Coronary Artery Disease (CAD) Prediction using Machine Learning Method with Ensemble Technique

Waqas Mehmood* Abdul Khaliq†

***Abstract*—Coronary Artery Disease (CAD) prediction is a crucial and challenging task in the medical field. CAD if detected at later stages can affect heart and blood vessels which can lead to blockage in veins and lead to a life threatening condition. Our research paper encompasses the different classifications algorithms and final results are generated using ensemble technique which helped in generating perfect diagnosis of CAD. Ensemble methods including Logistic Regression, Naive Bayes Classifier, K-Nearest Neighbor (KNN), Decision Tree Classifier, Support Vector Machine and then Multi-model training was implemented to create a predictive model. Use of Machine Learning in CAD related risk assessment enables early detection and prevention life threatening impact of CAD.**

I. Introduction

CAD is a health challenge at the global level. Disease of heart is jam or block your coronary clotting supply paths usually carried out by the development of greasy material called plaque. Coronary vein infection is additionally called coronary illness, ischemic coronary illness, and coronary disease. CAD ensues when plaque builds up in a patient's arteries as plaque continues to build [1]. ML predictive analysis is considered to be the best approach to prevent severe complications affecting both the heart and blood vessel. It can help in timely and precise identification of CAD which reduces risk of life threatening conditions due to heart disease. This research focuses on the application of machine learning classification algorithms, specifically use of ensemble techniques to predict CAD. Ensemble techniques involve combining multiple ML algorithms to enhance overall predictive performance. This research explains the effectiveness of ensemble methods including Random Forests, Gradient Boosting and Multi-model training in constructing a robust predictive model for CAD. The primary purpose of this research is to assess the efficacy of ensemble learning in CAD prediction, which can help in early detection and accurate diagnosis. By incorporating multiple classification algorithms within the ensemble framework, we aim to enhance the model's sensitivity and specificity, leading to a more precise CAD prediction tool. We used Cardiovascular Disease dataset containing 1024 patients' record to train our model. The methodology section details data sources, preprocessing steps, and the implementation of ensemble techniques. The results and discussions focus on the performance of the proposed CAD prediction model, highlighting its potential impact on early detection and risk assessment. Our proposed algorithm is used for early prediction of CAD with a higher accuracy level.

Capital University of Science and Technology, waaqasmehmood@gmail.com

Capital University of Science and Technology, engg.abdulkhaliq@gmail.com

II. Literature Review

Extensive work has already been done in this field yet there is gap for improvement. In previous studies, result accuracy on test data has never reached more than 88%. Considering this a critical decease, accuracy must be increased. Table below contains the comparative analysis of accuracy result generated through previous work.

TABLE I: Summary of Papers

Year	Reference	Title of Paper	Classifier	Accuracy
2018	[7]	Estimate of HD using ensemble learning and Subdivision Swarm Optimization	Naïve Bayes, K-nearest neighbor	82.6%
2019	[3]	ML-based coronary artery disease diagnosis: An inclusive review	Ensemble model	86.93%
2019	[1]	HD prediction using machine learning analytical approach and Random forest algorithm	Random forest model	76%
2020	[5]	A data-driven approach for predicting the heart disease by using logistic regression	Logistic regression	83.12%
2020	[4]	A innovative intelligence system based on machine learning system for coronary heart disease prediction	XGBOOST	82%
2021	[6]	Design and Application of CAD prediction using NB	Naïve Bayes	79%
2022	[2]	CAD calculation using supervised type ML algorithm	RF, LG, SVM, KNN	88% and 85%

All the results mentioned in Table I, research work have used the Cleveland dataset in their studies. We also used it with another type of dataset obtained from the UCI dataset repository, having 14 attributes with 1024 different patient's record Some of above mentioned papers used ensemble technique which improved the overall accuracy but still there is gap for improvement in accuracy. The method we used to improve the overall accuracy is to use ensemble of algorithms generating satisfactory results on test data and ignore the algorithms for which accuracy and confusion matrix of test data is not good.

III. Methods

In machine learning, we can use different classifiers or their combinations to help us perform classification on given datasets. Here in our work, we want to predict the number of patients with or without heart disease by training our model on given dataset by running eight algorithms on our data set. The reason we use eight algorithms is that it will allow us to get better and more reliable predictions. We ensemble 5 classifier and performed hard voting classifier and then we ensemble 3 classifier with hard voting as shown in Fig 1.

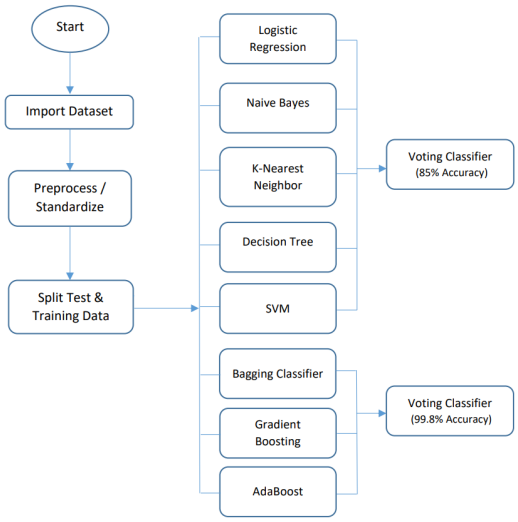

Fig. 1: Proposed Methodology

This dataset is obtained from UCI repository having 14 attributes with 1024 different patient's record are shown in Table II.

IV. Results

After applying all the machine learning classifier on the dataset we get this results shown in the Table III. Initially we used the Logistic Regression , Naive

TABLE II: Attribute Description

No.	Attribute	Description	Values
1	age	Age of patient	Age numeric values
2	gender	Gender	1-Male, 0-Female
3	C_P	Chest-pain	1 = typ-ang, 2 = aty-pAng, 3 = non-ang
4	trest_bps	Blood-pressure in rest	BP values
5	CHOL	Dietary fat	Cholesterol values
6	F-B-S	Sugar in blood	1 = present, 0 = Absent
7	restecg	Echocardiography at rest	1=Abnormal-ECG, 0=Normal
8	oldpeak	Exercise related to rest	Different values
9	slope	ST depression slope	0, 1, 2 represent different slopes
10	CA	Vessels	0, 1, 2, 3 values represent how your arteries are affected
11	thal	Thalassemia	0, 1, 2, 3 represent colored Fluoroscopy vessels
12	thalach	Patient maximum heart rate	Heart rate values
13	exang	Angina with exercise	1=present, 0=absent
14	target	Heart Disease	1=effected person, 0=healthy

Bayes , KNN , DT and SVM classifiers to predict the Disease and there individual results are given in Table III. After that we combined these classifiers using hard voting classifier and predictions were not improved, So then we used Adaboost, Bagging and Gradient Boost ensemble techniques using Decision Trees and by using this method, the predicted results were significantly improved which are given in Table III. In our proposed methodology, we combined these ensemble classifiers using hard voting classifier and predicted results were perfect on test data. Results comparison is shown in Fig 2 in form of Accuracy, Precision, Recall and F1 Score. Results shows that VC-2 is giving the best result.

V. Conclusion

In this research study, we have utilized eight machine learning algorithms on a dataset from the UCI repository, without implementing any feature

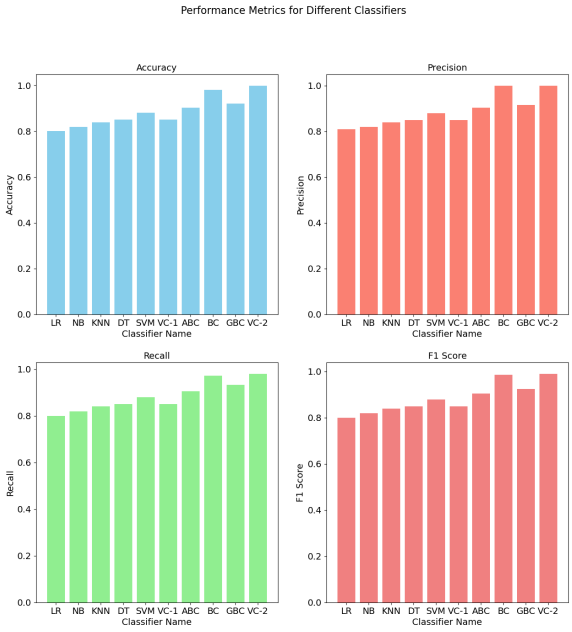

Fig. 2: Performace Comparison

TABLE III: Machine Learning Algorithms Results on Test Data

Classifier Name	Accuracy	Precision	Recall	F1 score
Logistic Regression	0.80	0.81	0.80	0.80
Naive Bayes Classifier	0.8195	0.82	0.82	0.82
K-Nearest Neighbor	0.8390	0.84	0.84	0.84
Decision Tree Classifier	0.85	0.85	0.85	0.85
Support Vector Machine	0.88	0.88	0.88	0.88
Voting Classifier with LR, NB, KNN, DT, SVM	0.85	0.85	0.85	0.85
AdaBoost Classifier	0.9024	0.9047	0.9047	0.9047
Bagging Classifier	0.98	1	0.9714	0.9855
Gradient Boosting	0.92	0.9158	0.9333	0.9247
Voting Classifier with AdaBoost, Bagging, GB	0.9987	1	0.98	0.99
Unseen Instance Test :	Correctly Classified			

selection technique. Our research was conducted using the Python programming language with Jupiter Notebook. Initially, we applied our proposed classifiers to the datasets and calculated various metrics including accuracy, precision, recall, and F1 score. Subsequently, we combined five classifiers, namely Logistic Regression, Naive Bayes Classifier, K-Nearest Neighbor, Decision Tree Classifier and Support Vector Machine to evaluate the accuracies and compare the results with our ensemble model. Performance of model after ensemble was not significantly improved. After that we applied Random Forest, AdaBoost, Bagging and Gradient Boosting classifiers and ensemble them which significantly improved the results and accuracy on test data reached to 100%.

Bagging and Gradient Boosting classifiers and ensemble them which significantly improved the results and accuracy on test data reached to 100%. Bagging and Gradient Boosting classifiers and ensemble them which significantly improved the results and accuracy on test data reached to 100%. Bagging and Gradient Boosting classifiers and ensemble them which significantly improved the results and accuracy on test data reached to 100%. Bagging and Gradient Boosting classifiers and ensemble them which significantly improved the results and accuracy on test data reached to 100%.

REFERENCES

[1] Moloud Abdar, Elham Nasarian, Xujuan Zhou, Ghazal Bargshady, Vivi Nur Wijayaningrum, and Sadiq Hussain. Performance improvement of decision trees for diagnosis of coronary artery disease using multi filtering approach. In *2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*, pages 26–30. IEEE, 2019.

[2] Ghulab Nabi Ahmad, Shafi Ullah, Abdullah Algethami, Hira Fatima, and Syed Md Humayun Akhter. Comparative study of optimum medical diagnosis of human heart disease using machine learning technique with and without sequential feature selection. *ieee access*, 10:23808–23828, 2022.

[3] Roohallah Alizadehsani, Moloud Abdar, Mohamad Roshanzamir, Abbas Khosravi, Parham M Kebria, Fahime Khozeimeh, Saeid Nahavandi, Nizal Sarrafzadegan, and U Rajendra Acharya. Machine learning-based coronary artery disease diagnosis: A comprehensive review. *Computers in biology and medicine*, 111:103346, 2019.

[4] Wenbing Chang, Yinglai Liu, Xueyi Wu, Yiyong Xiao, Shenghan Zhou, and Wen Cao. A new hybrid xgbsvm model: application for hypertensive heart disease. *Ieee Access*, 7:175248–175258, 2019.

[5] Ibomoiye Domor Mienye, Yanxia Sun, and Zenghui Wang. An improved ensemble learning approach for the prediction of heart disease risk. *Informatics in Medicine Unlocked*, 20:100402, 2020.

[6] Vardhan Shorewala. Early detection of coronary heart disease using ensemble techniques. *Informatics in Medicine Unlocked*, 26:100655, 2021.

[7] Indu Yekkala, Sunanda Dixit, and MA Jabbar. Prediction of heart disease using ensemble learning and particle swarm optimization. In *2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon)*, pages 691–698. IEEE, 2017.