International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Comparative Analysis for Prediction of Coronary Artery Disease Using Machine Learning Algorithms

Ravi Hosur[1], Pavan Mahendrakar[2], Mahesh Nagaral[3], Anand Hiremath[4], Shashidhar B. Patil[5]

Submitted: 25/11/2023 **Revised**: 30/12/2023 **Accepted**: 10/01/2024

Abstract: Cardiovascular disease, another name for heart disease, is linked to a number of conditions that affect the heart. Over the past few decades, heart disease has consistently been the leading cause of death. Numerous risk factors for heart disease are also identified, as well as the need of early disease management. This study includes a number of heart disease-related characteristics as well as models based on machine learning techniques like Nave Bayes, Convolution Neural Networks, and Logistic Regression. All previous trials relate to utilising a subset of 14, but we used the publicly accessible UCI heart disease database, which has 76 features. The purpose of this study is to estimate a patient's risk of getting heart disease. We have applied three machine learning classifiers for comparative analysis. In comparison to CNN and the Naive Bayesian algorithm, Logistic Regression has a higher accuracy of 93.22 percent.

Keywords: *Coronary Artery Disease (CAD); Naïve Bayes; Convolution Neural Network (CNN); Logistic Regression; Machine Learning, Generative Adversarial Networks; K-Nearest Neighbor (KNN).*

1. Introduction

Coronary Artery Disease, also known as CAD, is the most dangerous and death-causing chronic disease that quickly accelerates in both economically undeveloped and developed countries and may even cause death. This study shows that machine learning algorithms can be implemented with high accuracy and help predict coronary artery disease (CAD) using publicly available datasets. We limited our analysis to Logistic Regression, CNN, and Nave Bayes [NB] algorithms. These were decided based on the literature review and the attributes of the techniques. The machine learning algorithms were used to train and analyze the data set containing the different patients' test results. These algorithms were also tested for the accuracy of plotting a graph using matplotlib. The importance of absorbing valuable information from raw data has very good after-effects in many fields of life, such as the medical area, the business area, and more.

Often, with aging, atherosclerosis occurs. Therefore, the expanded categorization and models can give an error-free prediction of heart diseases by using the machine learning algorithms for coronary heart disease and reduce the number of incorrect identifications that may harm patients.

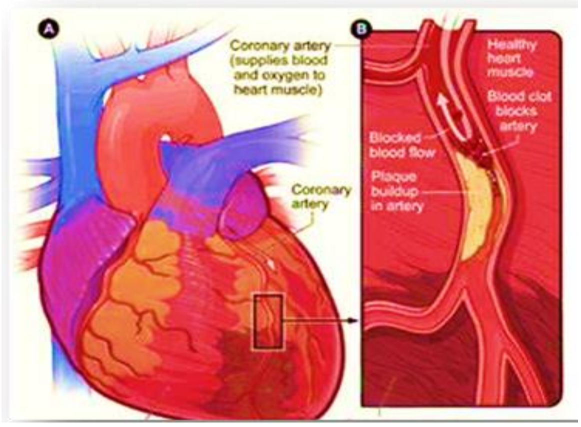

Fig. 1. (a) Coronary Arteries of the Heart; (b) Blockage (Plaque Built up) in the coronary artery. [14]

2. Related works

Md. Razu Ahmed et al. [1] described five supervised learning methods based on ML algorithms. Later, the action of five divisions is in contrast to the evaluation of performance using a ten-fold cross-validation method and confusion matrix. Therefore, they proposed a four-tier cloud architecture for the prediction of heart disease and evidence-based chasing based on real-time. Based on certain parameters, the authors used five ML techniques for the possible early detection of heart disease. Real-time data from health care centres and patients is collected and, in addition, the application is able to perform early identification of heart dis-ease based on the cloud application.

1,2,3,4 Department of Computer Science and Engineering,
BLDEA's V. P. Dr. P. G. Halakatti College of Engineering and Technology, Vijayapur586103, Karnataka, India
5 Department of ComputerApplications (MCA),
BLDEA's V. P. Dr. P. G. Halakatti College of Engineering and Technology, Vijayapur586103, Karnataka, India
Email Id: hosuravi@gmail.com,
pavandmahendrakar@gmail.com,
s.nagaral@bldeacet.ac.in,
hiremath.a@gmail.com,
mca.patil@bldeacet.ac.in

Asma Baccouche et al. [2] studied the binate classification of CAD datasets taxed in a phrase of size. Several important features and the distribution of rewards The main focus of the proposed study is to break down four types of heart disease. The classification of the problem is done by plotting different models of neural networks. The methodology substructure involves recurrent architecture for a neural network, a unidirection-al neural network, including a convolution neural network. The authors also conducted many experiments to pick the best model for every outcome class label. Despite the quality of the structured dataset, nonetheless, the designed model attains high-performance classification. The author also concluded that further work may also include the implementation of other neural network designs such as “attention-based recurrent neural networks” or “generative adversarial networks (GAN)”.

Yang Peili et al. [3] developed a Deep Learning approach for the management of CAD in the early days. CHD [coronary heart disease] is the usual chronic disease alarming people’s fitness as well as lives. To simulate the CHD research early moral, we have studied the clue challenges and objections about fact control regarding the CHD early caution research built on the DL technique and an evolved system of data control which blends the patient data and the Deep Neural Network model data. Deep learning created conditions for the re-searcher's broad awareness as more control objects appeared in DL tasks. The data on the CHD of prior detection of the patient is not only used for training purposes but also used for support analysis. The Find-specifies-ancestor and tracking algorithms are outlined to demeanor the stock for deep learning model management. Based upon the experiment inquiry, to choose between the document-oriented NoSQL database and MongoDB database for pdmdims, CHD patient data is used as the guide. The patient data schema and the model data are sketched and moved to one relational schema and a “document-oriented NoSQL schema”. Ex-amine and investigate CHD patient data and evidence, as well as the DNN model data vision, and successfully present it at an exhibit.

Aditi Gavhane et al. [4] proposed a method to predict heart disease using machine learning. The output predicted by this system will be either YES or NO as a result. The status of the heart leading to CAD is given by the system. If a patient gets a positive test result, then for future diagnosis, he/she needs to approach the cardiologist.

Md Ashraful Alam et al. [5] developed the technique for early detection of coronary artery blockage using image processing. Good health is the priority of mankind. Thus, the detection of diseases like CAD will be a blessing if predicted early. There is always room for some mistakes if someone prefers the help of a doctor or a manual method for the diagnosis of heart disease. The proposed system finds the artery blockage in the initial phase. The priority of this arrangement is to find the artery blockage with fewer faults in degree. Ac-cording to the parameters of the patient’s past health and pre-trained dataset, the risk of CAD is predicted.

Vardhan Shorewala et al. [6] used K-nearest neighbors in addition to the statistical approach to play a vital role in the effective selection of data from the dataset. The average accuracy of 71.92% has been acquired by analyzing the base model, while 73.97% accuracy is acquired by the neural network approach. The effective-ness in raising the base model accuracy is provided by boosting, stacking, and bagging. The most productive ensemble technique involves a heterozygous model, achieving 75.1% accuracy. This also involves a support vector machine, a random forest classifier as a meta-classifier with logistic regression, and stacking KNN. Currently, the limitations for studying ensemble techniques are bagging, stacking, and boosting.

Abderrahmane Ed-daoudy et al. [7] worked on the Spark and Cassandra frameworks that are used for monitoring arterial disease. The introduced system targets the application of a real-time classification model based on the attributes of artery-disease for continuous observation. This system is mainly divided into two subparts, namely data storage and streaming processing, and visualization. The Spark framework is the basis for monitoring heart disease.

Aravind Akella et al. [8] developed machine learning algorithms for predicting coronary artery disease. They have predicted heart disease with the help of ML algorithms. Seeing that ML algorithms are showing a great impact on predicting and diagnosing heart diseases, the author has used the six ML algorithms (regression tree, support vector machine, linear regression, random forest, neural network, nearest neighbor, neural net-work, and k-nearest neighbor). These algorithms work perfectly when accuracy is greater than 80%, and the accuracy of the nearest neighbour should be more than 93%. Here we use a dataset that is stored in a store-house from which the dataset can be downloaded and maintained by the many centers for Intelligent Systems and Machine Learning (the USA, University of California, CA).

Yar Muhammad et al. [9] discussed several contemporary classification algorithms regarding their experimental results. First, all the classification representations are checked against the full feature space. Those models are “Adaboost (AB), Gradient Boosting (GB), Random Forest (RF), Logistic Regression (LR), Nave Bayes (NB), Artificial Neural Network (ANN), Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Decision

Tree (DT), Extra-Tree Classifier (ETC)”). Four feature selection algorithms have been applied to select outstanding and extremely alternative features from feature space. These are (“Minimal Redundancy Maximal Relevance (mRMR), Least Absolute Shrinkage and Selection Operator (LASSO), Correlation-Based Filter (FCBF), Fast Correlation-Based Filter (FCBF)”). In this paper, the author used some materials and methods.

Ying Fang et al. [10] proposed a method for identifying coronary artery disease, and the calculations will be derived from the machine learning models for the considered morphological features. So, the result of the experiment shows that the polynomial-SVM model performs very well among all regarded machine learning models. Moreover, there are two more necessary features in predicting coronary artery disease; they are (AER) the area expansion ratio and (n) the exponent of vessel diameter (n). This study also showed a new way to predict coronary artery disease using a combination of imaging-based morphological measurement methods and some machine learning techniques.

Conclusion from the Literature Survey: Most researchers work on different methodologies for the prediction of coronary artery disease. It is also evident that a machine learning algorithm is considered the best method for the prediction of “coronary artery disease (CAD)”.

3. Dataset and Attributes

The dataset is publicly as UCI Heart Disease Dataset [11,15]. It gives us the particular details, which have 4,000 pages of documentation and 14 features or attributes. These characteristics/attributes include gender, age, resting blood pressure, type of chest pain, exercise-induced angina, maximum heart rate, sugar blood, electrocardiographic results, ST depression caused by exercise, number of major vessels, slope of peak exercise, and a target classification ranging from 0 to 2, with 0 indicating no heart disease. The provided dataset is in the form of a CSV format.

Table I. Attributes and its Description from UCI Heart Disease Dataset.

Attributes	Description
age	age in years
sex	1 = male, 0 = female
cp	chest pain type: 1: typical angina, 2: atypical angina, 3: non-anginal pain, 4: asymptomatic
treetops	resting blood pressure in mm Hg
choli	serum cholesterol in mg/dl
fbs	fasting blood sugar > 120 mg/dl: 1 = true, 0 = false
restecg	0: normal, 1: having ST-T wave abnormality, 2: left ventricular hypertrophy
thalach	maximum heart rate achieved
Exang	exercise induced angina (1 = yes; 0 = no)
Oldpeak	ST depression induced by exercise relative to rest
Slope	the slope of the peak exercise ST segment: 1: upsloping, 2: flat, 3: downsloping
Ca	number of major vessels: (0-3) colored by flourosopy
Thal	3: normal, 6: fixed defect, 7: reversable defect
Target	diagnosis of heart disease: (0 = false, 1 = true)

The educational data is component to the heart disease of an individual, so it is released. Further the experiments are carried out with this pre-processing of the data.

4. Overview of Methods

A flowchart is a diagrammatic representation that describes a workflow or process. The methodology involves five major steps, and they are: Fetching of data; Wrangling; Analysing the data; Modelling and Appraisal.

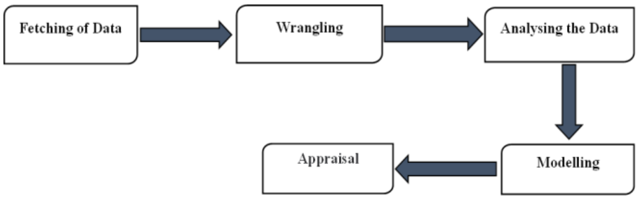

Fig 3. Flow chart for prediction of coronary artery disease**.**

An optimised random forest classifier is necessary for the following reasons:

4.1 Fetching of Data: Fetching, originally, is the procedure of extracting data from a database and making it accessible to the application. Fetching is the retrieval of data by a hardware device, script, or software program. After being retrieved, the data is displayed on screen or moved to an alternate location.

4.2 Wrangling: Wrangling is basically handling the missing data. It is the procedure of converting raw data into different formats with the need to make it more relevant and valuable for analytics.

4.3 Analysis of data and modelling: Analysis of data is a process of cleansing, converting, inspecting, and modelling data with the aim of finding out useful, informing conclusions, supporting decision-making and information. Modelling is the process of training and identifying certain types of data. It is basically training the algorithms of machine learning to use the labels from the features.

4.4 Appraisal: It is the process under which the process is undertaken to evaluate the performance of the used algorithms. The input of the algorithms to the output is pre-defined performance goals. The essence of building this system is to presage the risk of future heart disease. The system has been trained using different feature selection algorithms, such as backward elimination, recursive feature elimination, and logistic regression as a machine-learning algorithm.

Table II. Attributes and Its Description from UCI Heart Disease Dataset.

Attributes	Variable	Type
Age	continuous	int
Sex	categorial	int
Cp	categorial	int
Trestbps	continuous	float
Chol	continuous	float
Fbs	categorial	int
Restecg	categorial	int
Thalach	continuous	float
Exang	categorial	int
Oldpeak	continuous	float
Slope	categorial	int
Ca	continuous	int
thal	categorial	int
target	categorial	int

	Age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63.0	1.0	1.0	145	233	1	2	150	0	2.3	3	0.0	6.0	0
1	67.0	1.0	4.0	160	286	0	2	108	1	1.5	2	3.0	3.0	2
2	67.0	1.0	4.0	120	229	0	2	129	1	2.6	2	2.0	7.0	1
3	37.0	1.0	3.0	130	250	0	0	187	0	3.5	3	0.0	3.0	0
4	41.0	0.0	2.0	130	204	0	2	172	0	1.4	1	0.0	3.0	0

Logistic Regression: Among the supervised classification algorithms is logistic regression. The predictive analysis process known as linear regression relies on the concept of probability. The underlying logistic function is used to calculate the probabilities and apply them to the dependent variable in order to weigh the affinity between it and one or more independent variables. Here, the risk factors are the independent variables, the dependent variable is ten-year CHD, and the logistic function is a sigmoid function. The logistic regression hypothesis has been constrained to the range of 0 and 1, or 0 h (x) 1, by using the sigmoid function as a cost function.

The cost function in the logic of regression is characterised as:

$$“\mathcal{C}(\mathrm{h}\theta(\mathrm{x}), \mathrm{y}) = \; -\log(h\theta(x)) \; if \; y = 1$$

$$-\log(1\text{-}h\theta(x)) \; if \; y = 0”$$

The proper presentation of data is crucial for logistic regression. Therefore, key features from the given data set are selected utilising both the backward elimination and recursive elimination strategies in order to create an extra-capable model.

Convolutional Neural Network Algorithm: CNN uses the spatial correlations present in the input data. Every layer of the neural network concurrently connects to some input neurons. “Four different types of layers make up a convolutional neural network: the convolutional layer, the pooling layer, the ReLU correction layer, and the fully connected layer”.

“F1 Measure = 2 (Precision *Recall/Precision + Recall)”

The Naive Bayes Algorithm: The Naive Bayes classifier operates according to the Bayes theorem's definition of conditional probability. The best algorithm was discovered to be Nave Bayes, which was then followed by neural networks and decision trees. Additionally used for disease prediction are artificial neural networks. Diagnoses have been made using supervised networks, which may be trained using the Back Propagation Algorithm.

“P(X⁄Y) = P(Y⁄X) × P(X) / P(Y)”

5. Dataset and Experiments

5.1 Dataset: It gives us the particular details which have 4000 documentation and 14 features/attributes. These features/attributes contain sex, age, resting blood pressure, chest pain type, exercise induced angina, maximum heart rate, sugar blood, electrocardiographic results, ST depression induced by exercise, number of major vessels, slope of the peak exercise and the target classifying from 0 to 2, where 0 indicates the absence of the heart disease. The provided dataset is in the form of CSV format.

5.2. Data Preparation: The dataset includes 644 observations at risk for heart disease out of a total of 4240 observations, 388 of which have missing data. For data preparation, two unique experiments were completed. In order to check it for the first time, we declined the lost data, leaving only 3751 data and 572 observations at risk for heart disease.

Attributes	Description
age	age in years
sex	1 = male, 0 = female
cp	chest pain type: 1: typical angina, 2: atypical angina, 3: non-anginal pain, 4: asymptomatic
treetops	resting blood pressure in mm Hg
choli	serum cholesterol in mg/dl
fbs	fasting blood sugar > 120 mg/dl: 1 = true, 0 = false
restecg	0: normal, 1: having ST-T wave abnormality, 2: left ventricular hypertrophy
thalach	maximum heart rate achieved
Exang	exercise induced angina (1 = yes; 0 = no)
Oldpeak	ST depression induced by exercise relative to rest
Slope	the slope of the peak exercise ST segment: 1: upsloping, 2: flat, 3: downsloping
Ca	number of major vessels: (0-3) colored by flourosopy
Thal	3: normal, 6: fixed defect, 7: reversable defect
Target	diagnosis of heart disease: (0 = false, 1 = true)

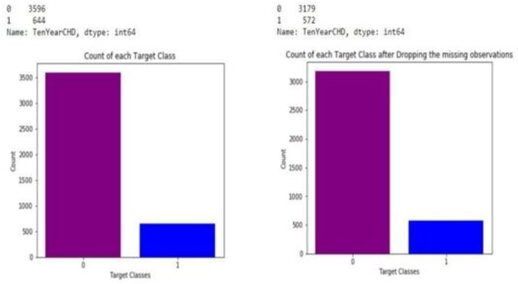

Fig 4: (A) Bar Graph of the Classes Before Dropping.
(B): Bar Graph of the Classes After dropping

Which results in the reduced number of the observations permitting the irrelevant training to this model. So, with the mean value of the observation we progressed with imputation of data and then scaling them using the Simplest Imputer and the Standard Scaler modules od Sklearn.

	male	age	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heart
0	1.153113	-1.234283	-0.988276	-0.758062	-1.758000e-01	-0.077014	-0.671241	-0.162437	-0.940825	-1.196267	-1.083027	0.287258	0.34
1	-0.867217	-0.417664	-0.988276	-0.758062	-1.758000e-01	-0.077014	-0.671241	-0.162437	0.300085	-0.515399	-0.159355	0.719668	1.59
2	1.153113	-0.184345	1.011863	0.925410	-1.758000e-01	-0.077014	-0.671241	-0.162437	0.187275	-0.220356	-0.243325	-0.113213	-0.07
3	-0.867217	1.332233	1.011863	1.767146	-1.758000e-01	-0.077014	1.489778	-0.162437	-0.263965	0.800946	1.016227	0.682815	-0.90
4	-0.867217	-0.417664	1.011863	1.177931	-1.758000e-01	-0.077014	-0.671241	-0.162437	1.089756	-0.106878	0.092555	-0.663554	0.75
...	...	...	...	...	...	...	...	...	...	...	...	...	
4235	-0.867217	-0.184345	1.011863	0.925410	2.069493e-17	-0.077014	-0.671241	-0.162437	0.254961	-0.061487	-0.915087	-0.933810	0.67
4236	-0.867217	-0.650984	1.011863	0.504542	-1.758000e-01	-0.077014	-0.671241	-0.162437	-0.602395	-0.265747	0.344466	-1.631584	0.84
4237	-0.867217	0.282295	-0.988276	-0.758062	-1.758000e-01	-0.077014	-0.671241	-0.162437	0.728764	0.051991	0.008585	-1.064025	0.34
4238	1.153113	-1.117623	-0.988276	-0.758062	-1.758000e-01	-0.077014	1.489778	-0.162437	-1.166445	0.392425	1.268138	-0.049334	-0.73
4239	-0.867217	-1.234283	1.011863	1.767146	-1.758000e-01	-0.077014	-0.671241	-0.162437	-0.918263	0.029296	0.260496	-1.201610	0.75

4240 rows × 14 columns

Fig. 5: Dataset after Scaling and Imputing.

5.3 "Exploratory analysis: Correlation Matrix visualization Before Feature Selection shows"

Fig. 6: Correlation Matrix Visualization.

Thus, it displays that there is no single feature which has a much higher correlation with our target value. Even here few of the features which are having the negative correlation with the target value and some are having the positive correlation. Also, the data is visualized through plots and then the bar graph.

5.4 Feature Selection: Feature Selection using the above algorithms: Further for selecting the most relevant features the data will be passed through the function that gives the following result:

Logit Regression Results

Dep. Variable:	TenYearCHD	No. Observations:	4240
Model:	Logit	Df Residuals:	4234
Method:	MLE	Df Model:	5
Date:	Mon, 09 Mar 2020	Pseudo R-squ.:	-0.5700
Time:	08:42:03	Log-Likelihood:	-2835.5
converged:	True	LL-Null:	-1806.1
Covariance Type:	nonrobust	LLR p-value:	1.000

	coef	std err	z	P>|z|	[0.025	0.975]
male	0.1053	0.033	3.178	0.001	0.040	0.170
age	0.2626	0.035	7.505	0.000	0.194	0.331
cigsPerDay	0.1294	0.034	3.812	0.000	0.063	0.196
prevalentStroke	0.0813	0.038	2.124	0.034	0.006	0.156
diabetes	0.1055	0.035	3.046	0.002	0.038	0.173
sysBP	0.2244	0.035	6.370	0.000	0.155	0.293

Fig 7: Result from Feature Selection using Backward Elimination Method According the result above the columns were dropped.

	male	age	cigsPerDay	prevalentStroke	diabetes	sysBP
0	1.153113	-1.234283	-0.758062	-0.077014	-0.162437	-1.196267
1	-0.867217	-0.417664	-0.758062	-0.077014	-0.162437	-0.515399
2	1.153113	-0.184345	0.925410	-0.077014	-0.162437	-0.220356
3	-0.867217	1.332233	1.767146	-0.077014	-0.162437	0.800946
4	-0.867217	-0.417664	1.177931	-0.077014	-0.162437	-0.106878
...	...	...	...	...	...	...
4235	-0.867217	-0.184345	0.925410	-0.077014	-0.162437	-0.061487
4236	-0.867217	-0.650984	0.504542	-0.077014	-0.162437	-0.265747
4237	-0.867217	0.282295	-0.758062	-0.077014	-0.162437	0.051991
4238	1.153113	-1.117623	-0.758062	-0.077014	-0.162437	0.392425
4239	-0.867217	-1.234283	1.767146	-0.077014	-0.162437	0.029296

4240 rows × 6 columns

Fig 8: Dataset After Dropping Columns after Feature Selection

Feature Selection using "Recursive Feature Elimination and Cross-Validated selection method":

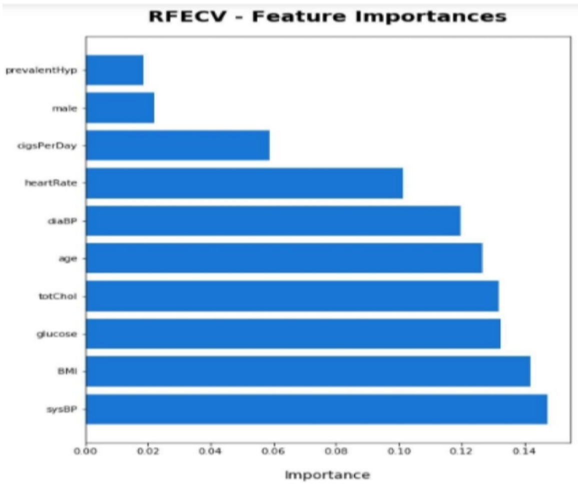

Fig 9: Top 10 important features supported by RFECV.

5.5 Training and testing: Ultimately the resulting data split into 80% train and 20% test data, which will be further proceed to the Logistic Regression model to fit, predict and score the model.

6. Results and Discussion

Backward eradication proved to be the method that produced the greatest outcomes for us when comparing several methods for feature selection and testing. Backward Elimination with and without KFold, as well as Recursive Feature Elimination with Cross Validation, were the different techniques that were tested. The accuracy of these techniques ranged between 92% and 93.22%, with the highest value being 93.22%.

Even so, the accuracy of both techniques was comparable, but we observed that Backward Elimination had a higher rate of True Negative misclassifications and exhibited greater accuracy dissimilarity compared to RFEV. Backward Elimination has an accuracy of 84%, and RFEV has an accuracy of 86%. Furthermore, there have been 0.99 and 1 recalls, respectively. There are less misclassifications in RFECV than in Backward Elimination, according to precision and recall measurements.

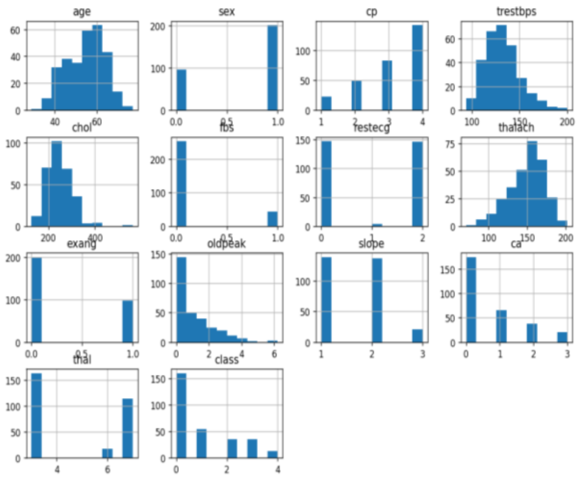

Fig 10: Histogram generated for each attributes by CNN.

Observations: From the above Fig 10. The following observation is drawn.

- Since males are more likely to develop heart disease than women, sex has a significant impact on the target variable.
- Those who reported having usual or asymptomatic chest pain appear to have the highest risk of developing heart disease.
- Fasting blood sugar (fbs) levels greater or lower than 120 mg/dl cannot accurately predict the target.
- A normal appears to have the greatest influence on the target according to the results of the resting electrocardiography (resecg).
- The aim is impacted quite strongly by the presence of exang (exercise-induced angina).
- Where responders have an upward-sloping peak exercise slope, the likelihood of having a cardiac condition is significantly lower.
- The target variable may be well predicted by the variable thal.

Fig 11. dipicts ROC curve for multi-class classifier illustrate the diagnostic ablitiy of the classifiers used in the system.

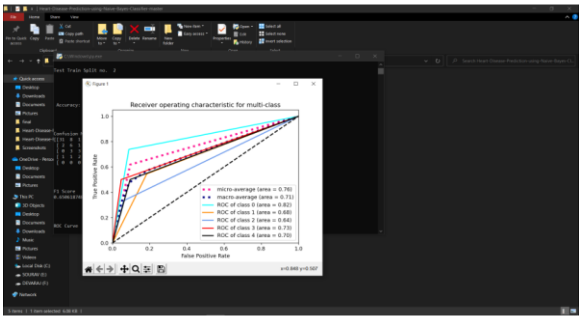

Fig 11: Receiver operating characteristic for multiclass classifier.

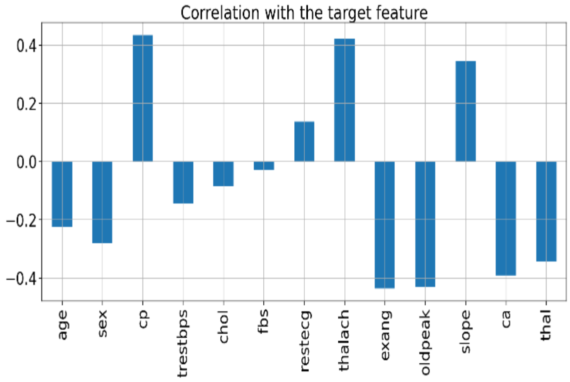

Fig 12: Graph: Correlation of attributes with the target feature.

In this project we have used three algorithms namely, Logistic regression, Convolutional Neural Network and Naive Bayes to predict the presence of coronary artery disease. We also have calculated performance metrics like precision, recall and F1-Score for each algorithm.

“Precision = (TP) / (TP +FP)”

“Recall = (TP) / (TP+FN)”

“F1-Score = 2(Precision * Recall) / (Precision +Recall)”

- “True positive: the test is positive and the patient has the condition”.
- “FP False positives occur when a test is positive although the patient does not have the condition”.
- “TN True negative means that both the patient's test results and the disease are negative”.
- “FN False Negative (test result is negative despite the patient having the illness)”.

The resulting data pre-processed and split into 80% train and 20% test data. The Logistic Regression, CNN and Naïve Bayesian algorithms are applied. Utilizing the confusion matrix, the performance metrics indicated above were determined. The accuracy score obtained for Logistic Regression, CNN and Naive Bayes classification techniques is shown below in Table 3.

Table III. Comparison of Machine Learning algorithms for the Prediction Accuracy

Sl. No.	Algorithm	Precision	Recall	F1 score	Accuracy
1.	**Logistic Regression**	**0.85**	**0.86**	**0.85**	**93.22 %**
2.	CNN	0.86	0.85	0.79	78 %
3.	Naïve Bayes	0.66	0.70	0.67	64.47 %

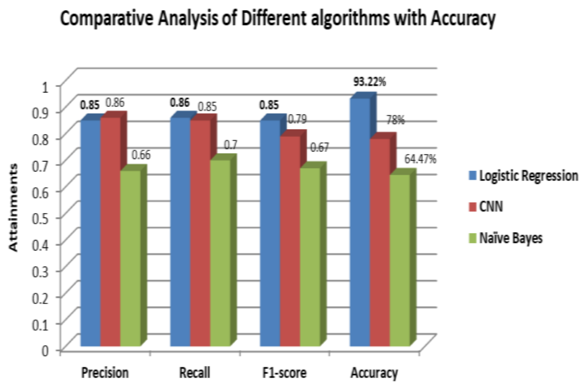

Fig 13. Graph for comparative analysis of different algorithms and its accuracy.

7. Conclusion

Coronary Artery Disease, which is also known as CAD, is the most damaging and death-causing chronic disease that quickly accelerates in both economically undeveloped and developed countries. This study shows that with high accuracy, machine learning algorithms can be implemented and help in predicting CAD using publicly available data sets. It uses the existing dataset from the Kaggle database-UCI heart disease database. Three machine learning classifiers have been implemented, namely: Logistic Regression, Naive Bayesian and CNN. The Logistic Regression algorithm shows a better accuracy of 93.22% in comparison with CNN and the Nave Bayesian algorithm

References

[1] M. R. Ahmed, S. M. Hasan Mahmud, M. A. Hossin, H. Jahan and S. R. Haider Noori, "A Cloud Based Four-Tier 253 Architecture for Early Detection of Heart Disease with Machine Learning Algorithms," 2018 IEEE 4th International 254 Conference on Computer and Communications (ICCC), 2018, pp. 1951-1955, doi: 10.1109/CompComm.2018.8781022. 255

[2] Asma Baccouche , Begonya Garcia-Zapirain, Cristian Castillo Olea, Adel Elmaghraby, "Ensemble Deep Learning Models 256 for Heart Disease Classification: A Case Study from Mexico", Information 2020, 11(4), 207; 257 https://doi.org/10.3390/info11040207. 258

[3] Y. Peili, Y. Xuezhen, Y. Jian, Y. Lingfeng, Z. Hui and L. Jimin, "Deep learning model management for coronary heart 259 disease early warning research," 2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis 260 (ICCCBDA), 2018, pp. 552-557, doi: 10.1109/ICCCBDA.2018.8386577. 261

[4] A. Gavhane, G. Kokkula, I. Pandya and K. Devadkar, "Prediction of Heart Disease Using Machine Learning," 2018 262 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2018, pp. 263 1275-1278, doi: 10.1109/ICECA.2018.8474922. 264

[5] Md. Ashraful Alam, Mohsinul Bari Shakir, Monirul Islam Pavel, "Early detection of coronary artery blockage using 265 image processing: segmentation, quantification, identification of degree of blockage and risk factors of heart attack" 266 (Conference Presentation) May 2019 , DOI:10.1117/12.2517452, Conference: Micro- and Nanotechnology Sensors, 267 Systems, and Applications XIAt: Baltimore, Maryland, USA. 268

[6] Vardhan Shorewala, Early detection of coronary heart disease using ensemble techniques, Informatics in Medicine 269 Unlocked, Volume 26, 2021, 100655, ISSN 2352-9148, https://doi.org/10.1016/j.imu.2021.100655. 270

[7] Ed-daoudy, A., Maalmi, K. A new Internet of Things architecture for real-time prediction of various diseases using 271 machine learning on big data environment. J Big Data 6, 104 (2019). https://doi.org/10.1186/s40537-019-0271-7 272

[8] Aravind Akella, Sudheer Akella, "Machine learning algorithms for predicting coronary artery disease: efforts toward an 273 open source solution", FUTURE SCIENCE, VOL. 7, NO. 6., 29 Mar 2021https://doi.org/10.2144/fsoa-2020-0206. 274

[9] Yar Muhammad, Muhammad Tahir, Maqsood Hayat & Kil To Chong, "Early and accurate detection and diagnosis of 275 heart disease using intelligent computational model", Scientific Reports volume 10, Article number: 19747 (2020). 276

[10] Chen, Xueping & Fu, Yi & Lin, Jiangguo & Ji, Yanru & Fang, Ying & Wu, Jianhua. (2020). Coronary Artery Disease 277 Detection by Machine Learning with Coronary Bifurcation Features. Applied Sciences. 10. 7656. 10.3390/app10217656. 278.

[11] Shah, D., Patel, S. & Bharti, S.K. Heart Disease Prediction using Machine Learning Techniques. SN COMPUT. SCI. 1, 345 (2020). https://doi.org/10.1007/s42979-020-00365.

[12] A. Singh and R. Kumar, "Heart Disease Prediction Using Machine Learning Algorithms," 2020 International Conference on Electrical and Electronics Engineering (ICE3), 2020, pp. 452-457, doi: 10.1109/ICE348803.2020.9122958.

[13] Md Mamun Ali, Bikash Kumar Paul, Kawsar Ahmed, Francis M. Bui, Julian M.W. Quinn, Mohammad Ali Moni, Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison, Computers in Biology and Medicine, Volume 136, 2021, 104672, ISSN 0010-4825, https://doi.org/10.1016/j.compbiomed.2021.104672.

[14] https://thenationalpilot.ng/2019/01/21/silent-ischemia-heart-disease. 279

[15] https://www.kaggle.com/datasets/5ff0a500aa39c3c772e989cddb13bc693039f062affc01e6447599368944b7f6.280