Ensemble Model for Early Prediction of Coronary Artery Disease

Sarthak Mukherjee[1], Sounak Maiti[2], Swastik Das[3], Samit Karmakar[4], Amit Kumar Das[5], Mili Sarkar[6]
Electronics & Communication Engineering Dept.
Institute of Engineering & Management, University of Engineering & Management
Kolkata, India
Email: [1]Sarthak.Mukherjee2023@iem.edu.in, [2]Sounak.Maiti2023@iem.edu.in, [3]Swastik.Das2023@iem.edu.in, [4]iem.samit@gmail.com, [6]amit.das@iem.edu.in, [7]milisarkar393@gmail.com
ORCID: https://orcid.org/0000-0002-0858-5168

2026 9th International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech) | 979-8-3315-6339-4/26/$31.00 ©2026 IEEE | DOI: 10.1109/IEMENTECH202669403.2026.11434094

***Abstract*—coronary artery disease (CAD) has become a major factor in today's world with respect to its fatal effects on humans worldwide. This disease stands out to be the primary cause of deaths of about 9.14 million people worldwide, as confirmed by WHO in 2022. Hypertension, diabetes, dyslipidemia, smoking, and obesity account for the major causes of CAD. Since there are many cases of CAD without prior symptoms, there is a significant need of an early-risk prediction system using machine learning so that presence or absence of CAD can be identified based on these symptoms. A soft-voting ensembled classifier model along with combined XGBoost, Balanced Random Forest and Gradient Boosting resulting in an accuracy of 90.5% helps in predicting discharge, dama or expiry of the patient based on the symptoms.**

Keywords—Coronary Artery Disease, Machine Learning, Soft Voting Ensemble, XGBoost, Balanced Random Forest, Gradient Boosting.

I. Introduction

Coronary Artery Disease (CAD) is a common type of heart disease seen in humans. It arises due to the insufficient supply of blood and oxygen to the myocardium of the heart. It involves the blockage of the coronary arteries, thereby leading to the inability of supply of blood to the heart muscles. It gradually leads to the formation of plaques in the coronary arteries, resulting in a reduced blood flow. Fats, cholesterol, and other substances build in and on the walls of the coronary arteries, resulting in a condition called atherosclerosis. The formation of plaques makes the coronary arteries narrow in size.

Though traditional treatments such as angiography, treadmill stress testing, and ECG tests are considered valuable in the treatment of coronary artery diseases, they have certain limitations. They involve the requirement of special equipment, highly skilled medical laboratory technologists, and complex procedures. This requirement makes the treatment of CAD of such a huge population in the world a hectic task. It has been observed that there are many patients who show late symptoms or no prior symptoms at all with respect to this disease. As a result, they are being subjected to late diagnosis and medical tests for this disease. This decreases their chances of surviving from this disease, thereby resulting in many deaths all over the world.

Due to all these challenges, the medical world is in urgent need of some data analysis and machine learning models so that the presence of CAD can be detected and predicted correctly despite the absence of those prior symptoms in the human body. Recently, CAD has become a burden to medical units worldwide. In many developing countries, it is becoming very difficult to arrange highly skilled cardiologists, special machinery, and emergency medical units. Most of the existing machine learning models predict only the presence or absence of CAD, but our model helps in the prediction of three categories of the outcome of the patient. This helps in bridging the gap between pure diagnosis treatments and clinical prognosis predictions.

Machine learning techniques have emerged as a significant approach in recent times in the field of medical analytics. It helps in predicting complex relationships among various features, contributing to the decisions of considering only important factors of coronary artery diseases. Using ensemble learning techniques, we have combined multiple algorithms to reduce overfitting and maintain a decent accuracy. The hospitals can now use these early predictions of the outcome of the patient to decide the preference of high-risk and low-risk patients in the context of emergency treatments. They can also organize and support more doctors if they notice high cases of expiry.

In this study, we have developed an ensemble voting classifier model which combines XGBoost Classifier, Balanced Random Forest Classifier, and Gradient Boosting Classifier. This model not only helps in classifying the three categories of outcome of the patient, namely discharge, dama (discharged against medical advice) and expiry, but also helps in identifying the top 20 features that contribute to the classifier model. This explains the importance of data science and data analytics in the field of medical sciences besides machine learning in today's world. It also figures out the correlation analysis of each numerical feature variable with the target variable. This concludes with the fact that as time progresses, artificial intelligence along with a high level of statistics is very necessary in prediction of coronary artery diseases.

II. Methodology

A. Dataset description

The dataset used in this study consists of **6,611** patient records with **53** clinical and demographic attributes, collected for coronary artery disease outcome analysis. Each record corresponds to a single patient admission and includes laboratory measurements, medical history, diagnostic indicators, and treatment-related variables.

The features can be broadly categorized into four groups:

- **Demographic Features:** Age, gender, ICU stay duration, and lifestyle factors such as smoking and alcohol consumption.

- **Clinical Histories:** Diabetes, hypertension, chronic kidney disease (CKD), cardiomyopathy, prior CAD, and neurological conditions.
- **Laboratories and diagnostic Parameters:** Hemoglobin, leukocyte count, platelet count, serum urea, creatinine, glucose levels, B-type natriuretic peptide (BNP), and ejection fraction.
- **Cardiac Condition Indicators:** Stable angina, acute coronary syndrome (ACS), ST-elevation myocardial infarction (STEMI), heart failure subtypes (HFrEF, HFNEF), arrhythmias, and valvular disorders.

The target variable, outcome, is a multi-class label representing the patient's discharge status: discharge, DAMA (discharged against medical advice), and expiry. This structured representation enables the model to capture both short-term clinical severity and long-term disease progression, making it suitable for outcome-based CAD prognosis.

B. Data Preprocessing

The dataset does not contain any missing values. The label encoding technique is applied to the target variable 'outcome'. As label encoding involves the process of encoding in an alphabetical order:

- [0] as 'Dama'.
- [1] as 'Discharge'.
- [2] as 'Expiry'.

Using the Column Transformer pipeline, standardization technique has been applied on the numerical columns and One Hot Encoding technique has been applied on the categorical columns. It also prevents accidental loss of data. The transformer fits the training set and then applied consistently to both training and test sets.

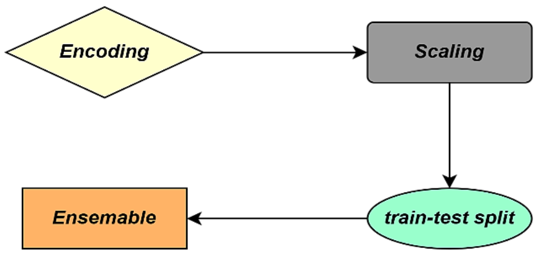

Fig. 1: Data preprocessing and feature transformation analysis

C. Evaluation Metrics

Multiple evaluation metrics are used to improve the performance of the ensemble voting technique and prevent the machine learning model from overfitting. Accuracy is used to determine the total measure of correct predictions. Precision, recall, and F1-score are also used to provide deeper insights for all three categories of the target variable 'outcome'. In addition to it, a confusion matrix is also used to evaluate the level of misclassifications that are performed across the 'dama', 'discharge' and 'expiry' classes, thereby providing us with the prediction strengths and weaknesses.

III. Result and Performance Analysis

This part fundamentally presents the experimental outcomes obtained from the provided machine learning model for coronary artery disease prediction. The performance of the mentioned model is evaluated considering various essential factors such as classification metrics, confusion matrix analysis and features importance interpretation to assess both predictive accuracy and clinical relevance.

A. Overall Model Performance & ROC Graph

The enlisted soft-voting ensemble model includes the integration of certain critical components such as XGBoost, Balanced Random Forest and Gradient boosting classifiers, demonstrating strong predictive capability when evaluated on a particular test dataset. The model attained an overall classification accuracy of 90.5% illustrating a high level of correctness in predicting the outcomes across the three stated classes: discharge, dama and expiry. This satisfactory performance rate highlights the effectiveness and accuracy of combining multiple learning algorithms to reduce bias, variance and overfitting. The provided Receiver Operating Characteristic (ROC) curve for this ensemble model was plotted using a one-vs-rest system to assess its capability across multiple outcome classes. This particular model achieved the highest (AUC) of 0.98 for "Expiry" class whereas the AUC of "DAMA" is low up to 0.80 showcasing desirable predictive capability. In general, this ROC graph confirms that the proposed ensemble model effectively distinguishes between different clinical outcomes with appreciable performance.

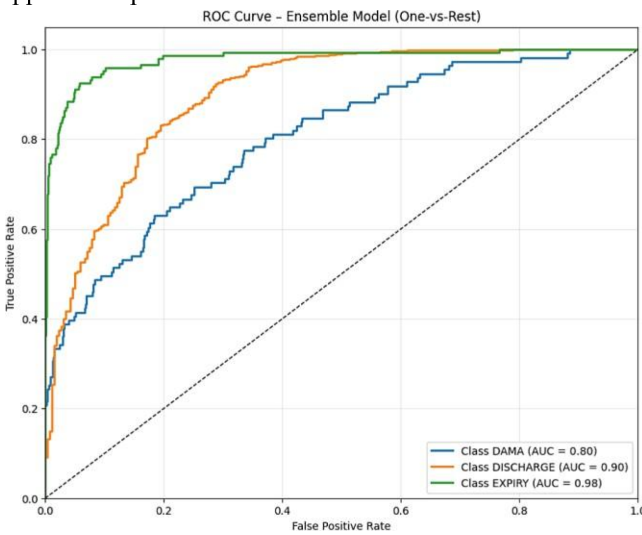

Fig. 2: Multi-Class ROC Curves

B. Class-Wise Performance Evaluation

- The class wise evaluation basically represents the exceptional performance of the ensemble model in predicting the discharge class as well as the other two classes. The discharge class mainly constitutes the majority of the provided dataset. High precision and recall values mainly indicate that most of the discharged patients are correctly identified with minimum false predictions.
- For the expiry class, the ensemble model attains a high recall, which is critical from a clinical view as it reduces the risk of missing high-risk patients. Although the dama class exhibited slightly lower performance metrics as compared to the other classes the predictions remain within the acceptable limits showcasing the inherent overlap of clinical characteristics between dama and discharge cases.

C. *Confusion Matrix Analysis*

A confusion matrix primarily provides a view of the prediction in the depth of behavior of the provided ensemble model by illustrating correct and incorrect classifications across all the outcome classes. A strong concentration of values along the diagonal elements of the matrix indicates that the majority of predictions were accurate, reflecting the high overall performance of the proposed ensemble model.

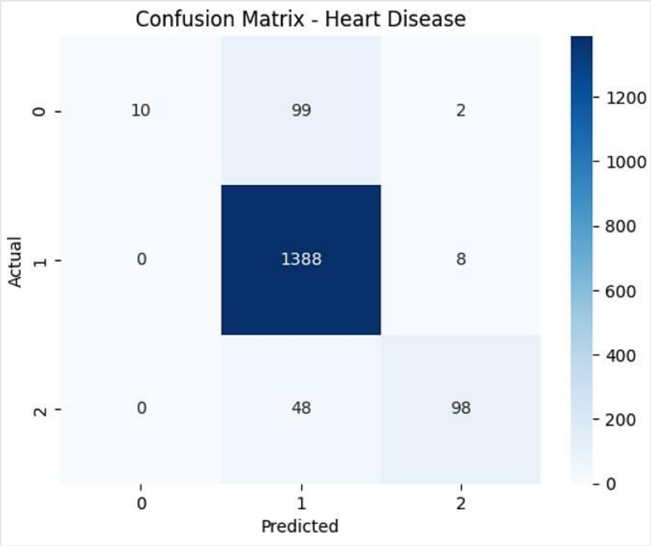

Fig. 3: Confusion matrix of ensemble model

Misclassifications were mainly observed between the dama and discharge classes. This is expected as patients falling under these categories often share similar clinical profiles and the final outcome may depend on certain non-clinical factors such as personal choice, financial constraints, or availability of post-hospital care. Despite this overlap, the number if incorrect predictions remained low, demonstrating the model's ability to distinguish subtle differences in patient conditions which plays a very crucial role in clinical aspect.

D. *Feature Importance and Clinical Interpretation*

- Feature importance analysis was performed in order to interpret the decision-making process of the ensemble model and identify the most influential predictors of the patient outcomes.

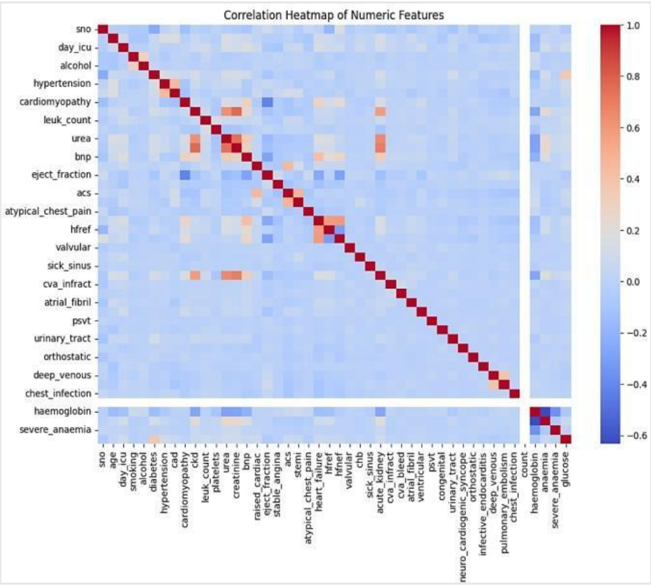

Fig. 4: Correlation heatmap of numerical clinical features

- Ejection fraction emerged as the most crucial significant predictor, reflecting its direct relationship with cardiac function and disease severity. Reduced ejection fraction is strongly associated with heart failure, arrhythmias and increased mortality risk.
- Various renal function indicators, including serum creatinine and blood urea, were highly influential, emphasizing the strong relation between cardiovascular and renal health in CAD patients.
- As per the conducted detailed analysis it has been observed that "Age" factor played a crucial role, as older patients often exhibit reduced physiological resilience and multiple comorbidities, increasing the probability of adverse outcomes. Conditions such as diabetes, mellitus, heart failure and acute coronary syndrome (ACS) further contributed significantly to model predictions aligning with established clinical evidence regarding CAD progression.
- The strong connection between model-identified features and established clinical evidence confirms the medical validity of the proposed approach.
- By showcasing transparent and clinically meaningful predictions, the mentioned ensemble model demonstrates its potential as a reliable decision-making system for an outcome-based diagnosis of coronary artery disease.
- Certain significant laboratory parameters such as hemoglobin levels and platelet count indirectly influenced predictions, suggesting the significance of anemia, inflammation and coagulation status in disease progression.

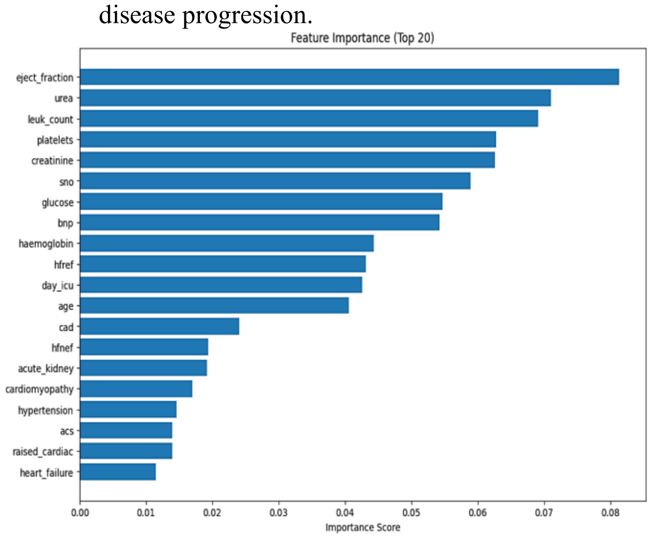

Fig. 5: Top 20 clinical features importance bar graph

The clarity offered by feature importance analysis supports explainable artificial intelligence (XAI) principles, facilitating clinical trust and adoption for the coronary artery disease prediction system through the proposed ensemble model.

IV. Future Scope

The enlisted machine learning-based framework for coronary heart disease (CAD) prediction demonstrates promising and appreciable results; however, several enhancements and advancements can further improve its clinical applicability, robustness, accuracy and efficiency. The future scope of this particular research is outlined under the following aspects explaining certain spheres of development which can lead to the overall progress of the machine learning model in the coming years.

A. *Dataset Expansion and Model Generalization*

Collection of multi-center datasets from various hospitals to improve population diversity at an appreciable aspect.

- Validation of the model on various international datasets for global applicability.
- Evaluating model performance across different age groups and genders.

Incorporating various socio-economic factors, Influencing patient outcomes.

B. *Advanced Modeling Techniques*

Integration of deep learning models such as LSTM and GRU for temporal prediction. Use of attention mechanisms to highlight clinically important features. Use of hybrid modeling combining statistical and machine learning approaches. Create circumstances resulting in the development of clinician-friendly explanation dashboards. Adopting learning models that evolve with incoming data. We can also utilize model compression for faster inference in clinical settings. Robustness testing against various adversarial and noisy inputs can also be very effective in use as well. Validation of explainability results with suitable cardiologist feedback. Incorporation of fairness-aware learning to reduce bias.

C. *Clinical Deployment and Real-Time Decision Support*

Based on the work there are some various types of clinical deployment and corresponding real-time decision support.

- Integration with electronic health record (EHR) systems.
- Development of hospital dashboards for outcome risk visualization.
- Can be used in emergency triage and ICU prioritization.
- Early warning system for high-risk patient deterioration.
- Mobile and cloud-based deployment can be adapted for remote healthcare centers.
- Support for physician-in-the-loop decision making.
- Deployment in low-resource and rural healthcare settings can be of significant help.
- Cost-benefit analysis for healthcare institutions and scalability for high-volume hospital environments.

V. Conclusion

This paper presents an efficient and effective machine learning model for outcome prediction in coronary artery disease using soft-voting ensemble classifier. Utilizing the integration of XGBoost, Balanced Random Forest and Gradient Boosting models has provided a significant predictive accuracy while maintaining a stable performance across classes such as discharge, DAMA and expiry. The model also successfully captures complex clinical relationships and identifies key features aligned with the established medical knowledge. The results generated from the mentioned model demonstrate the potential of ensemble learning techniques to support early risk detection and outcome-based prognosis in CAD patients.

The authors would like to express their heartiest gratitude to the Institute of Engineering & Management (IEM), Kolkata for providing the academic environment and reliable resources necessary to carry out this research. The authors are thankful to Mr. Satyajit Chakraborty, Director, IEM for his continuous encouragement and institutional support. Also the authors would like to extend their special thanks to Prof Amit Kumar Das, Prof Samit Karmakar and Prof Mili Sarkar for their valuable guidance, constructive feedback and technical insights throughout the course of this work.

VI. References

[1] Coronary Artery Disease Prediction Using Machine Learning, Milan Sharma; Rakesh Kumar; Meenu Gupta; Ronakkumar Bathani, 2025 3rd International Conference on Disruptive Technologies (ICDT), 07-08 March 2025, 10.1109/ICDT63985.2025.10986455.

[2] Vascular Ultrasound Data during Enhanced External Counterpulsation for Predicting Coronary Artery Disease, Yahui Zhang; Dilang Yang; Xu Zhao; Yang Luo; Yinfen Wang; Chubin Zhong, 2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 14-18 July 2025, 10.1109/EMBC58623.2025.11254099.

[3] An Optimized Semantic Segmentation Deep Learning Model for Automated Coronary Artery Calcification Prediction, Nathan Yee, 2024 IEEE 5th International Conference on Pattern Recognition and Machine Learning (PRML), 19-21 July 2024, 10.1109/PRML62565.2024.10779669.

[4] Applying Healthcare Analytics to Diagnose and Predict Coronary Artery Disease Using Machine Learning Techniques, Atreyi Pramanik; Pranchal Rajput; Srinivas Aluvala, 2023 International Conference on Advanced Computing & Communication Technologies (ICACCTech), 23-24 December 2023, 10.1109/ICACCTech61146.2023.00104.

[5] Predictive Models of Coronary Artery Disease Based on Computational Modeling: The SMARTool System, Antonis I. Sakellarios; Panagiota Tsompou; Panagiotis Siogkas; Vassiliki Kigka; Ioannis Andrikos; Nikolaos Tachos, 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 23-27 July 2019, 10.1109/EMBC.2019.8857040.

[6] Segmentation-Based CCTA Image Analysis to Support the Diagnosis of Coronary Artery Disease, Giuseppina Settanni; Raffaele Vaira; Michele Spedalieri; Mariapia Colatruglio; Nicola Savino; Angelo Galiano, 2024 IEEE International Humanitarian Technologies Conference (IHTC), 27-30 November 2024, 10.1109/IHTC61819.2024.10855144.

[7] Coronary Artery Disease Prediction Using Facial Features, Rutuja Khedkar; Mayur Ashok Jagtap; Nikhil Vinesh Bhoje; Vipul Narendra Patil, 2024 OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 4.0, 05-07 June 2024, 10.1109/OTCON60325.2024.10687745.

[8] Ensemble Machine Learning and Its Validation for Prediction of Coronary Artery Disease and Acute Coronary Syndrome Using Focused Carotid Ultrasound, Ankush D. Jamthikar; Deep Gupta; Laura E. Mantella; Luca Saba; Amer M. Johri; Jasjit S. Suri, IEEE Transactions on Instrumentation and Measurement (Volume: 71), 31 December 2021, 10.1109/TIM.2021.3139693.

[9] Coronary Heart Disease Prediction Using GKFCM with RNN, Varun Malik; Ruchi Mittal; Ajay Rana; Irfan Khan; Pankaj Singh; Bashar Alam, 2023 6th International Conference on Contemporary Computing and Informatics (IC3I), 14-16 September 2023, 10.1109/IC3I59117.2023.10398020.

[10] Towards Machine Learning Techniques for Enhanced Prediction of Coronary Artery Disease, Abhishek Kukreti; Srabanti Maji, 2024 IEEE International Conference on Contemporary Computing and Communications (InC4), 15-16 March 2024, 10.1109/InC460750.2024.10649400.

[11] Prediction of coronary artery disease using logistic regression, Debabrata Swain; Paawan Sharma; Vinay Vakharia; Tapash Kumar Tanty, 2022, 10.1515/9783110734652.