Combined Tree-Based Ensemble Learning for Predicting Coronary Artery Disease

Arun Kumar M
(24MDT1131)
Vellore institute of technology, Chennai, arunkumar.m2024a@vitstudent.ac.in

Abstract **- The study developed a predictive model for angiographic Coronary Artery Disease(CAD) on the enormous Coronary Artery Disease data set. Tree-based ensemble learning methods Random Forest, Gradient Boosting Machine (GBM), Extreme Gradient Boosting Machine (XGBM), Light Gradient Boosting Machine (LGBM), AdaBoost, and Cat boost are compared together in this paper by using repeated random train test split that allows consistency in the data set. Additionally, I created a meta-model by combining these tree-based models using the stacking method and developed a web application using Stream-lit. This data set from IEEE Data Port contains all clinical. This can help clinicians to rightly detect CAD. It may prove a potentially useful alternative methods of diagnosis for clinical purposes.**

Keywords **- Coronary Artery Disease (CAD), Machine Learning (ML), Predictive model, Gradient Boosting Machine (GBM), Extreme Gradient Boosting Machine (XGBM), Light Gradient Boosting Machine (LGBM), AdaBoost, Cat-boost**

1. Introduction

CAD is a disease that affects a large number of the population throughout the world, mainly in industrialized countries. According to the reports from the World Health Organization, cardiovascular diseases are responsible for more than 31% of the mortality in the world. Traditional diagnostic methods such as angiography are very expensive, which creates an increasing demand for inexpensive diagnostic methods.

Recent machine learning developments have shown great promise in building predictive models using clinical data for CAD detection. In this work, tree-based ensemble learning were deployed for enhancing the accuracy and reliability of CAD predictions because of small data-set. Several methods from this family have been covered within the scope, namely, Random Forest, GBM, XGBM, LightGBM, Cat-boost, and XGBoost. Further, I developed a meta-model by applying all these tree-based methods in a stacking approach and then finally a web application in Stream-lit for ease of use by clinicians. Our results demonstrate that these methods can significantly aid the clinician in correctly diagnosing CAD and may also supersede the conventional old classical diagnostic methods.

2. Literature Review

2.1 In 1993, Wilson, Peter WF, and Jane C. Evans published their study titled "Coronary artery disease prediction" in American Journal of Hypertension. Coronary artery disease continues to be one of the major health problems; it is among the leading causes of death in the United States. Several studies, such as the Framingham Heart Study, have identified many risk factors for CAD that can be used in its prediction and prevention. The cohort initiated the Framingham Heart Study in 1948 comprising of 5,209 men and women aged between 30 and 62 years. Every two years, all individuals in the original cohort have been interviewed since the start of the study, while the Offspring cohort, comprising 5,135 subjects related to the original participants, has been followed up approximately every four years. The baseline for CAD estimation was made during the eleventh biennial examination of the original cohort and the first examination of the Offspring group, between 1968 and 1975. Only individuals free of cardiovascular diseases at baseline were included in analyses. Some of the established key CAD risk factors are: age, gender, cigarette smoking, blood cholesterol levels, high-density lipoprotein (HDL) cholesterol, blood pressure, left ventricular hypertrophy, or LVH, and diabetes mellitus. These risk factors were analyzed in a middle-aged group of 2,590 women and 2,983 men. Results showed that 12-year incidence of CAD increased with age. The Framingham study particularly emphasized the increasing relative risk (RR) associated with risk factors. The relative risk for CAD among those with ECG-LVH was indeed greater than that from smoking, yet the prevalence of ECG-LVH was much lower. While this demonstrates a means to evaluate both relative and absolute effect of risk factors on CAD, it also points out a limitation in their use: namely, that assuming prevalence is uniform across populations, RR directly estimates the increase in rate due to one factor, but indirectly estimates the effect of another owing to its association with another factor through the relative rate.

2.2 In 2004, Mark J. Pletcher, et al. published the systematic review and meta-analysis "Using the coronary artery calcium score to predict coronary heart disease events." The CAC score is an important prognostic measure used for stratifying risk of CHD events, potentially very relevant for primary prevention strategies in high-risk patients. This can measure CAC and is being proposed to extend the predictive space beyond traditional risk factors; however, added predictive value remains under investigation. A total of 13 articles were chosen to be included in the review. Data extraction was conducted by two abs tractors who procured information on study demographics, follow-up methods, CAC measurement protocols, and outcome adjudication. This entailed a very close scrutiny, and hence only studies that met the very strict inclusion and exclusion criteria were included. This improved the reliability of results that might arise from the meta-analysis of selected studies. In the meta-analysis of the selected studies, the CAC score was an independent predictor of the CHD events. A summary adjusted relative risk of 2.1 was associated with scores of 1 to 100, and higher scores presented even more marked risk estimates ranging from 3.0 up to 17.0. The variability in the estimates was based on the differences in methods used, focusing on how the studies were conducted, which includes outcome adjudication and the way risk factors were measured. Literature review underscores the use of the CAC score to predict CHD events in asymptomatic patients. The addition of this measurement will be there in updated guidelines, making careful risk assessment more enhanced to increase the identification of patients at high risk, thus improving preventive strategies in clinical practice.

2.3 In 2004, Caren Marzban, in his work "The ROC Curve and the Area under It as Performance Measures," presented some evidence concerning relationships among ROC curves, AUC, and their own underlying distributions of forecasts. The ROC curve and the area under it are the main means of describing performance regarding the binary classification model. The paper stresses that the skill of forecasts of binary observations can be effectively gauged within the joint probability distribution of forecasts and observations. Indeed, such an approach enables the forecast quality to be comprehensively diagnosed by a number of diagrams, including the so-called ROC curve especially popular in meteorology and medicine. A ROC curve is defined as the parametric plot of the hit rate, or true positive rate, against the false alarm rate, or false positive rate, as the decision threshold varies. The diagonal line of the ROC space shows random forecasts; concavity in the curve is taken to be a measure of performance. The AUC value varies between 0 and 1, where 0.5 means the predictions are random and 1 represents perfect predictions.

2.4 In 2016, Verma Luxmi, Sangeet Srivastava, and P. C. Negi investigated it in the study "A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data". CAD is called coronary artery disease, a significant health concern of the world, which causes several serious consequences such as heart attacks and cardiac arrest. World Health Organization reports that cardiovascular diseases caused 31% of all deaths registered in all regions of the globe in 2012 and CAD significantly contributes to it. For such a paramount and menacing disease, the conventional diagnosis involves invasion-a technique as simple as cutting into the interior of a blood vessel for viewing using angiography or other processes that are expensive and require technical know-how. Therefore, most researchers turned their attention and searches to non-invasive methods with the help of data mining. The article particularly proposes and applies readily available clinical data in predictions of CAD. Data from 335 suspected CAD patients at Indira Gandhi Medical College, Shimla, India, have been taken into consideration for testing the proposed hybrid model. According to the study, diagnoses acquired through non-invasive methods can be reproducible and objective; therefore, they are very important and useful in the clinic.

2.5 In 2017, Praveena, M., and V. Jaiganesh, "A literature review on supervised machine learning algorithms and boosting process," International Journal of Computer Applications, 2016. The field of supervised machine learning has witnessed a tremendous growth in many directions in recent years. Decision trees and Support Vector Machines (SVM) are but good examples. This paper synthesizes some of the recent research findings on both of the above algorithms and discusses how the boosting processes improve their performances. Boosting techniques were combined, including Ada-boost, in order to showcase the significant boost of predictive accuracy of supervised machine learning algorithms. The findings relate the effectiveness of clustering or classification tasks with a direct link to implementing boosting processes and reduce overall computational complexity.

2.6 In 2017, Lin, Weiwei, et al. "An ensemble random forest algorithm for insurance big data analysis" presented the study in IEEE Access. In insurance, one of the main problems is the imbalance in data. The ratio of positive versus negative cases in this industry is huge, sometimes even 100:1. This leads to biased predictions from the classical models, such as logistic regression and SVM, which always lean toward the majority class by achieving high recall but low precision. Recent developments have hence brought about improvements in SMOTE algorithms through the introduction of SMOTE-RSB and Borderline-SMOTE, which also focus on refining the sampling process as a means of achieving better model performance. Those are important for applications in various fields, including earthquake prediction and virus detection, where the classification accuracy is decisive. The effectiveness of these methods of computational intelligence is usually estimated on grounds of the F-Measure metric. It shows that the

ensemble methods, and especially the ensemble random forest algorithm, outperform the traditional classification algorithms on accuracy and performance when applied to the imbalanced insurance data.

2.7 In 2017, Osisanwo, F. Y., et al. published a study titled "Supervised machine learning algorithms: classification and comparison" in the International Journal of Computer Trends and Technology (IJCTT). There are several supervised learning algorithms, most of which are actually applied to the classification task. These include Linear Classifiers, Logistic Regression, Naïve Bayes Classifier, Perceptron, Support Vector Machine (SVM), Decision Trees, Random Forests (RF), and Neural Networks, and so on. Each has its unique characteristics regarding strengths and weaknesses, which leads them to be better suited for certain types of data and applications. The empirical study of this paper indicates that SVM can outperform the other compared algorithms as far as precision and accuracy are concerned. Naïve Bayes and Random Forest also gain a strong performance, although it could not reach the level of precision against SVM. This gives us an idea of considering evaluating an algorithm not only as standalone but also depending on specific datasets because the approach can be subjective to data qualities and characteristics. The literature also discusses the potential issues associated with machine learning classification, such as the fact that parameters may require fine-tuning and sufficient instances in a dataset. The best algorithm for one dataset may not be so good for another, which emphasizes the importance of understanding the underlying attributes of the data.

2.8 In 2018, Naushad, Shaik Mohammad, et al. conducted a study titled "Machine learning algorithm-based risk prediction model of coronary artery disease" published in Molecular Biology Reports, this conducted study included a heterogeneous population. Sourced from 648 study subjects, of whom 364 are CAD cases and 284 are healthy controls, demographic data and conventional risk factors were investigated. Ethical clearance was followed to conduct the study with informed consent from all individuals. The genomic DNA extraction using whole blood samples is an essential step in analyzing genetic polymorphisms. Three risk prediction models were built; these were Ensemble Machine Learning Algorithms (EMLA), Multifactor Dimensionality Reduction (MDR), and Recursive Partitioning (RP). The best results in CAD risk prediction were achieved in EMLA with 89.3%, whereas percentage stenosis reached an impressive 82.5%. The most relevant predictors discovered were hypertension, alcohol intake, and specific polymorphisms in EMLA.

2.9 In 2019, Abdar, Moloud, et al. published research studies is titled "A new machine learning technique for an accurate diagnosis of coronary artery disease" in Computer Methods and Programs in Biomedicine. Prediction of Coronary Heart Disease (CHD) is one of the most discussed topics in the medical and data science communities. Many research efforts have been reported in this area with different approaches to enhance the predictive accuracy with the help of ML techniques. In the previous literature, various algorithms of ML are used to classify CHD risk. Of them, SVM, ANN, and Decision Trees are mentioned. A recent literature reported that the SVM has a high accuracy of 92.1 percent, followed by ANN with 91.0 percent, followed by 89.6 percent of DT when the dataset of 502 instances is used. The studies often involve different datasets; for example, one dataset contains 13 attributes such as sex, blood pressure, and cholesterol levels, all of which were improved by adding smoking and obesity as factors. This approach thus may give a more precise risk factor compared to CHD. A number of researchers have comparative studies to evaluate the performance of diverse ML techniques. In fact, a study performed by Karthiga et al. proved that out of all the techniques considered, DT actually demonstrates better performances than Naïve Bayes in case of having a public dataset with 573 records to perform the accurate analysis. Another study further reiterates this fact whereby ANN method stands as the one that represents the highest predictive accuracy among all techniques followed. ROC curves have been very helpful in graphically representing the performances of various models. According to a report, the area under the ROC curve of ANN was about slightly over 80%, which was higher than those of the Naïve Bayes and DT algorithms.

2.10 In 2019, Mienye, Ibomoiye Domor, Yanxia Sun, and Zenghui Wang reviewed the "Prediction performance of improved decision tree-based algorithms." Decision tree algorithms have been applied to a wide range of areas such as retail, banking, education, and health due to their ability to process big volumes of data and extract meaningful patterns from this data. The fact that the data mining techniques have taken roots from machine learning, artificial intelligence, and statistics guarantees all-around versatility for these algorithms. The paper does show a conviction that further research is needed in developing practical applications of enhanced decision tree algorithms in real life. Presently, many enhancements remain isolated within academic circles, and their practical utility is yet to be fully realized. It is also a call for more research on applying evolutionary algorithms in optimal feature selection to drastically improve the performances of algorithms using decision trees when dealing with large data-sets.

2.11 In 2019, Vabalas, Andrius, et al. published a study titled "Machine learning algorithm validation with a limited sample size" in PloS One. Application of ML in the studies on autism: Studies applying ML to the knowledge domain for differential diagnosis of individuals with or without autism have increased significantly. In fact, a literature search reported 55 studies that used ML to

classify individuals as autistic or not. The studies overwhelmingly reported accuracy as a measure of performance, which ensured easy interpretation. This literature shows a negative relationship between sample size and reported classification accuracy. Most of the papers analyzed had a median sample size of 80; those with smaller sample sizes tended to have higher accuracy rates compared to larger sample sizes. The trend also holds generally for brain disorders, including autism: in particular, too small sample sizes tend to make estimates of the performance optimistic. One diagnostic plot showed that the sample size was not normal, and hence subjected to log-transformation to normality:. After transformation, a strong negative correlation was found between log-transformed sample size and reported accuracy ($r(53) = -0.70$, $p < 0.001$). This indicates that approximately half of the variance in reported accuracy may be explained by sample size alone. The importance of using robust validation methods also stands out from the literature review. K-fold Cross-Validation (CV) yielded biased performance estimates, especially at small sample sizes. In contrast, nested CV and train/test split approaches gave more robust results regardless of sample size.

2.12 In 2019, Ayatollahi, Haleh, Leila Gholamhosseini, and Masoud Salehi conducted research work in "Predicting coronary artery disease: a comparison between two data mining algorithms." The disease is the leading cause of death in the world. According to WHO, the mortality rate from heart diseases may reach 23 million by 2030 and calls for urgent predictive measures. Recent advancements in data mining technologies have led to novel pathways towards CAD prediction. The health care industry uses vast volumes of information, yet valuable information cannot be extracted easily. Data mining may, therefore, extract less evident patterns and risk factors associated with CAD and thus support the judgements of health care experts. ANN and SVM algorithms were used to advance the predictive values. That is, integration of techniques in data mining into the prediction of CAD fosters a promising approach to the improvement of patient outcome. It is important to carry out comparative analyses of such algorithms as ANN and SVM in order to effectively infer which methods are successful for use within CAD prediction. Further research needs to explore other data mining algorithms that can further improve predictive capabilities in CAD scenarios.

2.13 In 2020, Dipto, Imran Chowdhury, et al.“CAD Prediction Using Several Machine Learning Algorithms” This study uses several machine learning algorithms to improve CAD prediction. Amongst them are the algorithms of Logistic Regression, Support Vector Machine, and Artificial Neural Networks. These will be compared in performance metrics considering accuracy and area under the curve. That is where, usually, such algorithms get analyzed by means of methods like confusion matrices, stratified K-fold cross-validation, and ROC curves. For example, as was evidenced by one study, the average AUC for models applied was 0.98, and predictive accuracy for the models was high. It has also been pointed out that the application of the SMOTE algorithm for class balance correction on the data sets used will provide better accuracy for the models.

2.14 In 2020, Chen, Xueping, et al. presented “Coronary artery disease detection by machine learning with coronary bifurcation features”.reconstructed great abilities of machine learning for detecting CAD by extracting coronary bifurcation features from CCTA images and analysis using various classifiers. The conducted study focused on algorithms such as logistic regression (LR), decision tree (DT), linear discriminant analysis (LDA), k-nearest neighbors (K-NN), artificial neural network (ANN), and support vector machine (SVM). Of these, the polynomial-SVM resulted in the highest classification accuracy up to 100% in the identification of CAD that was best optimized through grid search methods.

2.15 In 2020, Gola D, et al. “Polygenic risk scores outperform machine learning methods in predicting coronary artery disease status”. Coronary artery disease is, among others, deemed one of the primary causes of death globally and has profound effects on public health. This condition involves the deposit within the walls of the arteries, which decreases blood flow and availability of oxygen for the heart and could result in some severe conditions such as myocardial infarction and cardiac arrhythmias. The classic risk assessment models, such as HeartScore and the Framingham Risk Score, are based on clinical variables. However, addition of genetic information through the use of polygenic risk scores has assisted these models. With recent advancements, genome-wide polygenic risk scores of millions of genetic variants have been able to predict CAD risk with favorable precision independent of established associations. Hence, this study has the potential for PRS in diagnosing CAD with a high level of predictability for timely clinical intervention.

2.16 In 2020, Chen, Rung-Ching, et al. published a study titled “Selecting critical features for data classification based on machine learning methods” in the Journal of Big Data. Feature selection is one of the most important operations within the training process of machine learning especially while dealing with large-dimensional data. This paper reviews a few studies that ventured to discuss the relevance of selecting features on classification datasets of data. Methods such as RF, SVM, and KNN are discussed here. The proper selection of critical features is quite necessary for increased accuracy in the model, as well as in computation. In fact, it has been proven that feature selection improves the performance of a classifier by enhancing the quality of data reduced noise and redundancy. In today's world of ecological, health, and finance, many datasets contain variables which are not

improving the model's predictability. Hence, it brought Random Forest among the strong methods of feature selection and classification. This algorithm is recognized for the fact that it can handle both nominal and continuous attributes. Therefore, it is versatile across the various domains. A number of research studies reported high accuracy by RF. In this area, some studies showed accuracies up to 98.57% with optimal feature sets. The variable importance analysis provided by RF attracted attention since it can identify the features by which classification could be done efficiently. There are a lot of comparisons of RF with other machine learning models, such as SVM and KNN. These comparisons point out how many classifiers the RF often outperforms concerning accuracy at the same time, features selection method and choice of features can make a big difference in conclusions.

2.17 In 2021, Vujović, Ž. published a study titled "Classification model evaluation metrics", International Journal of Advanced Computer Science and Applications. Analysis of classification models forms an important aspect of machine learning as applied to medical data classification, for example in the hepatitis C virus dataset analyzed in this paper. Review of literature suggests the involvement of different models and their related metrics in the performance estimation of the classification algorithms. Paper concentrates on four classification models: BayesNet, NaiveBayes, Multilayer Perceptron, and J48. These models are well known within the literature, due to the fact that they are often used in many applicable classification tasks. BayesNet uses a probabilistic graphical model to represent a set of variables and their conditional dependencies, which is very helpful in cases containing uncertain data. NaiveBayes simplifies computation by assuming independence among features, that makes it the most appropriate in high-dimensional datasets. Decision Tree-based J48 Again, the way through which a decision tree-based J48 provides for clear visualization of the decision-making process and, thereby, contributes to understanding the outcome of classification. Review of 16 metrics applied to WEKA software for evaluating these models of classification also happens. Some of the relevant metrics are Matthews Correlation Coefficient, Area of Receiver Operating Characteristic and Precision-Recall Curve Area. In fact, MCC should be used in cases where large classes exist without any strictness of the case of different class sizes and it works well even in those imbalanced sizes. ROC curves represent how true positives versus false positives vary with each other. The PRC is quite informative about unbalanced datasets and may be useful to analyze the precision and recall of the model. In fact, the evaluation metrics for deploying these models on the hepatitis C dataset were not satisfactory: all the metrics were reporting poor classification performance for all of them. MCC levels were at the level of pure random predictions, and ROC Area scores were on the brink of becoming unadequate. This underscores the need to investigate further into the nature of the dataset, possibly data preprocessing, for improved reliability in the outcome of the classification.

2.18 In 2021, Sammout et al. in their paper proposed a “clinical decision support system for the diagnosis of coronary artery disease by using ensemble learning”. CAD is one of the foremost causes of death in many parts of the world, and diagnostic facilities should be ensured for it. Literature makes different points considering the issue of coronary artery disease diagnosis: a focus on the integration of machine learning techniques and clinical decision support systems. According to the World Health Organization, CAD has the highest mortality rates among all non-communicable diseases; therefore, the necessity for accuracy in diagnosis is of vital importance. These are some of the conventional diagnostic techniques, the use of ECG and angiograms, which are usually expensive and invasive, thereby creating a need for a search into other alternatives for diagnosis. The more recent studies have attempted to introduce the use of algorithms in machine learning for the improvement of the sensitivity of the system for diagnosis. This has led to the development of ensemble learning methods, which combine several classifiers, to achieve the best results. Some of them have used a better ensemble model with the Ada-boost algorithm, which achieved very high accuracy for most sets of data. There is also an application of SVM when the data-sets are small.

2.19 In 2021, Wang, Chen et al. Presented “Development and validation of a predictive model for coronary artery disease using machine learning”. The authors have used a large real-world clinical data-set for developing models that can predict CAD risk with effectiveness. There has been much use of ML in health care, especially predictions regarding diseases like CAD. The ML models could be used for big data-sets for pattern identification as well as prediction and improve their diagnosis and treatment outcome. These predictive models were validated on several cohorts with excellent discrimination performance. For example, the AUC of the random forest model was 0.948 in the development cohort, and AUC values more than 0.940 were maintained in validation cohorts. It further indicates that the model is robust and sound to be used as a reliable predictor of CAD risk.

2.20 In 2021, Akella, Aravind, and Sudheer Akella present a review of the use of different ML algorithms to predict CAD disease in their article "Machine learning algorithms for predicting coronary artery disease: efforts toward an open source solution." Interest in the study was amplified by the fact that some clinical data-sets were becoming available for analysis, like the Cleveland data-set, whose data is free and was first introduced in 1988. The Cleveland data-set is considered one of the most popular data-sets at the UCI repository. Many studies have reported investigations concerning this data-set, applying a wide variety of ML

methods. Despite its popularity, the majority of the conducted studies focused on the analysis of only one ML method, providing scant comparison between different algorithms. The results showed that all six machine learning algorithms achieved an accuracy of more than 80%, and the best accuracy was of the neural network model, over 93% with a 0.93 recall. This indeed provides considerable evidence that ML models can have a high diagnostic value for CAD diagnosis, being one of the most critical clinical aspects.

2.21 In 2021, Muhammad et al. researched "Machine learning predictive models for coronary artery disease." CAD remains one of the leading health burdens around the world and especially among developed countries, accounting for over 31% of global mortality burdens according to the World Health Organization. In Nigeria, where it is not considered a great health threat, CAD accounted for 2.82% of the total deaths recorded in 2014, which is a growing number. Early detection and prediction of CAD will reduce mortality rates. The WHO estimated that a huge number of Nigerians have the potential to die prematurely from non-communicable diseases, among which CAD is included. This brings out the need for effective predictive models in identifying persons at risk. The application of ML has emerged as a strong tool in recent times in health care for diagnosis and the prediction of several diseases, including CAD. Some of the rapidly employed techniques include techniques such as data mining, deep learning, and expert systems, which improve diagnosis and decision-making. Therefore, automatic discovery of patterns in big data-sets makes ML suitable for the development of predictive models. Predictive models were developed for the present study using various ML algorithms, including SVM, KNN, random tree, Naive Bayes, gradient boosting, and logistic regression. Of these, the best result was obtained from the random forest model, with 92.04% accuracy and 92.20% ROC.

2.22 In 2022, Heydarian et al. presented "MLCM: Multi-label confusion matrix" in their paper which appeared in IEEE Access. In multi-label classification, one instance receives multiple labels, something quite common across various domains-text, images, and medical data. Besides, the traditional problems of classification have only one label assigned for every instance, while this overlapping nature of the labels in multi-label tasks requires more complex metrics for evaluation. In fact, the MLCM will range from ECG signals to movie posters for multi-label data-sets in order to extract metrics such as precision, recall, and F1-score of each class. This will not only help in evaluating the performance of each individual class but will also provide a better comparison with already existing metrics of the likes of sklearn.

2.23 In 2023, The study of Özbilgin, et al. on using machine learning techniques in the prediction of CAD from the iris, was titled "Prediction of coronary artery disease using machine learning techniques with iris analysis." Different contexts have been explored in the study of iridology-the study of the iris in diagnosing health problems. Present research has shown that changes in iris anatomy may reflect functional changes in a specific organ, and some iris features reflect diseases such as CAD. Several studies have used machine learning to improve the performance of disease detection using iris analysis. Performance usually involves accuracy, sensitivity, specificity, and AUC, among a number of Iris Analysis Techniques. The proposed model of the current study, using an SVM classifier, had succeeded in identifying CAD with a prediction accuracy rate of 93%, thus establishing iris analysis as a promising non-invasive diagnostic tool.

3. Methodology

3.1 Data Collection and Preprocessing

The dataset applied in this study was downloaded from IEEE DataPort. It has a mixture of various clinical, demographic, and relevant features associated with the dataset. Data preprocessing techniques such as treating missing values, encoding the categories, and normalizing the numerical features have been taken to maintain consistency and quality of data.

3.2 Model Development

The following ensemble learning methods by a tree-based algorithm were taken for predictive model development of CAD:

- Random Forest
- GBM
- XGBM
- AdaBoost
- LGBM
- Catboost
- XGBoost

The data was split using a stratified k-fold cross-validation approach in order to better obtain robust and unbiased performance metrics. Each model was trained and evaluated using Accuracy, Precision Score, F1 Score, Recall, and ROC curve.

3.3 Stacking Method

We also developed a metamodel through the stacking technique and aggregated predictions of the individual tree-based models to further enhance predictivity. In this approach, a secondary model is trained on outputs from base models, thereby exploiting their strengths but in an attempt to dampen any weaknesses that they might have.

3.4 Web Application Development

We created a web application using Streamlit to help in the real-life application of our prediction model. This will allow clinicians to have live predictions of CAD risk in a non-

invasive manner by placing a patient's input data on the application interface.

3.5 Evaluation Metrics

The performance of the models is evaluated using accuracy, precision, recall, F1-score, and area under receiver operating characteristic curve as popular standard metrics. These provide a balanced evaluation of the model that predicts it. The advanced techniques of machine learning are integrated into the present study into a user-friendly web application; therefore, it would provide a valuable tool to support early detection and diagnosis of CAD, thus improving patients' outcomes and reducing the reliance on invasive diagnostic methods.

4. REFERENCES

[1] Wilson, Peter WF, and Jane C. Evans. "Coronary artery disease prediction." American journal of hypertension 6.11_Pt_2 (1993): 309S-313S.
https://academic.oup.com/ajh/article/6/11_Pt_2/309S/111894

[2] Pletcher, Mark J., et al. "Using the coronary artery calcium score to predict coronary heart disease events: a systematic review and meta-analysis." Archives of internal medicine 164.12 (2004): 1285-1292.
https://jamanetwork.com/journals/jamainternalmedicine/articleabstract/217101

[3] Marzban, Caren. "The ROC curve and the area under it as performance measures." Weather and Forecasting 19.6 (2004): 1106-1114.
https://journals.ametsoc.org/view/journals/wefo/19/6/825_1.xml?tab_body=abstract-display

[4] Verma, Luxmi, Sangeet Srivastava, and P. C. Negi. "A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data." Journal of medical systems 40 (2016): 1-7.
https://link.springer.com/article/10.1007/s10916-016-0536-z

[5] Praveena, M., and V. Jaiganesh. "A literature review on supervised machine learning algorithms and boosting process." International Journal of Computer Applications 169.8 (2017): 32-35.
https://www.researchgate.net/publication/318486479_A_Literature_Review_on_Supervised_Machine_Learning_Algorithms_and_Boosting_Process

[6] Lin, Weiwei, et al. "An ensemble random forest algorithm for insurance big data analysis." Ieee access 5 (2017): 16568-16575.
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8007210

[7] Osisanwo, F. Y., et al. "Supervised machine learning algorithms: classification and comparison." International Journal of Computer Trends and Technology (IJCTT) 48.3 (2017): 128-138.
https://www.researchgate.net/publication/318338750_Supervised_Machine_Learning_Algorithms_Classification_and_Comparison

[8] Naushad, Shaik Mohammad, et al. "Machine learning algorithm-based risk prediction model of coronary artery disease." Molecular biology reports 45.5 (2018): 901-910.
https://link.springer.com/article/10.1007/s11033-018-4236-2

[9] Abdar, Moloud, et al. "A new machine learning technique for an accurate diagnosis of coronary artery disease." Computer methods and programs in biomedicine 179 (2019): 104992.
https://www.sciencedirect.com/science/article/abs/pii/S0169260718314585

[10] Mienye, Ibomoiye Domor, Yanxia Sun, and Zenghui Wang. "Prediction performance of improved decision tree-based algorithms: a review." Procedia Manufacturing 35 (2019): 698-703.
https://www.sciencedirect.com/science/article/pii/S235197891930736X

[11] Vabalas, Andrius, et al. "Machine learning algorithm validation with a limited sample size." PloS one 14.11 (2019): e0224365.
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0224365

[12] Ayatollahi, Haleh, Leila Gholamhosseini, and Masoud Salehi. "Predicting coronary artery disease: a comparison between two data mining algorithms." BMC public health 19 (2019): 1-9
https://link.springer.com/article/10.1186/S12889-019-6721-5

[13] Dipto, Imran Chowdhury, et al. "Comparison of different machine learning algorithms for the prediction of coronary artery disease." Journal of Data Analysis and Information Processing 8.2 (2020): 41-68.
https://www.scirp.org/journal/paperinformation?paperid=99402

[14] Chen, Xueping, et al. "Coronary artery disease detection by machine learning with coronary bifurcation features." Applied Sciences 10.21 (2020): 7656.
https://www.mdpi.com/2076-3417/10/21/7656

[15] Gola, Damian, et al. "Polygenic risk scores outperform machine learning methods in predicting coronary artery disease status." Genetic epidemiology 44.2 (2020): 125-138.
https://onlinelibrary.wiley.com/doi/full/10.1002/gepi.22279

[16] Chen, Rung-Ching, et al. "Selecting critical features for data classification based on machine learning methods." Journal of Big Data 7.1 (2020): 52.
https://link.springer.com/article/10.1186/s40537-020-00327-4

[17] Vujović, Ž. "Classification model evaluation metrics." International Journal of Advanced Computer Science and Applications 12.6 (2021): 599-606.
https://www.researchgate.net/publication/352902406_Classification_Model_Evaluation_Metrics

[18] Sammout, Rawia, et al. "A proposal of clinical decision support system using ensemble learning for coronary artery disease diagnosis." Wireless Mobile Communication and Healthcare: 9th EAI International Conference, MobiHealth 2020, Virtual Event, November 19, 2020, Proceedings 9. Springer International Publishing, 2021.
https://eudl.eu/pdf/10.1007/978-3-030-70569-5_19

[19] Wang, Chen, et al. "Development and validation of a predictive model for coronary artery disease using machine learning." Frontiers in Cardiovascular Medicine 8 (2021): 614204.
https://www.frontiersin.org/journals/cardiovascularmedicine/articles/10.3389/fcvm.2021.614204/full

[20] Akella, Aravind, and Sudheer Akella. "Machine learning algorithms for predicting coronary artery disease: efforts toward an open source solution." Future science OA 7.6 (2021): FSO698.
https://www.tandfonline.com/doi/epdf/10.2144/fsoa-2020-0206?needAccess=true

[21] Muhammad, L. J., et al. "Machine learning predictive models for coronary artery disease." SN Computer Science 2.5 (2021): 350.
https://link.springer.com/content/pdf/10.1007/s42979-021-00731-4.pdf

[22] Heydarian, Mohammadreza, Thomas E. Doyle, and Reza Samavi. "MLCM: Multi-label confusion matrix." IEEE Access 10 (2022): 19083-19095.
https://ieeexplore.ieee.org/abstract/document/9711932

[23] Özbilgin, Ferdi, Çetin Kurnaz, and Ertan Aydın. "Prediction of coronary artery disease using machine learning techniques with iris analysis." Diagnostics 13.6 (2023): 1081.
https://www.mdpi.com/2075-4418/13/6/1081