	0.94	0.94	0.94	0.997
Light Gradient Boost	0.97	0.97	0.94	1.0
K – Nearest Neighbors	0.97	0.97	0.94	1.0
Support Vector Machine	0.97	0.97	0.94	0.977

Table 7: Performance values of the Framingham Dataset

Model	Accuracy	f1 score	Precision	AUC
Logistic Regression	0.66	0.57	0.65	0.705
Decision Tree	0.75	0.75	0.76	0.779
Random Forest	0.88	0.86	0.88	0.945
AdaBoost	0.79	0.76	0.79	0.845
Gradient Boost	0.88	0.86	0.88	0.952
Extreme Gradient Boost	0.88	0.86	0.88	0.945
Light Gradient Boost	0.88	0.86	0.88	0.939
K – Nearest Neighbors	0.84	0.82	0.84	0.878
Support Vector Machine	0.87	0.85	0.87	0.926

The tables above show each algorithm's performance parameters for both datasets. From the numbers, we can figure out the following:

Most of the algorithms other than Logistic Regression, when applied on the Cleveland dataset have performed well and yielded good accuracy and AUC values.

- Decision Tree, Gradient Boost, Light Gradient Boost and K-Nearest Neighbours have achieved extraordinarily well with an AUC value of 1.
- The highest accuracy achieved is 0.98 by the Decision Tree model.
- The lowest accuracy achieved is 0.81 by the Logistic Regression model.
- With respect to Framingham dataset, the highest accuracy value is 0.88 and is achieved by Random Forest, Gradient Boost, Extreme Gradient Boost and Light Gradient Boost.
- The Logistic Regression model gave the lowest accuracy of 0.66.
- The Decision Tree model has achieved fairly well with an accuracy of 0.75.
- Though 4 models have achieved an accuracy of 0.88, the Gradient Boost model has outperformed all other models with an AUC of 0.952.

Hence it is inferred that Decision Tree model has given the best performance with respect to the Cleveland dataset and Gradient Boosting model has achieved the best with respect to the Framingham dataset. The following are the confusion matrix plot and AUC curve plot for these models.

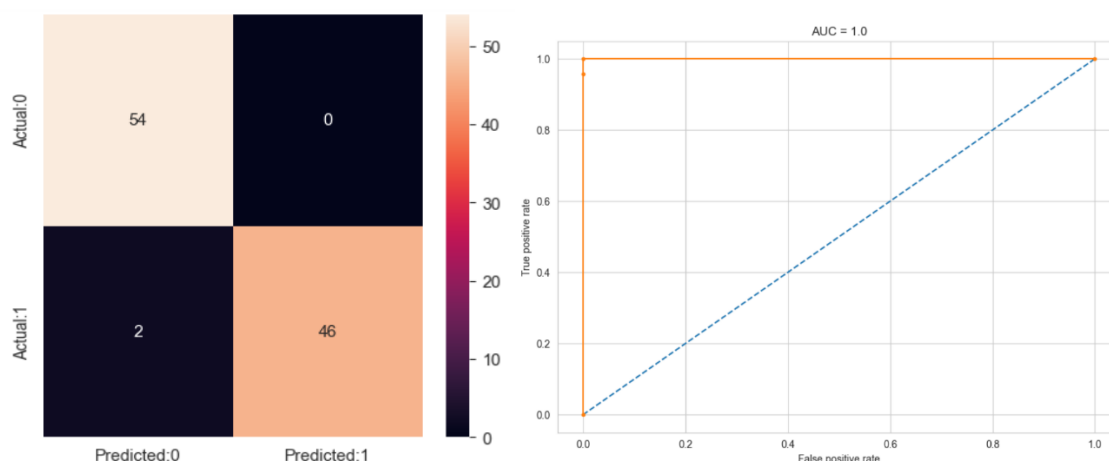


Fig. 24: Confusion Matrix and AUC curve for Cleveland Dataset (Decision Tree)

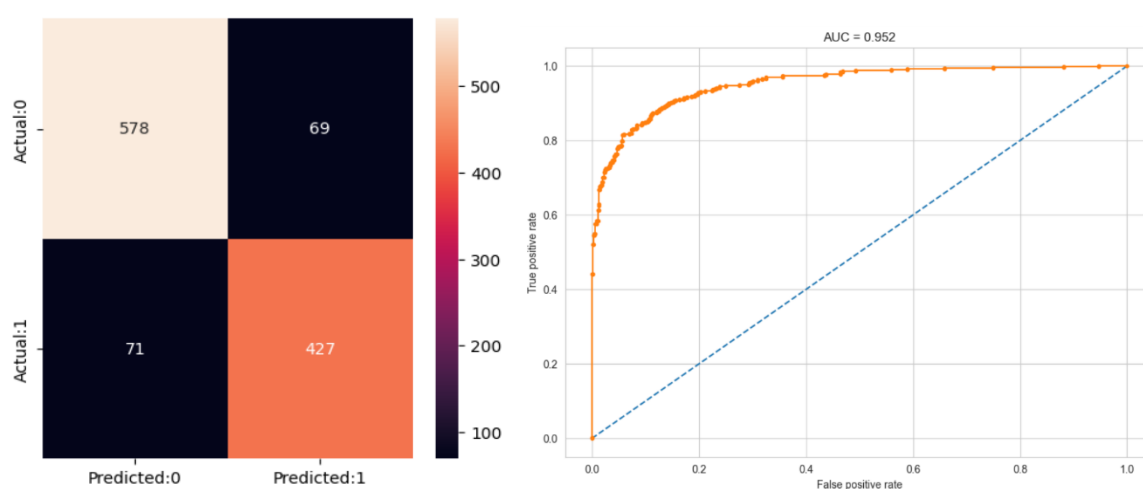


Fig. 25: Confusion Matrix and AUC Curve for Framingham dataset (Gradient Boost)

In the confusion matrix, the Actual value is the value of the target variable as in the dataset. The Predicted value is the value obtained for the same patient when the values are given to the trained model. While using the Decision Tree algorithm on the Cleveland dataset, only 2 patients who actually have the disease have been predicted to have no disease. Rest of the patients have been predicted accurately. While using Gradient Boost algorithm on the Framingham dataset, only 140 patients out of 1145 patients have been wrongly predicted. These details have been represented in the confusion matrix in the Figure 24 and Figure 25.

9. CONCLUSION :

In the course of this research, several different machine-learning strategies were applied to the analysis of two distinct datasets (the Cleveland Dataset and the Framingham Dataset), both of which were downloaded from Kaggle. Because each dataset had a unique combination of parameters, the pre-processing steps needed to be tailored specifically to that dataset. We were able to achieve different outcomes for each dataset by applying the various machine learning methods to the training dataset and then comparing the trained models using the testing dataset. The accuracy and area below the curve (AUC) measurements were used to assess the performance of the models. Although boosting algorithms performed well for both datasets, the Cleveland dataset benefited more from the application of decision trees in its analysis. Deep learning techniques need to be used to advance the research, and neural network models need to be constructed in command to create accurate prediction models. This can be helpful in determining whether or not the performance and correctness of the predictions can be upgraded further.

REFERENCE :

- [1] Baillargeon, B., Rebelo, N., Fox, D. D., Taylor, R. L., & Kuhl, E. (2014). The living heart project: a robust and integrative simulator for human heart function. *European Journal of Mechanics-A/Solids*, 48(1), 38-47. [Google Scholar](#)
- [2] Miao, K. H., & Miao, J. H. (2018). Coronary heart disease diagnosis using deep neural networks. *international journal of advanced computer science and applications*, 9(10), 1-9. [Google Scholar](#)
- [3] Coronary Artery Disease (CAD) https://www.cdc.gov/heartdisease/coronary_ad.htm. Retrieved on 04/03/2023.
- [4] Coronary Artery Disease <https://stanfordhealthcare.org/medical-conditions/blood-heart-circulation/coronary-artery-disease.html>. Retrieved on 04/03/2023.
- [5] Obstructive Coronary Artery Disease <https://stanfordhealthcare.org/medical-conditions/blood-heart-circulation/obstructive-coronary-artery-disease.html>. Retrieved on 04/03/2023.
- [6] Non-obstructive Coronary Artery Disease <https://stanfordhealthcare.org/medical-conditions/blood-heart-circulation/non-obstructive-coronary-artery-disease.html>. Retrieved on 04/03/2023.
- [7] Spontaneous Coronary Artery Dissection (SCAD) <https://stanfordhealthcare.org/medical-conditions/blood-heart-circulation/spontaneous-coronary-artery-dissection.html>. Retrieved on 04/03/2023.
- [8] What is Coronary Artery Disease? <https://www.healthline.com/health/coronary-artery-disease>. Retrieved on 04/03/2023.
- [9] Sindayigaya, L., & Dey, A. (2022) Machine Learning Algorithms: A Review. *Information Systems Journal*, 11(8), 1127-1133. [Google Scholar](#)
- [10] Dogan, M. V., Grumbach, I. M., Michaelson, J. J., & Philibert, R. A. (2018). Integrated genetic and epigenetic prediction of coronary heart disease in the Framingham Heart Study. *PLoS one*, 13(1), e0190549. [Google Scholar](#)
- [11] Swathy, M., & Saruladha, K. (2022). A comparative study of classification and prediction of Cardio-Vascular Diseases (CVD) using Machine Learning and Deep Learning techniques. *ICT Express*, 8(1), 109-116. [Google Scholar](#)
- [12] Li, J. P., Haq, A. U., Din, S. U., Khan, J., Khan, A., & Saboor, A. (2020). Heart disease identification method using machine learning classification in e-healthcare. *IEEE Access*, 8(1), 107562-107582. [Google Scholar](#)
- [13] Watson, D. S., Krutzinna, J., Bruce, I. N., Griffiths, C. E., McInnes, I. B., Barnes, M. R., & Floridi, L. (2019). Clinical applications of machine learning algorithms: beyond the black box. *Bmj*, 364(1), 1-5. [Google Scholar](#)
- [14] Pal, M., & Parija, S. (2021, March). Prediction of heart diseases using random forest. In *Journal of Physics: Conference Series*, 1817(1), 1-9. IOP Publishing. [Google Scholar](#)
- [15] Cherian, R. P., Thomas, N., & Venkitachalam, S. (2020). Weight optimized neural network for heart disease prediction using hybrid lion plus particle swarm algorithm. *Journal of Biomedical Informatics*, 110(1), 1-11. [Google Scholar](#)
- [16] Petch, J., Di, S., & Nelson, W. (2022). Opening the black box: the promise and limitations of explainable machine learning in cardiology. *Canadian Journal of Cardiology*, 38(2), 204-213. [Google Scholar](#)
- [17] Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, 17(1), 1-9. [Google Scholar](#)

- [18] Johnson, K. W., Torres Soto, J., Glicksberg, B. S., Shameer, K., Miotto, R., Ali, M., ... & Dudley, J. T. (2018). Artificial intelligence in cardiology. *Journal of the American College of Cardiology*, 71(23), 2668-2679. [Google Scholar](#)
- [19] Harini, C., & Anu, V. M. (2021). Clinical Decision Support Systems Using Sequential Pattern Mining Algorithms for Cardio Vascular Diseases. *REVISTA GEINTEC-GESTAO INOVACAO E TECNOLOGIAS*, 11(3), 756-770. [Google Scholar](#)
- [20] Priya, A., Garg, S., & Tigga, N. P. (2020). Predicting anxiety, depression and stress in modern life using machine learning algorithms. *Procedia Computer Science*, 167(1), 1258-1267. [Google Scholar](#)
- [21] Building an End-to-End Logistic Regression Model <https://www.analyticsvidhya.com/blog/2021/10/building-an-end-to-end-logistic-regression-model/>. Retrieved on 04/03/2023.
- [22] Decision Trees <https://www.ibm.com/topics/decision-trees>. Retrieved on 04/03/2023.
- [23] Understand Random Forest Algorithms with Examples <https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/>. Retrieved on 04/03/2023.
- [24] Gradient Boosting – What You Need to Know <https://datascience.eu/machine-learning/gradient-boosting-what-you-need-to-know/>. Retrieved on 04/03/2023.
- [25] Dealing with outliers using the Z-Score method <https://www.analyticsvidhya.com/blog/2022/08/dealing-with-outliers-using-the-z-score-method/>. Retrieved on 04/03/2023.
- [26] K-Fold Cross Validation and its Technique and its Essentials <https://www.analyticsvidhya.com/blog/2022/02/k-fold-cross-validation-technique-and-its-essentials/>. Retrieved on 04/03/2023.
- [27] A Comprehensive Guide on Hyperparameter Tuning and its Techniques <https://www.analyticsvidhya.com/blog/2022/02/a-comprehensive-guide-on-hyperparameter-tuning-and-its-techniques/>. Retrieved on 04/03/2023.
- [28] XGBoost (Extreme Gradient Boosting) in Machine Learning <https://medium.com/@jwbtfm/xgboost-extreme-gradient-boosting-in-machine-learning-3427b937b35c>. Retrieved on 04/03/2023.
- [29] GBM in Machine Learning <https://www.javatpoint.com/gbm-in-machine-learning>. Retrieved on 04/03/2023.
- [30] K-Nearest Neighbour (KNN) Algorithm for Machine Learning <https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning>. Retrieved on 04/03/2023.
- [31] Support Vector Machine Algorithm <https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm>. Retrieved on 04/03/2023.
- [32] Heart Disease Cleveland Dataset <https://www.kaggle.com/datasets/ritwikb3/heart-disease-cleveland>. Retrieved on 04/03/2023.
- [33] Heart Disease Framingham Dataset <https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset>. Retrieved on 04/03/2023.
