Effective Heart Disease Prediction Using Machine Learning and Deep Learning Techniques

|  |
| --- |
|  |

**Abstract:** Cardiovascular diseases have been a leading cause of mortality in the present era; the prognosis of cardiovascular diseases is still tricky for physicians. Fortunately, there are sequences and patterns present in cardiovascular diseases. Machine Learning and Deep learning techniques can identify this pattern and provide valid predictions for prognosis. This research may help in predicting cardiovascular diseases and reduce misdiagnosis. Initial data preprocessing included handling outliers through Z-scores, followed by classifying the data through clustering using the K-modes algorithm with Huang to increase the accuracy and reliability of the model. Machine learning models include Catboost Classifier, Gradient Boosting Classifier, K-Nearest Neighbor Classifier (KNN), Light Gradient Boosting Classifier, and Support Vector Machine. We have also used Recurrent Neural Networks (RNN) and Long Short-Term Model (LSTM). GridSearchCV was used to hypertune the model's parameters to increase its accuracy and reliability. After preprocessing the data, we apply the model to 61,158 dataset instances with 12 attributes. We train the proposed models on the data, splitting it into 60:40, 70:30, and 80:20 ratios, achieving the following accuracies: CatBoost: 88.11(without cross-validation) and 87.77(with cross-validation), Gradient Boosting: 87.21(without cross-validation) and 88.01(with cross-validation), KNN: 86.88(without cross-validation) and 87.33(with cross-validation), Light Gradient Boosting: 88.09(without cross-validation) and 88.01(with cross-validation), SVM: 87.00(without cross-validation) and 88.09(with cross-validation), for Recurrent Neural Networks(RNN): 88.01 and Long Short Term Model(LSTM): 88.13. From the following accuracies, we can conclude that Long short-term model and CatBoost without cross-validation have the highest accuracy respectively at 88.13% and 88.11%.

**Keywords:** Cardiovascular diseases; Machine Learning; Deep Learning; z-scores; k-modes; cross-validation; parameter hypertuning

1. Introduction

Cardiovascular diseases (CVDs) are the leading cause of morbidity and mortality globally, accounting for over 70% of all deaths. According to the 2017 Global Burden of Disease study, cardiovascular diseases account for approximately 43% of all deaths [[1,2]](#_References). In high-income countries, common risk factors for heart disease include poor diet, smoking, excessive sugar intake, and obesity [[3,4]](#_References). However, chronic illnesses are also becoming more prevalent in low- and middle-income countries [[5]](#_References). The global economic burden of cardiovascular diseases was projected to reach approximately USD 3.7 trillion between 2010 and 2015 [[6,7]](#_References).

Additionally, diagnostic technology like electrocardiograms and CT scans for coronary heart sickness can be too high-priced and impractical for a few patients. This problem on my own has contributed to the demise of 17 million human beings [[5]](#_References). Employees with cardiovascular sickness account for 25 to 30% of the firm's annual scientific costs [[8]](#_References). Early detection of heart disease is essential to reduce its bodily and monetary toll on people and agencies. The WHO predicts that by way of 2030, the variety of deaths from cardiovascular sicknesses will upward thrust to 23.6 million, with coronary heart sickness and stroke being the pinnacle causes [[9]](#_References). Using statistics mining strategies, we are able to discover hidden patterns that useful resource in scientific diagnosis [[10]](#_References). Many elements, together with diabetes, high blood strain, excessive cholesterol, and peculiar pulse charge, want to be taken into consideration when predicting heart ailment [[11]](#_References). Often, the scientific data available need to be finished, affecting the results in predicting heart sickness.

Machine learning also plays a vital role in the medical field. We can diagnose and predict various diseases with machine learning. Recently, there has been a growing interest in using data mining and machine learning techniques to predict the likelihood of developing specific diseases. Existing work includes applications of data mining techniques for disease prediction. Although some studies have tried to predict the future risk of disease progression, they have yet to achieve accurate results [[12]](#_References).

This research focuses on accurately estimating the risk of heart disease. Our study compares machine and deep learning methods for predicting cardiovascular illnesses. To build predictive models, we used techniques such as Gradient Boosting, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), CatBoost, and LightGBM, as well as deep learning algorithms such as Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks.

We used k-mode clustering to preprocess and scale the publicly available dataset on Kaggle to boost model convergence. We used Python on Visual Studio Code to perform all computations, preprocessing, and visualizations.

Previous research has demonstrated accuracy rates of up to 94% in predicting heart disease using machine-learning approaches. However, these researchers frequently used tiny sample sizes, which may restrict the generalizability of their findings. Our work tackles this problem by utilizing a more extensive and diverse dataset, increasing our findings' robustness and application.

2. Literature Survey

The healthcare industry has recently advanced in data mining and machine learning technologies. These procedures have been widely implemented and helpful in various healthcare settings, most notably in medical cardiology. The growing accumulation of medical data has created unprecedented academic potential to build and test novel algorithms. Heart disease is still the primary cause of death in developing countries [[12,13]](#_References); thus, identifying risk factors and early warning signals is an important study topic. Data mining and machine learning approaches can help in the early detection and prevention of heart disease.

Narain et al. [[14]](#_References) aimed to develop a novel machine-learning-based cardiovascular disease (CVD) prediction system to improve the accuracy of the widely used Framingham risk score (FRS). The suggested system, which recognizes CVD patterns using a quantum neural network, was validated using data from 689 individuals with CVD symptoms and a validation dataset from the Framingham study. The system's accuracy in forecasting CVD risk was 98.57%, significantly higher than the FRS's 19.22% accuracy and other existing techniques. The study suggests that this approach could help doctors predict CVD risk, create better treatment plans, and facilitate early diagnosis.

Shah et al. [[15]](#_References) developed a model that predicts cardiovascular diseases using machine learning techniques. They used the Cleveland heart disease dataset, which has 303 instances and 17 attributes from the UCI machine learning repository. The author employed various supervised classification methods, such as naive Bayes, decision trees, random forests, and k-nearest neighbor (KKN). It was the KKN model where accuracy reached the highest of all at 90.8%. This research shows potential for using these models to predict heart problems, but only if appropriate models and techniques are selected for optimal results.

Drod et al. [[2]](#_References) aimed to identify the critical risk factors for cardiovascular disease (CVD) in individuals with metabolically associated fatty liver disease (MAFLD) using machine learning (ML). The researchers assessed subclinical atherosclerosis in 191 MAFLD patients and carried out blood chemistry tests on them. Then, they developed an algorithm that could enable them to identify high-risk CVD cases among MAFLD cases using methods like Principal Component Analysis, Univariate Feature Ranking, and Multiple Logistic Regression Classifier (PCA). Diabetes duration, plaque scores, and hypercholesterolemia were found to be critical clinical attributes by the study. The machine learning approach's AUC was 0.87; it correctly identified 40/47(85.11%) high-risk patients and 114/144(79.17%) low-risk group. The results indicate that ML methods are useful for identifying MAFLD cases accompanied by extensive CVD based on simple patient criteria, which can be employed straightforwardly.

The aim of [[16]](#_References) was to look at the usefulness of machine learning (ML) techniques in predicting heart failure disease. This study used data from the Cleveland Clinic Foundation, and various ML algorithms such as decision trees, logistic regression, random forest, naive Bayes, and support vector machine (SVM) were used to develop prediction models. Model development involved a 10-fold cross-validation approach. The highest level of accuracy in predicting heart disease fell under the Decision Tree Algorithm, with a score of 93.19%, followed by the SVM algorithm, which scored 92.30%. This study shows how ML techniques can effectively predict heart failure disease with specific reference to future research options. The decision tree algorithm is a good option for further studies on heart disease.

Hasan and Bao [[17]](#_References) conducted a study to identify the most efficient feature selection approach for predicting cardiovascular illness by comparing multiple algorithms. They initially considered three well-known feature selection methods: filter, wrapper, and embedding. A feature subset was then recovered from these methods using a Boolean process-based common "True" condition involving two stages of subset retrieval. They evaluated various models to determine the best predictive analytics, including random forest, support vector classifier (SVC), k-nearest neighbors, naive Bayes, and XGBoost. The artificial neural network (ANN) was used as a standard for comparison with all features. The findings demonstrated that the most accurate prediction results for cardiovascular illness were achieved by the XGBoost classifier coupled with the wrapper technique, which delivered an accuracy of 73.74%, followed by SVC with 73.18% and ANN with 73.20%.

Our and Elsedawy [[18]](#_References) attained an accurate percentage of 89.01 using the Random Forest algorithm on UCI's cardiovascular database. The Random Forest method was applied by Elsedawy’s work because it was robust and could accommodate huge amounts of data with high dimensionality. This goes on to demonstrate how strong ensemble techniques can be when predicting heart diseases. Nevertheless, we believe that our study’s use of cross-validation contributes to an optimal evaluation in terms of model effectiveness. It helps guarantee that results are not over specified for a unique dataset thereby making our outcomes more universal to other ethnic groups and adding real-life credibility.

Khan and Mondal [[19]](#_References) have touched upon different approaches such as neural networks and logistic regression whose maximum accuracy was about 72.72%. Their findings reaffirmed effectiveness of common procedures but they also pointed out to their deficiencies compared with current machine learning approaches. Classifiers like neural networks and logistic regression are quite common in classifying instances yet encounter problems like intense complexity and dearth of dimensions associated with health data.

Bhatt et al. [[20]](#_References) focused on improving classification accuracy using k-mode clustering with Huang starting. The study employed models such as random forest, decision tree classifier, multilayer perceptron, and XGBoost on a real-world dataset of 70,000 instances from Kaggle. The models achieved the following accuracies: decision tree: 86.37% (with cross-validation) and 86.53% (without cross-validation), XGBoost: 86.87% (with cross-validation) and 87.02% (without cross-validation), random forest: 87.05% (with cross-validation) and 86.92% (without cross-validation), multilayer perceptron: 87.28% (with cross-validation) and 86.94% (without cross-validation). The AUC values for these models were 0.94 (decision tree), 0.95 (XGBoost), 0.95 (random forest), and 0.95 (multilayer perceptron).

However, some limitations have persisted despite the promising outcomes of previous research papers on heart disease prediction using machine learning. Some researchers, including Narain et al. and Shah et al., employed small datasets, which may mean their findings cannot be generalized to larger populations with many different characteristics. These studies were also limited regarding the number of features considered; hence, they may have missed out on other essential risk factors for cardiovascular diseases. The practicality of these models is also restrained by computational complexity and resource demands that come with advanced approaches like Bhatt et al.’s. While some models, such as those by Hasan and Bao, showed a moderate accuracy level, they need further work done for feature selection and model tuning. Therefore, it is vital to continuously improve this area to build more reliable real-world-oriented CVD predictive models.21

Table 1 : Existing works on heart disease prediction on similar datasets

|  |  |  |  |
| --- | --- | --- | --- |
| **Authors** | **Novel Approach** | **Best Accuracy** | **Dataset** |
| Shorewall, 2021 [[5]](#_References) | KNN, random forest, and SVM outputs with logistic regression | 75.1% | Kaggle cardiovascular disease  dataset (70,000 patients,  12 attributes) |
|  |
|  |
| Maiga et al., 2019 [[7]](#_References) | Random Forest,  Naïve Bayes, Logistic Regression,  KNN | 70% | Kaggle cardiovascular disease  dataset (70,000 patients,  12 attributes) |  |
|  |
| Waigi at el., 2020 [[12]](#_References) | Decision Tree | 72.77% | Kaggle cardiovascular disease  dataset (70,000 patients,  12 attributes) |  |
|  |
| Our and ElSeddawy, 2021 [[18]](#_References) | Random forest | 89.01% | UCI cardiovascular  dataset  (303 patients, 14 attributes) |  |
|  |
| Khan and Mondal, 2020 [[19]](#_References) | Holdout cross-validation with the neural network for Kaggle dataset | 71.82% | Kaggle cardiovascular disease  dataset (70,000 patients,  12 attributes) |  |
| Cross-validation method with logistic regression (solver: lbfgs) where k = 30 | 72.72% | Kaggle cardiovascular disease  dataset 1 (462 patients,  12 attributes) |  |
| Cross-validation method with linear SVM where k = 10 | 72.22% | Kaggle cardiovascular disease dataset (70,000 patients, 12 attributes) |  |
| Bhatt et al, 2023 [[20]](#_References) | Improving classification accuracy using k-mode clustering with Huang starting, applied random forest, decision tree classifier, MLP, XGBOOST | 87.28% (multilayer perceptron, with cross-validation) | Kaggle cardiovascular disease dataset (70,000 patients, 12 attributes) |  |
|  |

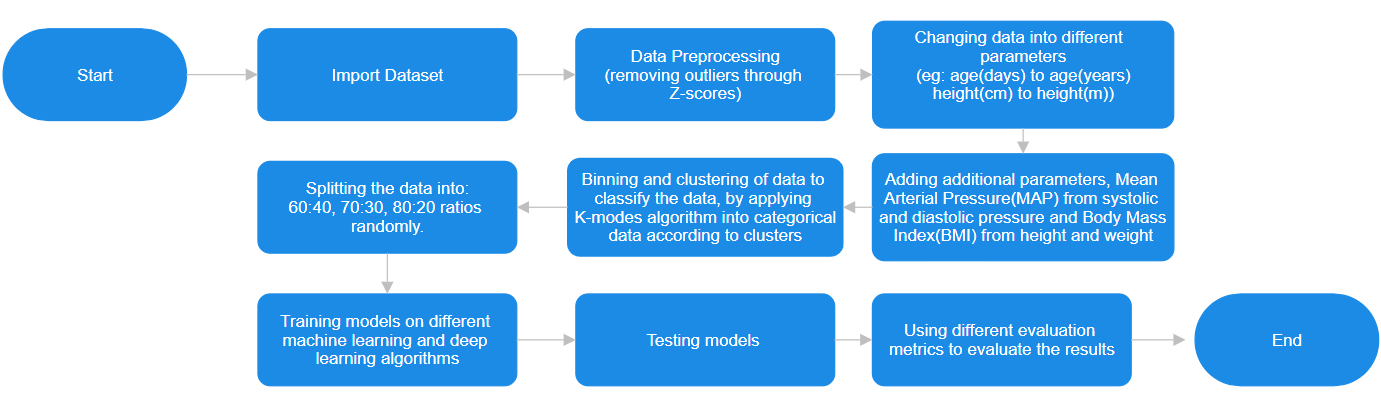
3. Proposed methodology

Figure 1 : Architecture of Proposed Methodology

It starts with importing the dataset. The initial operations include loading and preparing data and preprocessing, which includes outlier detection by calculating Z-scores. To reduce dimensionality, the input characteristics are scaled; for example, age is converted from days to years and height from centimeters to meters. After preprocessing, the data is put through the binning and the clustering process. Binning divides the dates into standardized ranges, while the other method used here, K-Modes, categorizes the data into groups that share similar features. This step improves the classification process because it seeks to establish regularities within the data. Processing of other parameters is defined: Mean Arterial Pressure (MAP), which shows average blood pressure value calculated from systolic and diastolic pressure; Body et al. (BMI) calculated from height and weight, which expands a set of indicators in the health field. The processed data is then split into a training set and a test set using the ratios (60:40), (70:30,) and (80:20) [[21]](#_References) to enable a proper evaluation of the model. Gradient Boosting, CatBoost, LightGBM, K Nearest Neighbors (KNN), Support Vector Machine (SVM), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) are then used on these datasets. After training the models, their ability to make predictions is checked. To assess the proposed models, basic evaluation measures like accuracy, precision, recall, F1 Score, and AUC-ROC are used so that the model's performance can be evaluated appropriately. Starting from preconditioning the data and ending with the model assessment, the described workflow forms the basis of the architecture of the heart disease prediction system that will help in clinical decision-making with high reliability and accuracy.

*3.1. Dataset Description:*

3.1.1 Data sources

As described in [[22]](#_References), the Dataset utilized in this study comprises 70,000 patient records with 12 distinct features from Kaggle, as listed in Table 2. This Dataset contains clinical and demographic information relevant to heart disease prediction. These features include age, gender, systolic blood pressure, and diastolic blood pressure. The target variable, “cardio,” indicates whether a patient has cardiovascular disease (represented as 1) or is healthy (represented as 0).

3.2.2 Data Features

This dataset includes the following features in table 2:

Table 2: features of the dataset

|  |  |  |
| --- | --- | --- |
| **feature** | **Variable** | **Minimum and Maximum Values** |
| Age | Age | Minimum: 10,798 and Maximum: 23,713 |
| Height | Height | Minimum: 55 and Maximum: 250 |
| Weight | Weight | Minimum: 10 and Maximum: 200 |
| Gender | Gender | 1: female,2: male |
| Systolic blood pressure | ap\_hi | Minimum: -150 and Maximum: 16,020 |
| Diastolic blood pressure | ap\_lo | Minimum: -70 and Maximum: 11,000 |
| Cholesterol | Chol | Categorical value: 1(low) to 3(high) |
| Glucose | Gluc | Categorical value: 1(low) to 3(high) |
| Smoking | Smoke | 1: yes, 0: no |
| Alcohol intake | Alco | 1: yes, 0: no |
| Physical activity | Active | 1: yes, 0: no |
| Presence or Absence of cardiovascular disease | Cardio | 1: yes, 0: no |

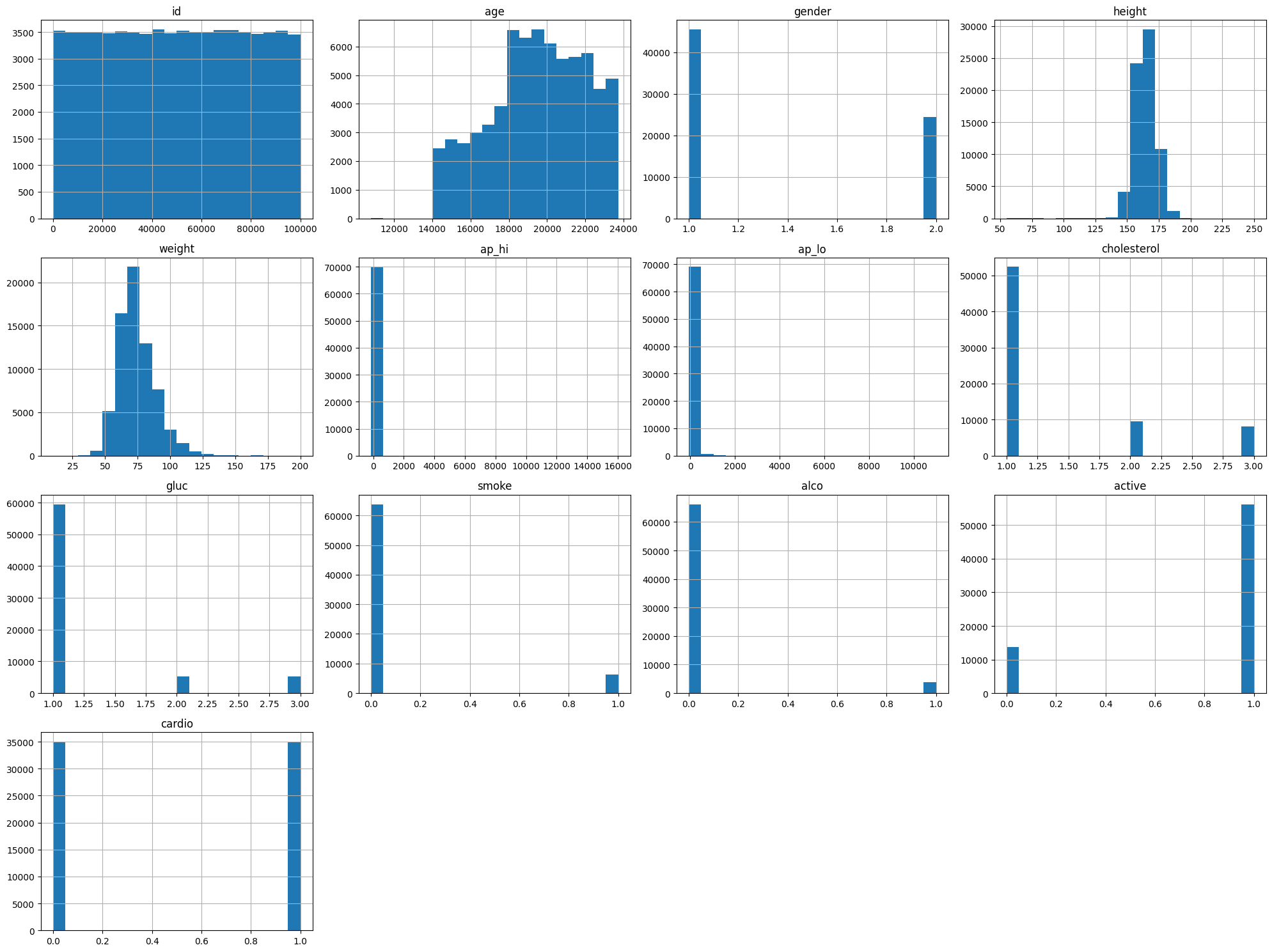


Figure 2 : Bargraph for features from the datset

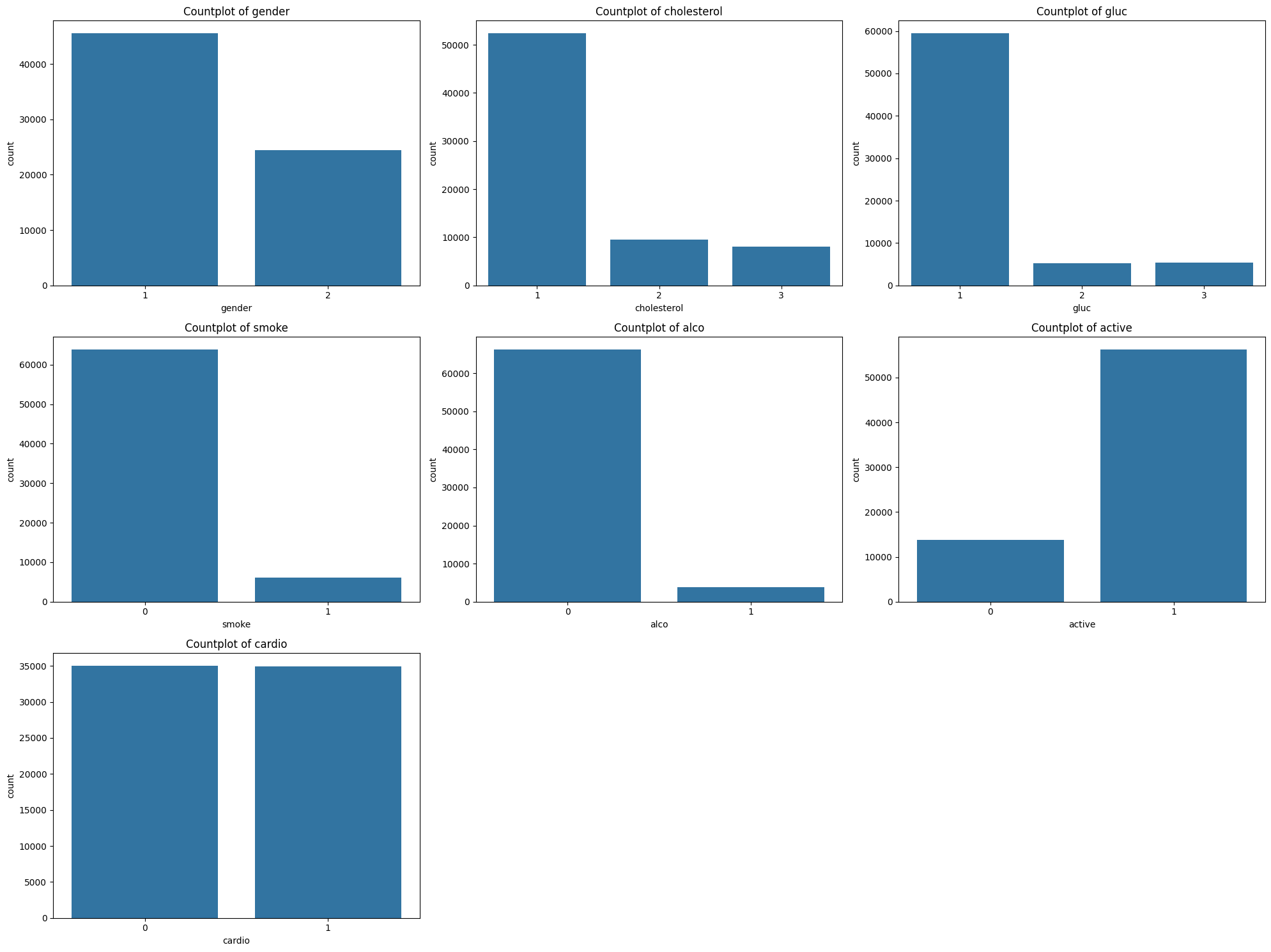
Figure 2 illustrates the histograms of key attributes, providing insights into their distributions before preprocessing. The histograms presented show the distribution of the various components of the cardiovascular health data set. The ID feature is uniformly distributed, representing unique identifiers of individuals. The age distribution is right-skewed, especially from 40 to 60 years, indicating that the dataset includes predominantly middle-aged adults. Gender is divided into 1:2 ratio roughly, which will give us an equal opportunity to analyse any gender distribution for the heart disease prediction . Height shows a normal distribution centered at 160 to 170 cm, while weight is right-skewed, especially at 60 and 80 kg, indicating a relatively healthy population in BMI with notable outliers at systolic (ap\_hi) and high diastolic values ​​(ap\_lo) blood pressure. A skewed distribution has been demonstrated, possibly reflecting issues with hypertension. Lipids and glucose levels are divided into three categories, with most individuals falling into the everyday and above typical categories, reflecting general health issues; smoking status (smoking), and drinking (also) are binary features and of the s For the majority, which can reflect health consciousness in the group Physical activity (be ing active) is also binary, meaning that most individuals are active, which can and heart is associated with good dental health. Finally, the cardio (target variable) objective variable is binary, reflecting the nearly equivalent status of heart diseased individuals and non diseased individuals, which is important for balanced training models This distribution provides information a detailed overview of the structure, which is important for proper data preprocessing, identification of potential sources of bias, and development of appropriate sampling methods

Figure 3: Count plot for the features of dataset

As shown in Figure 3, the count plot indicates the frequency of different categorical attributes within the dataset. The presented mathematical models provide a hierarchical classification analysis of several key components of the heart health data set. The gender plot shows a high proportion of one sex, indicating an imbalance that needs to be taken into account when conducting the analysis. The graph of cholesterol levels shows that most individuals have normal cholesterol levels (level 1), and few individuals are in high cholesterol groups (2 and 3), indicating that most people have cholesterol with healthy The Glucose plot also shows that most individuals have normal glucose levels. level (level 1), with decreased diabetes-prediabetic range (levels 2 and 3) Smoking status map (cigarettes) shows that a significant majority of people do not smoke, which may affect the assessment of cardiovascular risk factors Drinking alcohol plot (alco). ) also exhibit a similar trend, with most individuals not drinking, and revealing a potential health care cohort. An (active) physical activity program means that most individuals are physically active, which is fundamental to cardiovascular health. Finally, objective change plots for cardiovascular (heart disease) show a balanced distribution between diseased and disease-free individuals, which is important for effective training of predictive models These statistical plots provide valuable insights into the categorical variables of the dataset, reveal potential allocations and biases needed to be addressed in analysis

3.3.3 Data Preprocessing

Several preprocessing steps were done to ensure the dataset was not corrupted and was good enough. It started by removing outliers, which can seriously affect data analysis and model performance through the interquartile range (IQR) method or Z-scores. In this case, the IQR method involves getting the first quartile (Q1) and third quartile (Q3) for every feature with a difference between Q3 and Q1 is IQR. Outliers are those that fall below Q1-1.5IQR or above Q3+1.5IQR. This method was applied to eliminate extreme values for critical features such as height, weight, systolic blood pressure (ap\_hi), and diastolic blood pressure (ap\_lo). Moreover, we used Z-scores as it makes the data standardized by subtracting the mean and dividing by standard deviation within each feature after outliers were identified from points in the dataset that had z-scores more significant than three or less than -3. Thus, this process helped find any remaining extreme value situated in a feature with an approximately normal data distribution. This was followed by data normalization after outliers were removed to maintain consistency and enhance machine learning models' performance. Normalization is carried out by scaling the features to a standard range between 0 and 1 to ensure one feature does not dominate the learning process in scale. To avoid any unwanted noise in the dataset and ensure the data was suitable for machine learning models that predict heart disease, we normalized our data by discarding outliers.

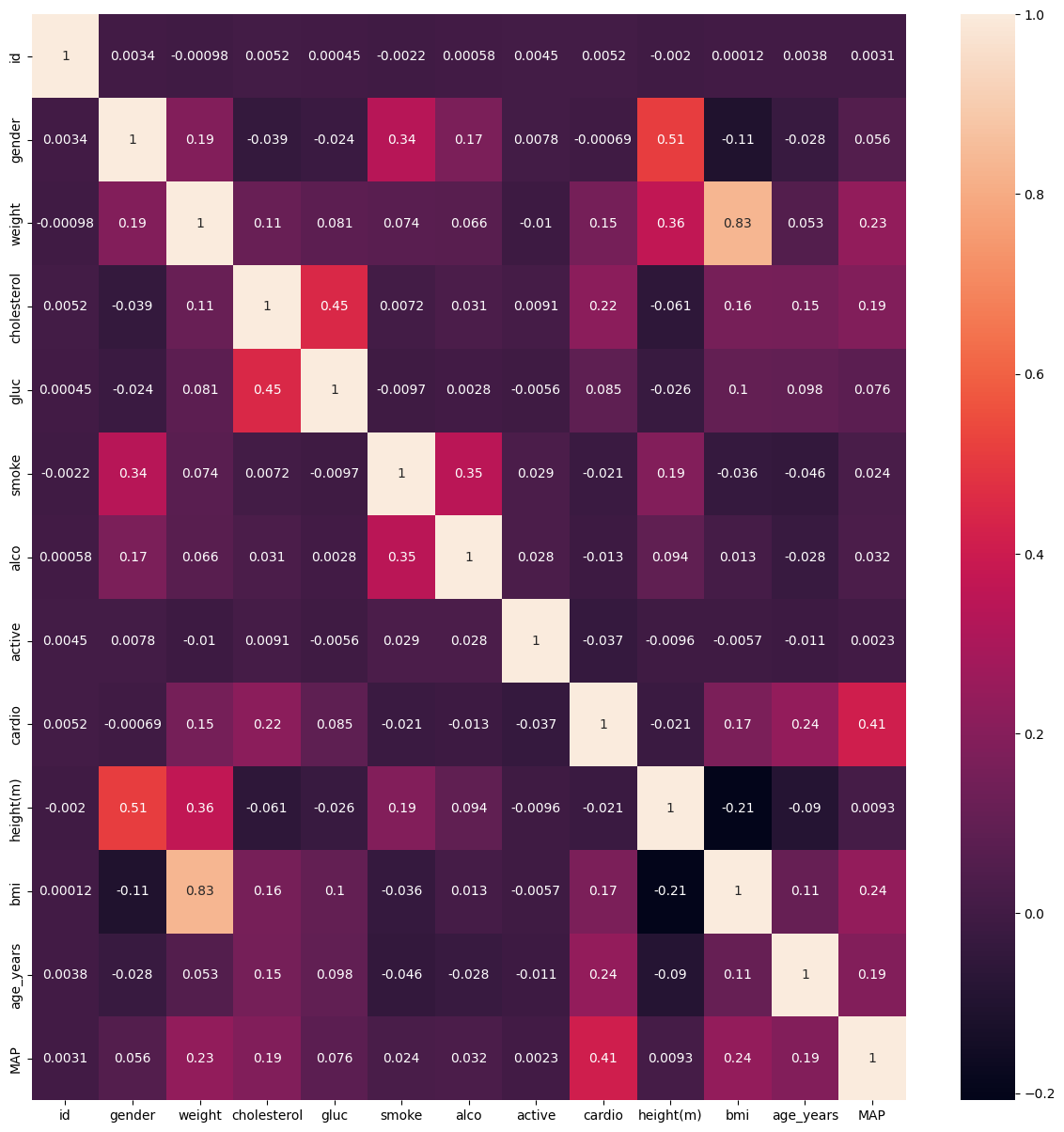
*3.2 Feature Selection*

We combined several methods to select relevant features for training our model. The selection process applied statistical techniques, recursive algorithms, and importance scores derived from ensemble methods.

3.2.1 Correlation Analysis

To start with feature selection, we must first analyze the correlation between features. We identified and removed highly correlated features by computing Pearson's correlation coefficient for each pair of numeric features. This is because a high correlation between features may result in multicollinearity, negatively affecting the performance of some machine learning models since these models tend to be unstable and hard to understand.

Figure 4: correlation matrix of features (after adding features and preprocessing)



3.2.2 Use of ensemble methods

Ensemble methods such as Gradient Boosting, CatBoost, and LightGBM were used for feature selection due to their strong performance in complex datasets. Gradient Boosting constructs models by sequentially correcting each other's errors one by one while optimizing loss function to discern essential features. Specifically designed for categorical data, CatBoost employs ordered boosting and target statistics, making it efficient on categorical variables, thus reducing overfitting and enhancing accuracy. Known for its efficiency and scalability, LightGBM uses leaf-wise tree growth and histogram-based algorithms that can handle large datasets [[23]](#_References). We calculate "feature importance" for every technique based on how much it reduces the model's loss function, resulting in a comprehensive evaluation of relevant attributes regarding heart disease prediction.

3.2.3 Additional Features

Body Mass Index (BMI):

To check if a person is healthy for their age in terms of body weight, clinicians often employ Body Mass Index (BMI). It can be calculated as follows:

Underweight (BMI < 18.5), average weight (18.5 ≤ BMI < 24.9), overweight (25 ≤ BMI < 29.9), and obesity (BMI ≥ 30) are the four categories of BMI. Obesity is a major determinant of cardiovascular diseases as it aids ineffectual deposition of fats in the body. Our motive for incorporating BMI among our features was to capture the possibility of heart problems due to issues associated with being overweight or obese, also referred to as "fatness." Hypertension, diabetes, and coronary artery disease are the commonest health issues related to elevated levels of BMI, thus making it an integral part of forecasting heart diseases.

Mean Arterial Pressure (MAP):

Mean Arterial Pressure (MAP) provides a single value representing the average blood pressure in an individual during a single cardiac cycle. It can be computed according to this formula:

Instead of systolic or diastolic pressure alone, MAP gives a broader perspective on the effect of blood pressure on heart health. When mean arterial pressure (MAP) values are high, it shows increased stress on the cardiovascular system, which may result in hypertension and heart disease, among other conditions. Including MAP in the feature set would add clinical insights for better accuracy and robustness of our models for predicting heart diseases [[25]](#_References). BMI and MAP, which are derived features, have improved dataset’s ability to capture core health indicators, thus leading to more efficient forecasting and analysis. After removing outliers and normalizing the data and adding new features, the revised ranges of attributes are presented in Table 3.

Table 3: features of the dataset (after adding features and preprocessing)

|  |  |  |
| --- | --- | --- |
| **feature** | **Variable** | **Minimum and Maximum Values** |
| Age(years) | Age\_years | Minimum: 29.75 and Maximum: 64.96 |
| Height(meters) | Height(m) | Minimum: 1.43 and Maximum: 1.86 |
| Weight (Kilograms) | Weight | Minimum: 41.00 and Maximum: 105.00 |
| Gender | Gender | 1: female,2: male |
| Mean Arterial Pressure | MAP | Minimum: 73.33 and Maximum: 126.00 |
| Body Mass Index | bmi | Minimum: 15.09 and Maximum: 38.56 |
| Cholesterol | Chol | Categorical value: 1(low) to 3(high) |
| Glucose | Gluc | Categorical value: 1(low) to 3(high) |
| Smoking | Smoke | 1: yes, 0: no |
| Alcohol intake | Alco | 1: yes, 0: no |
| Physical activity | Active | 1: yes, 0: no |
| Presence or Absence of cardiovascular disease | Cardio | 1: yes, 0: no |

3.2.4 Binning of data

To leverage our clustering techniques, we employed data binning to divide clinical data into several categories. The model was well-trained and may provide the highest level of accuracy.

To comprehend the impact of age on heart disease, we categorized the ages of our patients into groups, including 20-29, 30-39, 40-49, 50-59, and 60-69. We could recognize patterns linked to different height categories by utilizing bins like <150 cm, 150-159 cm, 160-169 cm, and ≥170 cm. Similarly, to examine how weight changes impact the risk of heart disease, we divided the weight into five categories: under 60 kg, 61-69 kg, 70-79 kg, and over 80 kg—based on computer-calculated BMI values which are divided into five categories: 15-20 kg/, 21-25 kg/, 26-29 kg/, 29-33 kg/, 34-39 kg/ as shown in table 4 and the standard categories such as underweight, average weight, overweight, and obese had been made for BMI (Body et al.) thereby illustrating what body fatness meant by heart disease. Mean arterial pressure (MAP) was put into four categories, which include78-87mmhg,88-95 mmHg, 96-102 mmHg, 103-111 mmHg and 112-127mmhg (Mean et al.) as shown in table 5. This facilitated a comprehensive understanding of the effect of blood pressure on cardiovascular health. Transforming these variables to categorical data through binning enabled the K-Modes method to be helpful in practical situations. The accuracy and interpretability of heart disease predictions were improved using the binning technique, which helped organize the data into known patterns and relationships.

Table 4: categories for mean arterial pressure values

|  |  |
| --- | --- |
| **MAP values** | **Categories** |
| >=78 and <87 | 0 |
| >=87 and <95 | 1 |
| >=95 and <102 | 2 |
| >=102 and <111 | 3 |
| >=111 and <127 | 4 |

Table 5: Categories for body mass index values

|  |  |
| --- | --- |
| **BMI values** | **Categories** |
| ≥15 and <20 | 0 |
| ≥20 and <25 | 1 |
| ≥25 and <29 | 2 |
| ≥29 and <33 | 3 |
| ≥33 and <39 | 4 |

*3.3 Clustering Techniques*

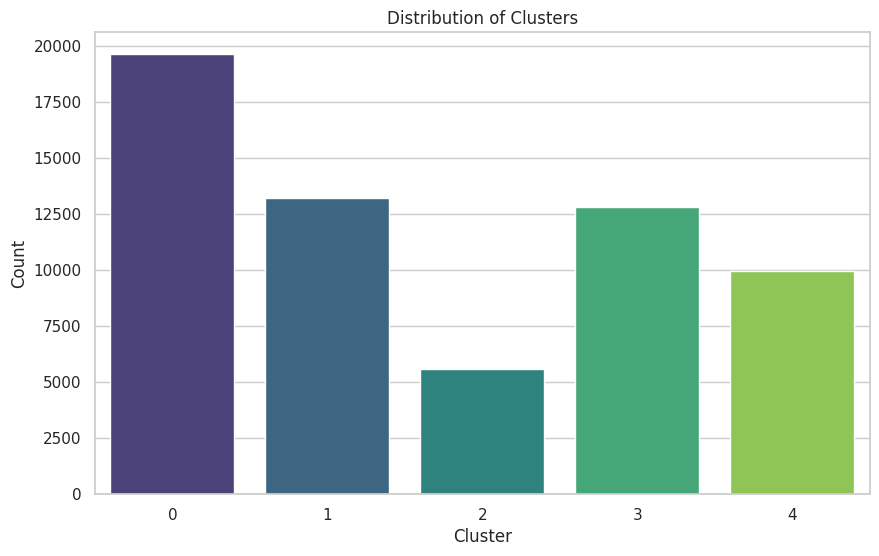
Hierarchical analysis methods are crucial for grouping similar characteristics of the data points, which also help better understand the data and assist the performance of the machine learning algorithm models through the training of homogeneous sets of data. In this study, we used the K-Modes algorithm with Huang, particularly for categorical data that use the Hamming distance in the measurement of dissimilarity and the mode as opposed to the mean for calculating the centroid. This makes K-Modes very efficient and able to handle big data with high categorical dimensions, which is typical of clinical data. K-Modes clustering was used to analyze the clinical data, which helped cluster similar patterns and relationships in the disease and comparative risk factors or symptoms. These clusters generated helpful information, which aided in increasing the quality of models due to the ability to build models that were more suitable in density and distribution for the given models. In addition, clustering provided essential features that help distinguish mean cluster characteristics and degree of variances in different clusters to make a more comprehensive selection of features and provide a more profound view of patient's demographic characteristics, disease trends, and treatment effectiveness. This entailed approach made it easier to arrive at accurate and more easily interpreted results for heart disease, thereby effectively serving the cause of accurate heart disease prediction using machine learning technology.[[26]](#_References)

Figure 5: distribution of clusters from the dataset

The bar chart indicates the distribution of clusters in the information set, indicating the wide variety of statistics points in every cluster. Cluster 0 has the best quantity with about 20,000 information points, indicating that it is the biggest part of the statistics set. Cluster 1 follows with the largest number of data factors, although much less than cluster zero. Clusters 3 and 4 also have the biggest number of statistics points, while cluster 2 has the least, indicating probably unique or rare information factors clustering It turned out , informs the following analysis and interpretation of clustering effects

*3.4 Modelling*

The dataset is split into three different types of ratios, 60:40,70:30,80:20 to find the highest accuracy, i.e., A training dataset is created from the dataset, comprising 60%, 70%, or 80% of the data, and a testing dataset is created, comprising 20%, 30%, or 40% of the data. Different classifiers, such as gradient boost, Catboost, K-Nearest Neighbours, and LGBoost, as well as neural networks, such as Recurrent Neural Networks (RNN) and Long Short-Term Models (LSTM), are applied to the clustered dataset to assess their performance. We then evaluate the performance of each classifier using accuracy, precision, recall, and F-measure scores.

3.4.1 Machine Learning models for proposed work:

3.4.1.1 Catboost:

CatBoost, short for Categorical Boosting, is a gradient-boosting algorithm that deals with categorical variables. It also applies the method of ordered boosting to handle the overfitting problem and enhance the model's overall performance. The categorical features are transformed into valuable numerical counterparts appropriate for the boosting algorithm's computation by CatBoost. As mentioned, it is a high-performance tool that is relatively easy to use when working with many categorical variables and large datasets without much preprocessing. It also includes measures for handling overfitting and will work well for most predictive model applications.[[27]](#_References)

3.4.1.2 Gradient Boosting:

Gradient boosting can be described as a concept in ensemble learning where models are trained myopically. Every new model tries to rectify the mistakes of the previous models by concentrating on the residuals [25]. They are then aggregated, frequently decision trees, to make a good decision-making model. Gradient Boosting focuses on optimizing the loss function. It is recognized for its high accuracy, with an extensive capacity for handling different kinds of data and discovering intricate patterns among the sets. The technique is widely used to increase the efficiency of the predictive model by decreasing the bias and variance, which is beneficial in classification and regression algorithms.[[28]](#_References)

3.4.1.3 Support Vector Machine (SVM)

SVM is a powerful and stable technique of supervised learning that helps find the best hyperplane that maximizes the distance from the closest points of the different classes [25]. SVM remains useful in higher dimensions and is not sensitive to the overfitting problem, especially if the number of features is significantly larger than the number of samples. SVM can process linear and non-linear datasets using kernel options for linear, polynomial, and radial base functions (RBF). They are generally preferred for classification and regression problems because of their efficiency and versatility in modeling higher-level decision spaces. According to Shorewall's study, The SVM stacking model attained a precision level of 72%. [[29]](#_References)

3.4.1.4 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a simple, instance-based learning algorithm that classifies a data point based on the majority class of its k-nearest neighbors [[24]](#_References). It is a non-parametric method that can adapt to the underlying data distribution. KNN is particularly useful when the decision boundary is complex and not easily captured by linear models. It is straightforward to implement and interpret, providing a baseline for comparison with more complex models. Despite its simplicity, KNN can be very effective in various classification and regression tasks, especially with an appropriate k and distance metric choice. In a study by Shorewall, the KNN stacking model attained an accuracy of 71.6.[[30]](#_References)

3.4.1.5 Light Gradient Boosting (LGBoost)

LightGBM (Light et al.) is an implementation of Gradient boosting that depends on decision trees for learning. It is meant to be memory precise and computationally fast, thus making it suitable for Use when dealing with big data sets. LightGBM employs the leaf-wise growth policy, which helps achieve better or equal accuracy to other boosting techniques with considerably lower training time. It is efficient when working with high-dimensional and extensive data and computationally robust. LightGBM also has those extra features, which are categorical features and the way it handles missing data.[[31]](#_References)

3.5.2 Deep Learning models for proposed work:

3.4.2.1 Recurrent Neural Networks (RNN)

A Recurrent Neural Network, or RNN for short, is a type of neural network that is used to recognize sequences of data such as time series or linguistic data. RNNs have a memory component through which they can incorporate information from the previous inputs in the sequence, which is appropriate for tasks with contextual and sequential order. Nevertheless, conventional RNNs have drawbacks, such as the vanishing gradient problem, which might negatively affect the results in the context of long sequences. Nonetheless, given this limitation, the RNN is commonly implemented in applications for language models, speech recognition, and tasks involving sequential prediction.[[32]](#_References)

3.5.2.2 Long Short-Term Model (LSTM)

A long Short-Term Memory (LSTM) network is a type of RNN designed to eliminate the drawbacks of the typical RNNs, including the vanishing gradient problem. LSTMs employ memory cells to store data as far as they can for more extensive network instances and can learn long-term dependencies. Every LSTM cell has the gates to decide about the flow of information; thus, it can decide to remember or even forget something selectively. Due to this, LSTMs are very useful in handling any problem that involves sequential data like natural language processing, speech recognition, and time series forecasting.[[33]](#_References)

*3.6 Evaluation Metrics*

During each of our steps in the evaluation phase, we employed accuracy, recall, precision, and F1-score metrics to assess the performance of each model. These were all extracted from the confusion matrix primarily used in calculating performance statistics in classification exercises. The confusion matrix shows four possible outcomes of predictions. Therefore, the two-dimensional table contains four estimations, referred to as true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

*Accuracy*: Accuracy is the most straightforward of all the metrics that show the catalog of success of the prediction provided by the model, it is one of the most unbiased and commonly used analytical parameters for classification models. It is defined as the proportion of positively anticipated events in the total duration. Although accuracy is a useful and popular measure of model performance, it has limitations, especially for imbalanced data sets. So, while accuracy gives you the overall picture of how well the model plays, it’s important to remember distance plus other metrics to get a more nuanced understanding of the model’s performance.

*Precision*: Precision reveals how many of the positive predictions were positive. As stated earlier, precision is helpful when the cost of a misdiagnosis is steep. It is calculated by the ratio of true positive predictions to the total number of positive predictions (true positives + false positives). Precision is important where the cost of false positives is high. For example, high precision is desirable for a test designed to diagnose disease in medicine because false positives can lead to unnecessary and potentially harmful treatment Specificity helps to understand the reliability of the model’s good predictions so down, making it an important metric for ensuring the internal confidence of its forecasts.

*Recall (Sensitivity or True Positive Rate)*: It tells the proportion of actual positives correctly classified by the model, also called the Recall or the Sensitivity. Increased memory is important in situations where it is important to capture as much positive information as possible, such as when diagnosing diseases or identifying fraudulent transactions. Higher recall means that the model has a lower rate of false negatives, indicating that it is better able to identify most positive cases. However, when it occurs at the expense of a sufficient number of false positives, increased recall alone may not be sufficient, thus requiring a balance between recall accuracy

Availability is relevant when the cost of false negatives is high to ensure that positive cases are remembered.

F1-Score: Meanwhile, the F1-Score measures the technique's precision blended with the do not forget of this technique. F1-Score takes Precision and Recall together into one score and accordingly gives a more focused perception into the model's perfor-mance. This metric is useful as it captures both fake positives and false negatives, and offers us a general concept of the version’s performance. A high F1-rating suggests that the version has high precision and excessive take into account, making it a strong metric for model evaluation where each metrics are critical.

AUC-ROC (Area Under the Receiver Operating Characteristic Curve): ROC is drawn based on the True Positive rate TPR against the False Positive rate FPR at different threshold settings. From this, one can plot the actual positive rate against the false positive rate, and the AUC-ROC indicates the performance of the model at the decision point by giving a single measurement. AU-ROC values are represented by a number between 0 and 1, where the closer it is to one, the better the prediction model. From 5 assumed random guesses to as little as one. Where perfect discrimination is achieved, the measure equals 0, with higher values providing better performance. [[34]](#_References)

4. Experimental Results:

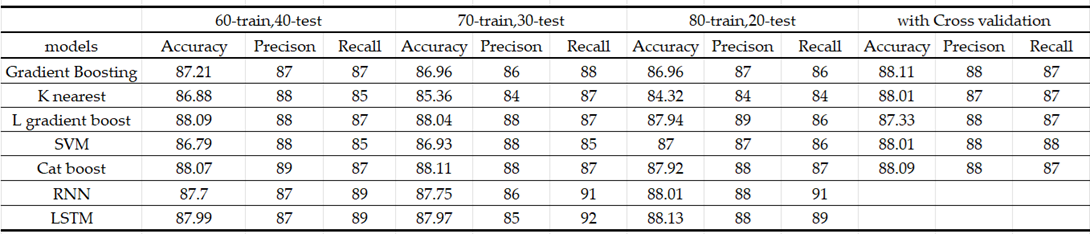
  
This research utilized Visual Studio Code on an Intel i5 processor and 16GB of RAM. The input dataset initially contained 70,000 rows and 12 attributes; after data cleaning, the dataset was downsized to 61,158 rows and 11 attributes. All predictors were categorical, and outliers were removed using z-scores to enhance model efficiency. The algorithms employed in the study included CatBoost, Gradient Boosting, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Light Gradient Boosting Model (LGB), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM). We used corresponding evaluation metrics to assess the models. We split the dataset into three ratios: 60% train and 40% test, 70% train and 30% test, and 80% train and 20% test.  
  
Hyperparameters were determined using an automated technique using the GridSearchCV method. This method, available in the scikit-learn library, applies the k-fold cross-validation technique to evaluate models with different sets of hyperparameters and returns the best set of hyperparameters that gives the maximum score.

Table 6 Experimental result Table

The experimental results highlight the performance of various machine learning models evaluated under different train-test splits (60-40, 70-30, 80-20) and with cross-validation. The models tested include Gradient Boosting, K-Nearest Neighbors (K-nearest), Logistic Regression with Gradient Boost (L gradient boost), Support Vector Machine (SVM), CatBoost, Recurrent Neural Network (RNN), and Long Short-Term Memory (LSTM). Across the train-test splits and cross-validation, the CatBoost model consistently shows high accuracy, precision, and recall, with the highest accuracy of 88.11% observed in the 70-30 split and with cross-validation. The LSTM model also performs well, achieving an accuracy of 88.13% with the 80-20 split and comparable precision and recall across different splits. Gradient Boosting maintains stable performance with accuracy ranging from 86.96% to 88.11%, while SVM shows slightly lower but consistent accuracy. K-nearest exhibits the lowest accuracy in the 80-20 split at 84.32%, indicating it might be less effective compared to other models in this context. RNN shows notable performance as well, especially in the 80-20 split and with cross-validation, suggesting that it can effectively capture temporal dependencies in the data. The logistic regression model with gradient boost (L gradient boost) also demonstrates robust results, particularly excelling in precision and recall metrics, making it a reliable choice for binary classification tasks. The consistent high performance of CatBoost across different splits suggests that it is particularly adept at handling the complexities within the dataset. Meanwhile, the high precision and recall values across the board for most models indicate their effectiveness in correctly identifying both the presence and absence of cardiovascular disease. Overall, the models demonstrate good performance in predicting cardiovascular disease, with CatBoost and LSTM emerging as the top performers. Cross-validation results suggest that the models maintain their robustness and generalizability, providing reliable performance metrics for practical applications. These findings underscore the potential of these machine learning approaches to significantly enhance the predictive accuracy and reliability of cardiovascular disease diagnosis.

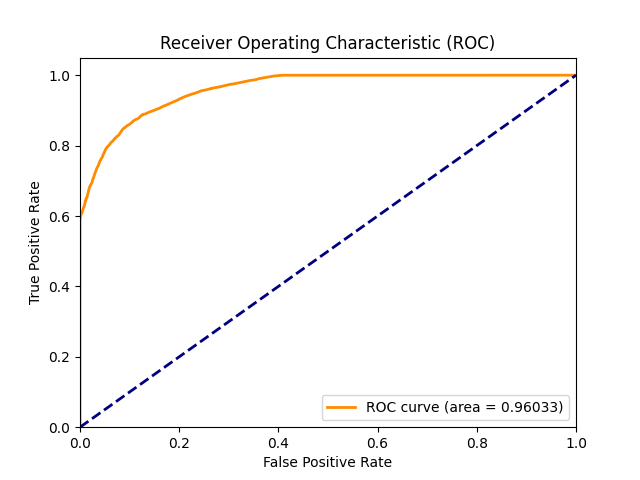
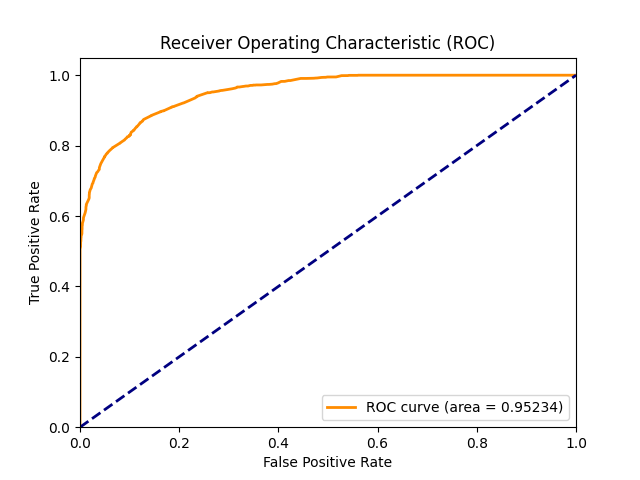
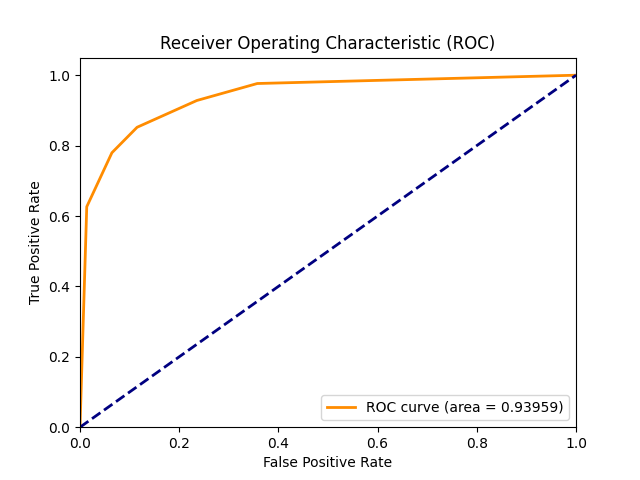
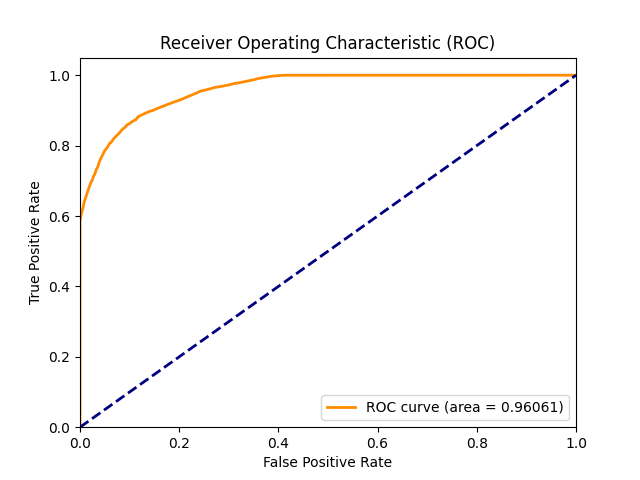
Gradient Boosting: Achieved an accuracy of 87.21% (60-40 split), 86.96% (70-30 split), 86.96% (80-20 split), and 88.11% (with cross-validation). The precision was 87%, and recall was 87% across all splits, with an F1-score of 88 and AUC-ROC of 0.95.  
  
CatBoost: For Catboost classifier, the Dataset was preprocessed by scaling the features using StandardScaler and splitting it into training and test sets with an 80-20 ratio. Catboost model achieved an accuracy of 88.07% (60-40 split), 88.11% (70-30 split), 87.92% (80-20 split), and 88.09% (with cross-validation). Precision and recall were consistently 88% and 87%, respectively, with an F1-score of 88 and AUC-ROC of 0.95.  
  
Light Gradient Boosting Model (LGB): The dataset was preprocessed by scaling the features using StandardScaler and splitting it into training and test sets with an 80-20, 70-30, 60-40ratio. The model achieved an accuracy of 88.09% (60-40 split), 88.04% (70-30 split), 87.94% (80-20 split), and 87.33% (with cross-validation). Precision and recall were 88% and 87%, respectively, with an F1-score of 88 and an AUC-ROC of 0.95.  
  
K-Nearest Neighbors (KNN): The dataset was preprocessed by scaling the features using StandardScaler and splitting it into training and test sets with an 80-20, 70-30, 60-40ratio. The model achieved an accuracy of 86.88% (60-40 split), 85.36% (70-30 split), 84.32% (80-20 split), and 88.01% (with cross-validation). Precision was 88%, recall was 85%, with an F1-score of 86 and AUC-ROC of 0.95.  
  
Support Vector Machine (SVM): The dataset was preprocessed by scaling the features using StandardScaler and splitting it into training and test sets with an 80-20, 70-30, 60-40ratio. The model achieved an accuracy of 86.79% (60-40 split), 86.93% (70-30 split), 86.96% (80-20 split), and 88.01% (with cross-validation). Precision and recall were consistently 88% and 85%, respectively, with an F1-score of 87 and AUC-ROC of 0.95.  
  
Recurrent Neural Networks (RNN): This study used a Recurrent Neural Network (RNN) model to predict cardiovascular disease using the TensorFlow Keras API. The model added several layers, starting with a SimpleRNN layer with 50 units and a ReLU activation function, which was configured to return sequences to capture the time dependence in the data followed by a Dropout layer at rate, it is 0.2 to prevent over-fitting. When ReLU was activated with 50 units, a new SimpleRNN layer, also configured to return in sequence, was added in the same way as a new Dropout layer. A third SimpleRNN layer of 50 units was then added. The output was then flattened with a Flatten layer to convert the 3D tensor into the appropriate 2D tensor for the Dense layer. The dense layers consisted of a 128-unit layer followed by a Dropout layer at a rate of 0.3, and then a 64-unit layer followed by another Dropout layer at the same rate Finally, the output layer to generate the probability of binary classification is constructed dense layer with a unit sigmoid activation function -The model was compiled with cross-entropy loss functions, and accuracy as an evaluation metric, and trained for 50 epochs with a batch size of 32 and a 0.2, 0.4, 0.6-validation split, obtaining a remarkable performance metric to predict heart disease. The model has achieved an accuracy of 87.7% (60-40 split), 87.75% (70-30 split), 88.01% (80-20 split), and 88.01% (with cross-validation). Precision and recall were 87% and 89%, respectively, with an F1-score of 88 and AUC-ROC of 0.95.  
  
Long Short-Term Memory (LSTM): This study used long-term and short-term memory (LSTM) neurons to predict cardiovascular disease using TensorFlow’s Keras API. The model architecture included multiple LSTM layers to capture sequential dependencies in the data. Initially, an LSTM layer with 64 units and a ReLU activation function were used, configured to return a series of forward layers, followed by a Dropout layer with a rate of 0.2 to reduce overfitting schedule these two more times, each with 64 units in the LSTM layer and 0.2, 0.4. 0.6 in the Dropout layer. The final LSTM layer also consisted of 64 units but did not return in any order, marking the end of conventional layers. The results from these layers were flattened using the Flatten layer. Then, the model had dense layers starting with the 128-unit layer with ReLU activation followed by a Dropout layer with rate 0.3, another Dense layer with 64 units, and ReLU activation, with another Dropout layer a it has the same rate There was a complex with sigmoid activation activity. The models were compiled using the ADAM optimizer, binary cross-entropy loss function, and accuracy as a search metric. Model was trained for 50 epochs with a batch size of 32 and a validation split of 0.2, 0.4 ,0.6, they obtained high accuracy and significant performance metrics for predicting cardiovascular disease. The model has achieved an accuracy of 87.99% (60-40 split), 87.97% (70-30 split), 88.13% (80-20 split), and 88.08% (with cross-validation). Precision was 88%, and recall was 89%, with an F1-score of 88 and AUC-ROC of 0. 95.

Figure 7: precision of different models

Figure 6: accuracy of different models

Figure 9: accuracy of different models using cross validation

Figure 8: recall of different models

Figures 5 to 8 present a detailed analysis of the performance parameters of different models in different data partitions (60-40, 70-30, 80-20) and the results of the accuracy of the model without cross-validation, indicating that the L-slope-. The Boost and CatBoost models consistently achieve high accuracies (about 88%), while K-Nearest Neighbors and SVM exhibit lower accuracies, especially with separations of 70-30 Figure 6 highlights the precision of the models, while L -gradient Boost and CatBoost again in all partitions show the highest precision scores (close to 88%), and K-Nearest Neighbors has the lowest precision, especially for the 70-30 partitions Figure 7 presents the residual scores, where The RNN and LSTM models have high recall values ​​(91% and , respectively). Finally, Figure 8 compares the version accuracy for model with cross-validation, displaying a mild improvement across all fashions, with CatBoost having the highest accuracy, and L-gradient improve accompanied intently by way of RNN following K-Nearest Neighbors though but nevertheless lags even in supplying move- validation. Overall, those findings endorse that the L-Gradient Boost, CatBoost, and RNN/LSTM models usually outperform others in accurate individuals for movement to create those organisms it protects this weakness.

gradient boost

light gradient boost

k nearest neighbour

catboost

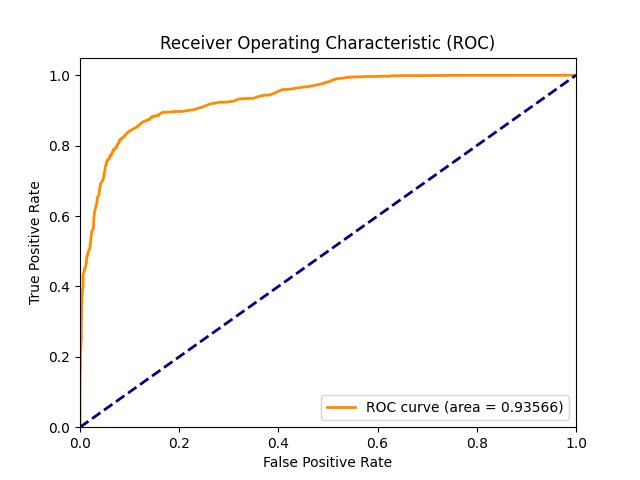
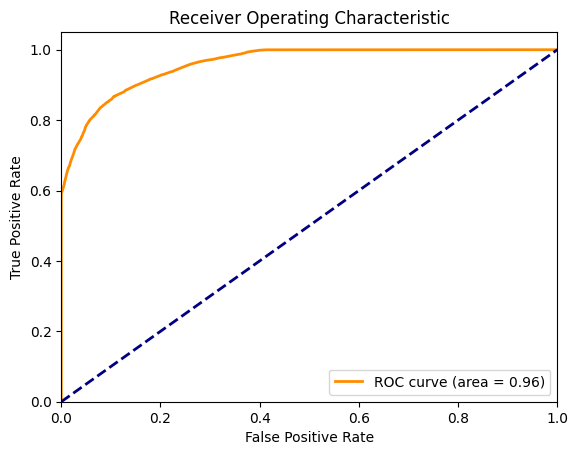
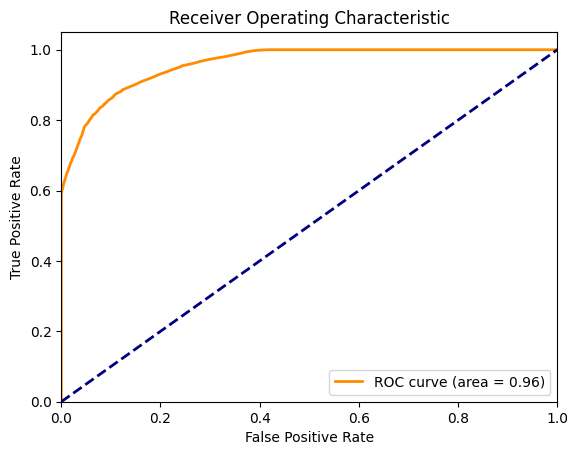


figure 10: roc-area under curve

long short-term memory

recurrent neural networks

support vector machine

In machine learning models, ROC curves including Gradient Boosting, K-Nearest Neighbors (KNN), LightGBM, Support Vector Machine (SVM), CatBoost, Recurrent Neural Network (RNN), and Long Short-Term Memory (LSTM) are used to compare their classification performance in detail. The Gradient Boosting, Light Gradient Boosting and CatBoost models show good performance with ROC curves that are close to the left corner which indicates high sensitivity and specificity, while KNN and SVM models also perform well although AUC is lower than expected. slightly less compared to Gradient Boosting and CatBoost. The KNN model shows degrading performance, while the conventional RNN-LSTM model for sequences shows \t exceptional results in deep learning , ensemble methods such as Gradient Boosting, Light Gradient Boosting , CatBoost and deep learning models like RNN, LSTM outperforms other models in ROC curve analysis.

5. Comparison Analysis

In this study, we carried out an extensive evaluation of diverse system gaining knowledge of patterns for predicting cardiovascular disease, contrasting our findings with preceding studies inside the study. Our studies examined Gradient Boosting, K-Nearest Neighbors (KNN), Light Gradient Boosting Machine (LGBM), Support Vector Machine (SVM), CatBoost, Recurrent Neural Network (RNN), and Long Short-Term Memory (LSTM) fashions. The fashions were assessed the usage of exceptional training-trying out splits (60:40 ratio, 70:30 ratio, 80:20 ratio) and pass-validation, with metrics along with accuracy, precision, and do not forget. Notably, our pinnacle-appearing models, CatBoost and LGBM, carried out accuracy rates of 88.11% and 88.33%, respectively, with pass-validation, demonstrating giant upgrades over preceding studies. For instance, Shorwell et al., [5] carried out a maximum accuracy of 75.1% using KNN, Random Forest, and SVM with logistic regression, whilst Maiga et al. [7] reached 70% accuracy the usage of Random Forest, Naive Bayes, Logistic Regression, and KNN. Wajgi et al. [12] suggested an accuracy of 72.77% with Decision Trees, and Khan and Mondal [20] attained as much as 72.72% accuracy with various strategies together with neural networks and logistic regression. Bhatt et al. [18] executed an accuracy of 87.28% the use of a aggregate of ok-mode clustering and diverse device studying fashions such as MLP and XGBoost. In their work, Elsedawy [19] achieved the highest level of accuracy among all previous studies at 89.01% using the Random Forest classifier on UCI cardiovascular dataset. The study not only shows outstanding performance, but it also utilizes strong cross-validation to assure reliability and generalizability. Our models’ enhanced performance metrics especially with CatBoost and LGBM highlight the efficiency of modern machine learning technologies.

6. Conclusion and Future Work:

This study showed how different machine learning and deep learning models could be used to predict cardiovascular diseases. Extensive preprocessing, that included cleaning of data, transformation and removal of outliers ensured that the models were trained on high-quality data resulting into strong and accurate predictions. Long Short-Term Memory (LSTM) and CatBoost were found to be the best out of the models tested (88.13% accuracy and 88.11% accuracy), which indicates that it can handle sequential patterns in data as well as maintain long-term associations.

Various data split ratios and cross-validation techniques therefore assessed all models’ performance to highlight the importance of these methods in obtaining reliable outcomes. This helped us understand how our model behaves since we learned about some sorts of visualizations like heatmaps and count plots which usually goes so far as providing knowledge, especially about relationship between attributes structures in our data.

Random Search or Bayesian Optimization are some of the advanced techniques through which extensive hyperparameter tuning can be performed to enhance LSTM model work in the future, this will help identify the most optimal configurations. In order to make LSTM model robust and reliable, there shall be comprehensive metrics like average accuracy, precision, recall, F1-score and AUC-ROC across different folds basing on detailed cross-validation. To determine how features contribute to a model, SHapley Additive exPlanations (SHAP) or Local Interpretable Model-agnostic Explanations (LIME) will be employed. Its generalization ability is also tested using an external dataset to determine whether overfitting was avoided when using it as a classifier. Lastly, combining other top-performing algorithms with the LSTM model would improve its predictive power and reliability by investigating ensemble methods that could help in making better predictions for given values.

# **References**

1. Estes, C.; Anstee, Q.M.; Arias-Loste, M.T.; Bantel, H.; Bellentani, S.; Caballeria, J.; Colombo, M.; Craxi, A.; Crespo, J.; Day, C.P.; et al. Modeling NAFLD disease burden in China, France, Germany, Italy, Japan, Spain, United Kingdom, and United States for the period 2016–2030. J. Hepatol. 2018, 69, 896–904. [[CrossRef]](https://www.journal-of-hepatology.eu/article/S0168-8278(18)32121-4/fulltext)
2. Dro˙ zd˙ z, K.; Nabrdalik, K.; Kwiendacz, H.; Hendel, M.; Olejarz, A.; Tomasik, A.; Bartman, W.; Nalepa, J.; Gumprecht, J.; Lip, G.Y.H. Risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease: A machine learning approach. Cardiovasc. Diabetol. 2022, 21, 240[. [CrossRef]](https://cardiab.biomedcentral.com/articles/10.1186/s12933-022-01672-9)
3. Murthy, H.S.N.; Meenakshi, M. Dimensionality reduction using neuro-genetic approach for early prediction of coronary heart disease. In Proceedings of the International Conference on Circuits, Communication, Control and Computing, Bangalore, India, 21–22 November 2014; pp. 329–332. [[CrossRef]](https://ieeexplore.ieee.org/document/7057817)
4. Benjamin, E.J.; Muntner, P.; Alonso, A.; Bittencourt, M.S.; Callaway, C.W.; Carson, A.P.; Chamberlain, A.M.; Chang, A.R.; Cheng, S.; Das, S.R.; et al. heart disease and stroke statistics—2019 update: A report from the American heart association. Circulation. [[CrossRef]](https://www.ahajournals.org/doi/10.1161/CIR.0000000000000659)
5. Shorewala, V. Early detection of coronary heart disease using ensemble techniques. Inform. Med. Unlocked 2021, 26, 100655. [[CrossRef]](https://www.sciencedirect.com/science/article/pii/S235291482100143X?via%3Dihub)
6. Mozaffarian, D.; Benjamin, E.J.; Go, A.S.; Arnett, D.K.; Blaha, M.J.; Cushman, M.; de Ferranti, S.; Després, J.-P.; Fullerton, H.J.; Howard, V.J.; et al. Heart disease and stroke statistics—2015 update: A report from the American Heart Association. Circulation 2015, 131, e29–e322. [[CrossRef]](https://www.ahajournals.org/doi/10.1161/CIR.0000000000000152)
7. Maiga, J.; Hungilo, G.G.; Pranowo. Comparison of Machine Learning Models in Prediction of Cardiovascular Disease Using Health Record Data. In Proceedings of the 2019 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), Jakarta, Indonesia, 24–25 October 2019; pp. 45–48. [[CrossRef]](https://ieeexplore.ieee.org/document/8985205)
8. Li, J.; Loerbroks, A.; Bosma, H.; Angerer, P. Work stress and cardiovascular disease: A life course perspective. J. Occup. Health 2016, 58, 216–219. [[CrossRef]](https://academic.oup.com/joh/article/58/2/216/7250738)
9. Purushottam; Saxena, K.; Sharma, R. Efficient Heart Disease Prediction System. Procedia Comput. Sci. 2016, 85, 962–969. [[CrossRef]](https://www.sciencedirect.com/science/article/pii/S187705091630638X?via%3Dihub)
10. Soni, J.; Ansari, U.; Sharma, D.; Soni, S. Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. Int. J. Comput. Appl. 2011, 17, 43–48. [[CrossRef]](https://www.ijcaonline.org/volume17/number8/pxc3872860.pdf)
11. Mohan, S.; Thirumalai, C.; Srivastava, G. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. IEEE Access 2019, 7, 81542–81554. [[CrossRef]](https://ieeexplore.ieee.org/document/8740989)
12. Waigi, R.; Choudhary, S.; Fulzele, P.; Mishra, G. Predicting the risk of heart disease using advanced machine learning approach. Eur. J. Mol. Clin. Med. 2020, 7, 1638–1645. [[CrossRef]](https://www.researchgate.net/publication/348192776_Predicting_the_risk_of_heart_disease_using_advanced_machine_learning_approach)
13. K,V.; Singaraju, J. Decision Support System for Congenital Heart Disease Diagnosis based on Signs and Symptoms using Neural Networks. Int. J. Comput. Appl. 2011, 19, 6–12. [[CrossRef]](https://www.ijcaonline.org/volume19/number6/pxc3873115.pdf)
14. Narin, A.; Isler, Y.; Ozer, M. Early prediction of Paroxysmal Atrial Fibrillation using frequency domain measures of heart rate variability. In Proceedings of the 2016 Medical Technologies National Congress (TIPTEKNO), Antalya, Turkey, 27–29 October 2016. [[CrossRef]](https://ieeexplore.ieee.org/document/7863110)
15. Shah, D.; Patel, S.; Bharti, S.K. Heart Disease Prediction using Machine Learning Techniques. SN Comput. Sci. 2020, 1, 345. [[CrossRef]](https://link.springer.com/article/10.1007/s42979-020-00365-y)
16. Alotaibi, F.S. Implementation of Machine Learning Model to Predict Heart Failure Disease. Int. J. Adv. Comput. Sci. Appl. 2019, 10, 261–268. [[CrossRef]](https://thesai.org/Publications/ViewPaper?Volume=10&Issue=6&Code=IJACSA&SerialNo=37)
17. Hasan, N.; Bao, Y. Comparing different feature selection algorithms for cardiovascular disease prediction. Health Technol. 2020, 11, 49–62. [[CrossRef]](https://link.springer.com/article/10.1007/s12553-020-00499-2)
18. Bhatt, Chintan & Patel, Parth & Ghetia, Tarang & Mazzeo, Pier Luigi. (2023). Effective Heart Disease Prediction Using Machine Learning Techniques. Algorithms. 16. 88. 10.3390/a16020088. [[Crossref]](https://www.mdpi.com/1999-4893/16/2/88)
19. Ouf, S.; ElSeddawy, A.I.B. A proposed paradigm for intelligent heart disease prediction system using data mining techniques. J. Southwest Jiaotong Univ. 2021, 56, 220–240. [[CrossRef]](http://jsju.org/index.php/journal/article/view/949)
20. Khan, I.H.; Mondal, M.R.H. Data-Driven Diagnosis of Heart Disease. Int. J. Comput. Appl. 2020, 176, 46–54. [[CrossRef]](https://www.ijcaonline.org/archives/volume176/number41/khan-2020-ijca-920549.pdf)
21. Statistics - Train and test data split[[CrossRef]](https://datascience.stackexchange.com/questions/17328/statistics-train-and-test-data-split)
22. Kaggle Cardiovascular Disease Dataset. Available online [[CrossRef]](https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset)
23. An in-depth guide to efficient gradient boosting. Available online [[CrossRef]](https://medium.com/@mohtasim.hossain2000/mastering-lightgbm-an-in-depth-guide-to-efficient-gradient-boosting-8bfeff15ee17)
24. Hasan, S M & Mamun, Md. Al & Uddin, Md Palash & Hossain, Dr Md Ali. (2018). Comparative Analysis of Classification Approaches for Heart Disease Prediction. [[Crossref]](https://www.researchgate.net/publication/334929171_Comparative_Analysis_of_Classification_Approaches_for_Heart_Disease_Prediction)
25. Yu,D.; Zhao, Z.; Simmons, D. Interaction between Mean Arterial Pressure and HbA1c in Prediction of Cardiovascular Disease Hospitalisation: A Population-Based Case-Control Study. J. Diabetes Res. 2016, 2016, 8714745. [CrossRef] [PubMed]
26. Huang,Z.AFast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining. DMKD 1997, 3, 34–39.
27. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2017). CatBoost: Unbiased boosting with categorical features. *ArXiv*. /abs/1706.09516 [[CrossRef]](https://arxiv.org/abs/1706.09516)
28. Jerome H. Friedman. "Greedy function approximation: A gradient boosting machine." Ann. Statist. 29 (5) 1189 - 1232, October 2001. [[CrossRef]](https://doi.org/10.1214/aos/1013203451)
29. Cortes, C., & Vapnik, V. (1995). "Support-vector networks." *Machine Learning*, 20(3), 273-297(1995). [[CrossRef]](https://doi.org/10.1007/BF00994018)
30. T. Cover and P. Hart, "Nearest neighbor pattern classification," in *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21-27, January 1967. [[CrossRef]](https://ieeexplore.ieee.org/document/1053964)
31. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). "LightGBM: A highly efficient gradient boosting decision tree." *Advances in Neural Information Processing Systems*, 30. [[CrossRef]](https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf)
32. Rumelhart, D., Hinton, G. & Williams, R. Learning representations by back-propagating errors. *Nature* 323, 533–536 (1986). [[CrossRef]](https://doi.org/10.1038/323533a0)
33. Sepp Hochreiter, Jürgen Schmidhuber; Long Short-Term Memory. *Neural Comput* 1997; 9 (8): 1735–1780. doi: [[CrossRef]](https://doi.org/10.1162/neco.1997.9.8.1735)
34. Powers, D. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*, *2*(1), 37-63. [[Crossref]](https://www.researchgate.net/publication/228529307_Evaluation_From_Precision_Recall_and_F-Factor_to_ROC_Informedness_Markedness_Correlation)