Coronary Artery Disease (CAD) Prediction using Machine Learning Method with ensemble technique

**Abstract-** Coronary Artery Disease (CAD) prediction is a crucial & 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.

**Keywords:** Machine Learning model, Heart disease, CAD, Heart failure, Machine Learning classifiers, Ensemble model, Decision Tree, Logistic regression, KNN, Naive byes

1. **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.

1. **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 1. Review of most research related articles**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Year | Reference | Title of Paper | Classifier | Accuracy |
| 2018 | [2] | 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: A inclusive review | Ensemble model | 86.93% |
| 2019 | [4] | 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 | [6] | A innovative intelligence system based on machine learning system for coronary heart disease prediction | XGBOOST | 82% |
| 2021 | [7] | Design and Application of CAD prediction using NB | Naïve Bayes | 79% |
| 2022 | [8] | CAD calculation using supervised type ML algorithm | RF, LG, SVM, KNN gives below then 88 % accuracy but Naïve Bayes Gives 85% accuracy | 88% and 85% |

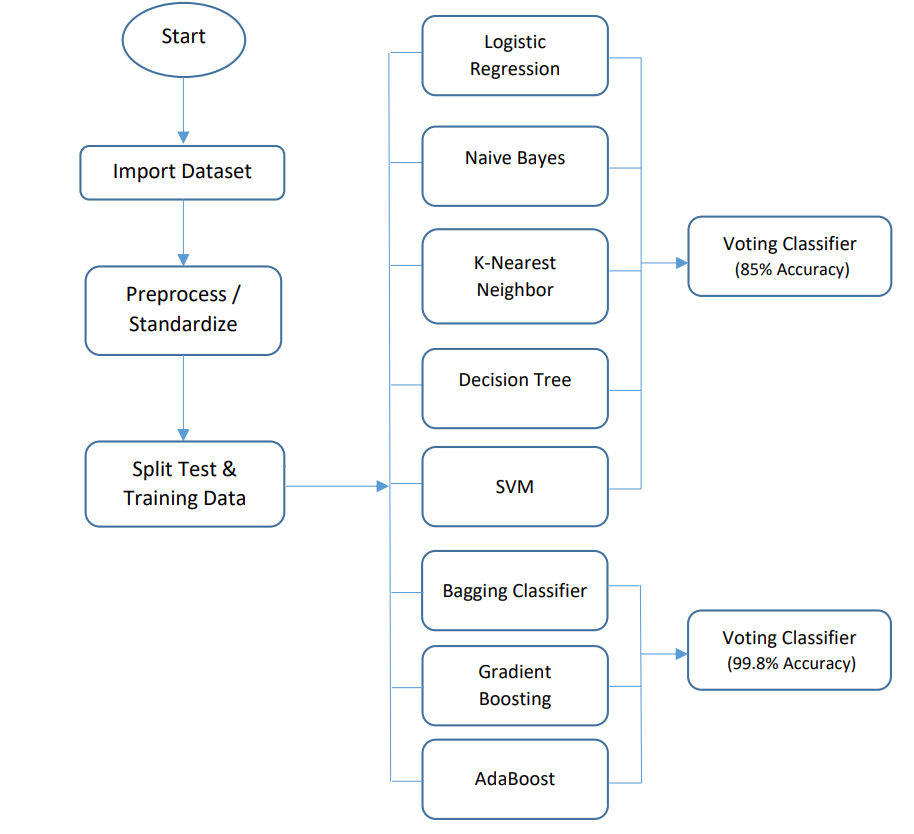
All of the above mentioned research work used have used the Cleveland dataset in their studies. We also used it with another type of dataset obtained from the UCI dataset repository, having 14 attributes with 1024 different patient’s record

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.

1. **METHODOLOGY**

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.

If we use more than one algorithm or classifier in our case eight of them we can compare them with each other and if we find that one classifier provides gives us precision far from another algorithm provided we can figure out that something is wrong. Maybe the algorithm itself isn't right for the job, or we need to change parameters to improve classification results. Therefore, the use of some algorithm is essential for any prediction-based system. The proposed methodology is shown in the Fig 1.



***Fig. 1 Research Methodology***

**Attributes Description**

This dataset is obtained from UCI repository having 14 attributes with 1024 different patient’s record are shown in table 2.

**Table 2. 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 = atypAng, 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 effected |
| 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 |

**Machine Learning Algorithms Results**

After applying all the machine learning classifier on the dataset we get this results shown in the table 3.

***Table 3. Machine Learning Algorithms Results***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **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** | | | |

1. **CONCLUSION**

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

[1] <https://vfast.org/journals/index.php/VTSE/article/view/1487/1243>

[2] I. Yekkala, S. Dixit, and M. A. Jabbar, “Prediction of heart disease using ensemble learning and Particle Swarm Optimization,” Proc. 2017 Int. Conf. Smart Technol. Smart Nation, SmartTechCon 2017, pp. 691–698, 2018, doi: 10.1109/SmartTechCon.2017.8358460. [6]

[3] R. Alizadehsani et al., “Machine learning-based coronary artery disease diagnosis: A comprehensive review,” Comput. Biol. Med., vol. 111, p. 103346, 2019, doi: 10.1016/j.compbiomed.2019.103346 [15]

[4] M. Abdar, E. Nasarian, X. Zhou, G.Bargshady, V. N. Wijayaningrum, and S. Hussain, “Performance improvement of decision trees for diagnosis of coronary artery disease using multi filtering approach,” 2019 IEEE Int. journel Comput. Commun. Syst. ICCCS 2019, no. Dm, pp. 26–30, 2019, doi: 10.1109/CCOMS.2019.8821633 [14]

[5] I. D. Mienye, Y. Sun, and Z. Wang, “An improved ensemble learning approach for the prediction of heart disease risk,” Informatics Med. Unlocked, vol. 20, p. 100402, 2020, doi: 10.1016/j.imu.2020.100402. [8]

[6] W. Chang, Y. Liu, X. Wu, Y. Xiao, S. Zhou, and W. Cao, “A New Hybrid XGBSVM Model: Application for Hypertensive Heart Disease,” IEEE Access, vol. 7, pp. 175248–175258, 2019, doi: 10.1109/ACCESS.2019.2957367. [16]

[7] V. Shorewala, “Early detection of coronary heart disease using ensemble techniques,” Informatics Med. Unlocked, vol. 26, no. July, p. 100655, 2021, doi: 10.1016/j.imu.2021.100655. [13]

[8] . G. N. Ahmad, S. Ullah, A. Algethami, H. Fatima, and S. M. H. Akhter, “Comparative Study of Optimum Medical Diagnosis of Human Heart Disease Using Machine Learning Technique with and Without Sequential Feature Selection,” IEEE Access, vol. 10, pp. 23808– 23828, 2022, doi: 10.1109/ACCESS.2022.3153047. [18]