INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 4, April 2026 | ISSN: 2456-4184 | IJNRD.ORG

SYSTEMATIC ANALYSIS OF CORONARY ARTERY DISEASE PREDICTION USING MACHINE LEARNING AND CLINICAL DATA

Mrs. P. Gokila,AP , Mrs. E. Kiruthika,AP, Mr. G. Karthik, Ms. M. Rosmitha , Mr. C. Sivaraj
Department of Computer Science and Engineering, INFO Institute of Engineering, Kovilpalayam, Coimbatore, India – 641107
Email: {gokilacseinfo,kiruthika9605,karthikganesh1326,rosmithamathiyalagan,sivarajcofficial}@gmail.com

***Abstract*— Coronary Artery Disease (CAD) continues to be a leading cause of death globally [19]. Hence, there is an urgent need for early, precise, and economically viable diagnostic tools. Recent breakthroughs in Machine Learning (ML) and Deep Learning (DL) have made it possible to automatically analyze medical images, biosignals, and clinical data to assist doctors in CAD diagnosis and prediction. This survey provides a thorough overview of the latest advances in ML- and DL-based automated analysis of medical images, biosignals, and clinical data for CAD detection, segmentation, classification, and risk assessment. The surveyed studies cover a wide range of modalities, such as Computed Tomography Coronary Angiography (CTCA), Invasive Coronary Angiography (ICA), Electrocardiogram (ECG), Phonocardiogram (PCG), and hybrid clinical datasets. Each study is evaluated in terms of methodology, strengths, and weaknesses. On the basis of the comparative study, the research gaps and findings are pointed out, emphasizing the need for efficient, interpretable, and computationally feasible CAD prediction models using non-invasive clinical data. This survey is intended to provide a systematic basis for developing improved CAD prediction models for practical clinical applications.**

***Index Terms*— Coronary Artery Disease, Machine Learning, Deep Learning, ECG, CT Angiography, Feature Engineering, Boosted Decision Trees, Medical Decision Support Systems.**

I. INTRODUCTION

Coronary Artery Disease (CAD) is a major cause of morbidity and mortality worldwide and contributes to a substantial number of cardiovascular deaths [19]. CAD is a consequence of the narrowing or occlusion of the coronary arteries by the accumulation of atherosclerotic plaques, leading to a reduction in blood supply to the heart muscle and potentially causing myocardial infarction or sudden cardiac death [19], [20]. Thus, the early and accurate diagnosis of CAD is essential for its effective management and prevention of serious cardiac events. Although conventional methods of diagnosis, such as invasive coronary angiography (ICA) and computed tomography coronary angiography (CTCA), have high diagnostic accuracy, they are costly, invasive, time-consuming, and require specialized expertise [3], [4], [9]. In recent years, machine learning (ML) and deep learning (DL) algorithms have been recognized as highly effective tools for the automatic detection, segmentation, classification, and prediction of CAD by learning complex patterns from medical images, physiological signals, and clinical data [12], [13], [21]. Convolutional neural networks and U-Net models have shown outstanding performance in image-related tasks, whereas signal processing techniques based on electrocardiogram (ECG) and phonocardiogram (PCG) signals have been proposed as non-invasive and cost-effective alternatives for large-scale screening [1], [2], [6], [10], [12], [17], [22], [25].Moreover, the traditional ML models like Boosted Decision Trees and Logistic Regression offer accurate and interpretable results when combined with efficient feature engineering [8], [14], [15], [33]. However, the current CAD diagnostic systems are challenged by the issues of computational complexity, lack of explainability, overdependence on imaging modalities, and poor generalization capabilities for different patient groups. This survey work aims to provide a detailed review of the recent ML- and DL-based solutions for the detection and prognosis of CAD, and based on their strengths and weaknesses, it highlights the research gaps that lead to the development of an improved, non-invasive, efficient, and explainable CAD prediction system.

II. LITERATURE SURVEY

Recent works have shown the potential of machine learning (ML) and deep learning (DL) algorithms in enhancing the diagnosis and prognosis of coronary artery disease (CAD) from medical images, physiological signals, and clinical data. Deep learning algorithms like U-Net have been successfully used for the efficient segmentation of the aorta and coronary arteries from CT coronary angiography images with high accuracy and low computational cost, although these algorithms are mainly used for anatomical analysis. Automated systems for the identification of the coronary arteries from CT images have also shown high clinical adaptability and labeling accuracy, but they do not focus on the prediction of disease severity. Invasive coronary angiography-based systems using convolutional neural networks have shown excellent classification accuracy for severe lesions, but their performance is not satisfactory for mild stenosis, and their invasive nature limits their use for large-scale screening. Attention-based deep learning models have further improved the performance of segmentation and localization tasks, although they increase the computational cost. To address the challenges associated with imaging-based approaches, some research works have attempted non-invasive signal-based solutions using electrocardiogram (ECG) and phonocardiogram (PCG) signals, in which dual-input neural networks and the analysis of variability of electromechanical

delay have achieved high accuracy. ECG-only models, which combine feature design with one-dimensional convolutional neural networks, have enhanced robustness and generalization capabilities for large datasets. Moreover, traditional machine learning models, specifically Boosted Decision Trees, have achieved better performance and explainability for CAD prognosis based on structured clinical data. Although promising, the current state-of-the-art methods are challenged by issues of explainability, efficiency, and generalization.

A. A Computationally Efficient Approach to Segmentation of the Aorta and Coronary Arteries Using Deep Learning

The present research work is devoted to the problem of CAD diagnosis in conditions of qualified radiologists' scarcity, especially in emergency conditions [1]. The authors describe a fully automatic two-dimensional U-Net CNN-based deep learning system for aorta and coronary arteries segmentation from CTCA images [1]. Two independent models are trained: one for segmenting jointly both aorta and coronary arteries and another one devoted only to coronary arteries [1]. Special attention has been paid to computational complexity: the proposed system is able to run on a common hospital server without using GPU acceleration [1]. Experimental results demonstrated that the Dice similarity coefficients equal to 91.20% and 88.80% for both tasks respectively [1].

Methodology: This paper presents a fully automated 2D U-Net deep learning approach for the segmentation of the aorta and coronary arteries from CTCA scans. Two different models are developed: one for the joint segmentation of the aorta and coronary arteries, and another for the segmentation of the coronary arteries alone. The approach aims to be computationally efficient and does not require the use of GPUs.

Advantages:The proposed approach shows strong segmentation performance with high Dice similarity coefficients of 91.20% and 88.80% for the targeted anatomical structures. The computational efficiency of the approach allows it to be run on standard hospital servers without the use of graphical processing units, making it applicable for real-time clinical use.

Limitations/Gaps:Although the proposed approach shows strong segmentation performance, it is limited to anatomical segmentation and does not involve coronary artery disease classification or prediction. Moreover, the approach is based solely on CT coronary angiography imaging, which is costly and involves radiation, thus limiting its applicability for repeated screening and population studies.

B. AI-Based Detection of Coronary Artery Occlusion Using Acoustic Biomarkers Before and After Stent Placement

In this paper, a non-invasive approach for CAD detection using signals of heart sounds has been investigated [2]. The Matching Pursuit method is used in this research for the analysis of phonocardiogram signals, represented as "atoms," which are then analyzed with a "DeepSets"-based deep learning method. The goal is to distinguish between pre- and post-heart sounds resulting from percutaneous coronary intervention (PCI) therapy [2]. The proposed model detects specific acoustic biomarkers related to coronary occlusion, represented as a classification problem with a high accuracy of 88.06%, using full cardiac cycle signals, and a smaller accuracy of 71.43%, using diastolic signals [2].

Methodology: The paper uses heart sound recordings and the Matching Pursuit technique for signal decomposition, and then uses the DeepSets-based deep learning model to classify CAD-related acoustic biomarkers before and after PCI.

Advantages:The proposed technique is non-invasive and economical, making it ideal for frequent and long-term patient observation. By detecting individualized acoustic biomarkers, the technique allows for individualized analysis and better monitoring of disease progression. Moreover, the study proves the efficient use of artificial intelligence methods for biosignal analysis, thereby establishing their potential in cardiovascular diagnosis.

*Limitations\Gaps:*The drawback of the study is that it uses a small data set with only 12 patients, thereby hindering the generalization and validity of the obtained results. Moreover, the performance of the model is highly uneven for different parts of the cardiac cycle, thereby making it less reliable for general use.

C. Automatic Identification of Coronary Arteries in Coronary Computed Tomographic Angiography

This paper discusses an algorithm for automatic coronary artery identification and labeling, which meets the criteria of the Society of Cardiovascular Computed Tomography (SCCT) [3]. This algorithm is capable of automatically identifying and labeling the main coronary arteries and their branches such as RCA, LAD, LCx, and their sub-branches in a matter of seconds for a given number of CCTA datasets [3]. The algorithm has been in use in over 100 hospitals and has been tested using 892 datasets, resulting in a labeling accuracy of 95.96% [3].

*Methodology:*The authors described an automatic algorithm for the identification and labeling of the coronary arteries according to SCCT guidelines. The algorithm is capable of automatically identifying multiple branches of the coronary arteries in seconds per dataset.

Advantages: The proposed algorithm has a high accuracy of 95.96%, ensuring high reliability in the automatic identification of the coronary arteries. The algorithm has been successfully implemented in over 100 hospitals, ensuring its applicability in real-world settings. Additionally, the fact that the system is fully automated and validated against expert annotations ensures its effectiveness in real-world settings.

*Limitations/Gaps:*Although the algorithm has a high accuracy in identifying the coronary arteries, it does not aim to predict the severity or progression of coronary artery disease. Additionally, the system relies on the availability of high-quality images of the coronary computed tomographic angiography, which may not be the case if there are artifacts in the images.

D. Coronary Artery Disease Classification with Different Lesion Degree Ranges Based on Deep Learning

The effect of lesion severity on the deep learning-based

classification by CAD using invasive coronary angiography Pimages is discussed by the authors [4]. Various types of CNN architectures, i.e., DenseNet, ResNet, MobileNet, and NASNet, are trained with lesion and non-lesion image patches using different thresholds for lesion severity [4]. The experimental outcome shows that, for high severity, the values achieved for F-measure and AUC are 92.7% and 98.1%, respectively [4].

*Methodology:*This paper assesses the performance of various CNN models (DenseNet, ResNet, MobileNet, and NASNet) for binary lesion classification tasks using ICA image patches with different lesion severity ranges.

*Advantages:*The proposed method has a high Area Under the Curve (AUC) of 98.1% and an F-measure of 92.7%, which indicates excellent classification accuracy. The proposed method offers a comprehensive assessment for various ranges of lesion severity, providing important information on the influence of different levels of coronary artery stenosis on the classification accuracy of the proposed method. This analysis emphasizes the correlation between lesion severity and classification accuracy, which helps to improve the understanding of deep learning-based diagnostic systems.

Limitations/Gaps: The proposed method uses invasive coronary angiography imaging, which makes it unsuitable for large-scale clinical applications. Furthermore, the proposed method has lower classification accuracy for smaller ranges of lesion severity, which makes it less effective for early or mild coronary artery disease.

E. Deep Learning-Based Segmentation and Localization in CT Angiography for Coronary Heart Disease Diagnosis

This paper proposes the Multi-stream Attention-guided Coronary Analysis Network (MACAN), which jointly segments and localizes the coronary arteries from the CT angiography images [5]. The proposed model leverages the dual-stream architecture, which fuses channel and spatial attention mechanisms to enhance feature learning [5]. Further, the proposed model has been tested on the ARCADE dataset, which has outperformed various state-of-the-art segmentation models [5].

Methodology: The proposed MACAN framework employs a dual-stream attention-guided network that integrates segmentation and localization tasks with channel and spatial attention mechanisms.

Advantages:The proposed approach integrates segmentation and localization tasks within a unified framework, which enhances the overall diagnostic accuracy. The incorporation of attention mechanisms improves feature extraction by emphasizing clinically relevant regions, leading to more precise anatomical analysis. Furthermore, the model demonstrates superior performance compared to several existing state-of-the-art segmentation techniques.

Limitations/Gaps: Despite its strengths, the model achieves only moderate classification accuracy and F1-score, indicating room for improvement in disease discrimination. In addition, the architectural complexity of the framework is relatively high, which increases computational requirements and may limit its practicality for real-time or resource-constrained clinical environment.

F.Dual-Input Neural Network Integrating Feature Extraction and Deep Learning for CAD Detection Using ECG and PCG

The paper proposed employing a dual input neural network, which combines hand-crafted feature extraction and deep learning-based representation of ECG and PCG signals, which were recorded simultaneously [6]. In the paper, the information gain ratio method of feature selection and deep learning on multi-channel signals were employed. The proposed system had achieved an accuracy of 95.62% [6].

Methodology: The proposed dual-input neural network integrates feature extraction and deep learning techniques for ECG and PCG signals.

*Advantages:*The combination of multi-modal physiological signals greatly improves the diagnostic capability, resulting in a high classification accuracy of 95.62%. The proposed model successfully integrates handcrafted feature extraction and deep learning features, providing a good balance between interpretability and accuracy. In addition, the system is non-invasive and feasible for clinical implementation.

*Limitations/Gaps:*The two-input structure raises the complexity of the model, which could result in increased computational complexity. Furthermore, the system requires the simultaneous acquisition of ECG and PCG signals, which may cause difficulties in data acquisition and could restrict the applicability of the system in scenarios where high synchronization is not easily accomplished.

G. Improved CAD Classification via Feature Engineering and 1D CNN

This paper is concerned with the enhancement of CAD detection via ECG signals using enhanced feature design and a one-dimensional CNN [7]. The approach ensures signal quality by identifying diagnostically reliable segments before model training [7]. The MIMIC-III dataset demonstrates better robustness and generalization compared to conventionally trained CNN-based models [7].

Methodology: The proposed methodology applies feature engineering to enhance the quality of the ECG signal prior to classification by a 1D-CNN model on the MIMIC-III dataset.

Advantages: The ECG-based diagnostic method is cost-effective and non-invasive, making it ideal for large-scale screening and patient follow-up. The addition of effective feature engineering improves the overall generalizability and robustness of the model against noisy signals, making it more reliable.

Limitations/Gaps: The model is highly dependent on the quality of the feature engineering process, which could be a hindrance to consistency across different models.

SUMMARY OF ENGAGEMENT DETECTIONSTUDIES

TABLE I

Figure.1. Comparative Analysis of Recent Machine Learning and Deep Learning Approaches for Coronary Artery Disease Detection

Title	Dataset Used	Pros	Cons	Performance Metrics
A Computationally Efficient Approach to Segmentation of the Aorta and Coronary Arteries Using DL [1]	CT Coronary Angiography (CTCA) datasets	Computationally efficient, deployable without GPU, suitable for hospital servers	Limited to segmentation, no CAD severity prediction	Dice Score: 91.20% (Aorta + Coronary), 88.80% (Coronary only)
AI-Based Detection of Coronary Artery Occlusion Using Acoustic Biomarkers [2]	Heart sound recordings from 12 PCI patients	Non-invasive, patient-specific monitoring, acoustic biomarker discovery	Small dataset, reduced accuracy for isolated diastolic signals	Accuracy: 88.06% (full cycle), 71.43% (diastolic)
Automatic Identification of Coronary Arteries in CCTA [3]	892 CCTA datasets (SCCT standard)	Fast artery identification, clinically deployed in hospitals	No disease or severity estimation	classification Accuracy: 95.96%
CAD Classification with Different Lesion Degree Ranges Based on DL [4]	ICA image patches from 42 patients	High accuracy for severe lesions, multiple CNN evaluation	Performance degrades for low-severity lesions, invasive imaging	AUC: 98.1%, F-measure: 92.7%
Deep Learning-Based Segmentation and Localization in CT Angiography (MACAN) [5]	ARCADE X-ray angiography dataset	Joint segmentation and localization, attention-guided feature learning	High computational complexity, moderate F1-score	F1-score: ~43% (Phase-1), ~41% (Phase-2)
Dual-Input Neural Network Using ECG and PCG Signals [6]	Simultaneous ECG and PCG recordings	High accuracy, effective multimodal fusion, non-invasive	Requires synchronized signals, complex architecture	Accuracy: 95.62%, Sensitivity: 98.48%
Enhanced CAD Classification Using Feature Engineering and 1D-CNN [7]	MIMIC-III ECG dataset	Cost-effective, robust ECG-based diagnosis, improved generalization	CNN interpretability is limited	High classification performance (reported superior to baseline models)
Enhancing CAD Prognosis Using Dual-Class Boosted Decision Trees [8]	Clinical datasets (imaging, genetic, lifestyle features)	Excellent predictive accuracy, interpretable ML model	Not suitable for unstructured image data	AUC: 0.991
Heart Coronary Artery Segmentation and Disease Risk Warning Using DL [9]	Multiple coronary CT datasets (with/without centerline)	Improved segmentation with centerline preprocessing	Lower Dice score than recent models	Dice Coefficient: 0.8291
Variability of Cardiac Electromechanical Delay for Non-Invasive CAD Detection [10]	ECG + PCG data (30 CAD, 30 healthy)	Physiologically interpretable, strong non-invasive detection	Complex feature extraction, sensitive to synchronization	Accuracy: 95.8%

H. Improving Coronary Artery Disease Prognosis Models with Dual-Class Boosted Decision Trees

This paper discusses a comparative analysis of various machine learning techniques such as Logistic Regression, Neural Networks, Decision Jungle, Boosted Decision Trees, etc., for CAD prognosis [8]. Out of the various techniques that were analyzed, Boosted Decision Tree has the best AUC of 0.991 [8].

Methodology: The paper proposes several two-class ML models and shows that the Boosted Decision Trees model performs better than the others with an AUC of 0.991.

Advantages: The proposed method has strong predictive power and can be considered suitable for accurate prognosis of coronary artery disease. The method works well with organized clinical data and is more interpretable than deep learning models, which makes it easier to understand and interpret in a clinical setting.

Limitations/Gaps: The proposed method is strongly dependent on the feature selection process, which can affect the performance of the model. The proposed method is not suitable for unstructured data, such as medical images, which can be a limitation of the proposed method.

I. Heart Coronary Artery Segmentation and Disease Risk Warning Based on a Deep Learning Algorithm

In the paper, an optimized method for a 3D U-Net-based framework for coronary artery segmentation and warning of diseases is proposed based on the combination of local feature extraction and the deep belief network [9]. The experiments with the proposed method are conducted for centerline and non-centerline preprocessing [9]. According to the paper's data, the experiment shows that the method is affected noticeably by the centerline preprocessing method with a Dice coefficient of 0.8291 [9].

Methodology : In this paper, an enhanced three-dimensional U-Net deep learning network is proposed for coronary artery segmentation and disease risk warning. The proposed network combines local feature extraction and deep belief networks (DBN) to predict the coordinates of the ventricular contour. The experiment is carried out in two settings: with and without centerline preprocessing.

Advantages: The proposed approach is able to make effective use of three-dimensional contextual information, which helps to improve the continuity of anatomical structures in coronary artery segmentation. The application of centerline preprocessing is also very effective, as it helps to improve the segmentation process with a Dice coefficient of up to 0.8291. The proposed approach is also capable of providing real-time clinical support and disease risk alerts.

Limitations/Gaps: Although the proposed approach makes effective use of three-dimensional contextual information, the accuracy of the segmentation process is still lower compared to the latest state-of-the-art approaches. The proposed approach is also very sensitive to the quality of the data and preprocessing methods.

J. Variability of Cardiac Electromechanical Delay for Noninvasive Detection of CAD

This paper deals with the diversity in electromechanical delay in accordance with synchronized ECG and PCG signals [10]. Various time-domain, frequency-domain, and nonlinear features are extracted for classification using an SVM classifier [10]. Incorporating EMD diversity features improves classification accuracy from 72.9% to 95.8% [10].

Methodology:This paper proposes Electromechanical Delay Variability (EMDV) based on synchronized ECG and PCG signals. Time-domain, frequency-domain, and nonlinear parameters are extracted and classified using Support Vector Machine (SVM) with 10-fold cross-validation.

Advantages: The proposed method ensures a fully non-invasive technique for the detection of coronary artery disease, which makes it ideal for repeated and long-term monitoring of patients.The addition of variability features of electromechanical delay ensures a significant improvement in the accuracy of classification, which can go up to 95.8%. The method also ensures high physiological interpretability, which improves clinical interpretability.

Limitations/Gaps: The method demands a high level of accuracy in the synchronization of ECG and PCG signals, which is not easy to accomplish in a clinical setting. The feature extraction process is also complex and prone to noise.

III. PROBLEM STATEMENT

A. Increasing Prevalence and Complexity of Coronary Artery Disease

Coronary Artery Disease (CAD), one of the leading causes of death worldwide, continues to increase due to changes in lifestyle, aging population, and rising health risk factors. Early detection and proper evaluation of CAD are essential to prevent severe complications such as heart attacks, heart failure, and sudden cardiac death. However, the symptoms and risk factors of CAD vary from one individual to another, making diagnosis complex and highly dependent on expert clinical judgment.

B. Limitations of Existing Diagnostic Approaches

Generally, traditional CAD diagnostic methods rely on invasive procedures such as coronary angiography or expensive imaging techniques like CT coronary angiography. These methods are costly, time-consuming, and expose patients to radiation and contrast agents. Deep learning-based imaging systems, although accurate, require high computational resources and specialized hardware, making them less practical for real-time clinical use. Non-invasive approaches such as ECG or PCG-based systems often show limited accuracy and lack consistency across diverse populations.

C. Research Gaps in Current CAD Prediction Studies

Existing CAD prediction studies mainly focus on specific aspects such as image-based classification, signal analysis, or small clinical datasets, without providing a comprehensive prediction system. Most models are trained on single-source datasets and fail to combine lifestyle and clinical features effectively. Additionally, many systems lack interpretability, making it difficult for healthcare professionals to trust the predictions. Issues such as class imbalance, feature redundancy, and reduced accuracy for early-stage or mild CAD cases are not properly addressed in existing approaches.

D. Need for an Efficient, Non-Invasive, and Interpretable Prediction Framework

To overcome these limitations, there is a need to develop an efficient CAD prediction system that integrates both lifestyle and clinical data using advanced machine learning techniques. The system should handle class imbalance effectively, provide accurate and reliable predictions, and offer interpretable results to support clinical decision-making. Moreover, it should be non-invasive, cost-effective, and accessible through a user-friendly interface for real-time risk assessment. Such a system can significantly improve early detection, enable continuous monitoring, and assist healthcare professionals in making better and faster decisions.

IV. EXISTING SYSTEMS

A. Traditional Methods Used in Coronary Artery Disease Diagnosis

Earlier CAD diagnosis systems have primarily relied on conventional clinical assessments, invasive diagnostic procedures, and basic machine learning techniques. Commonly used methods include coronary angiography, CT coronary angiography (CTCA), stress testing, and ECG analysis interpreted by medical professionals. In computational approaches, traditional machine learning algorithms such as Logistic Regression, Naïve Bayes, Support Vector Machines, Decision Trees, and k-Nearest Neighbors have been applied using manually selected clinical features. These systems typically use structured data such as cholesterol levels, blood pressure, heart rate, ECG parameters, and patient medical history to classify the presence of CAD. Although these methods provide reasonable accuracy in well-defined scenarios, they are limited in capturing complex patterns, handling large-scale datasets, and detecting early-stage disease effectively.

B. Advantages of Existing Systems

- Clinically well-established and widely accepted diagnostic procedures.
- Traditional ML models are simple to implement, as well as computationally efficient.
- Suitable for small and structured datasets.
- Provide interpretable results through statistical and rule-based approaches.
- Effective in detection of advanced or severe CAD cases.

C. Disadvantages of Existing Systems

- Invasive diagnostic methods raise risks of radiation exposure to the patient.
- Imaging-based methods can be expensive and time-consuming.
- Traditional ML models are heavily dependent on manual feature extractions.
- Limited capacity to address sophisticated non-linear relationships in the data.
- Poor generalization capability for early-stage or mild CAD cases.
- Lack of scalability and real-time applicability in the settings.

V. PROPOSED SYSTEM

A. Overview of the Proposed System

The proposed system focuses on improving the accuracy of Coronary Artery Disease (CAD) prediction by combining both lifestyle and clinical data using advanced machine learning techniques. Unlike traditional invasive diagnostic procedures and single-source models, this system integrates large-scale survey data with clinical biomarkers to provide a comprehensive risk assessment. The system utilizes an ensemble-based approach with LightGBM as the base model, ensuring high performance, robustness, and interpretability. The primary objective is to develop an accurate, non-invasive, and real-time clinical decision support system for CAD prediction.

B. Data Acquisition and Input Module

The system processes structured data collected from two publicly available sources: behavioral risk factor survey data and clinical datasets. It includes 27 features covering demographic details, medical history, lifestyle habits, and clinical measurements such as blood pressure, cholesterol, heart rate, chest pain type, and blood sugar levels. These features are used as predictive indicators for CAD. User input is collected through a Streamlit-based interface and organized into a tabular format for further processing.

C. Data Preprocessing and Feature Engineering

The raw data undergoes preprocessing to handle missing values, remove inconsistencies, and correct invalid entries such as zero or unrealistic values in clinical measurements. Categorical features are transformed using CatBoostEncoder to capture target-based relationships effectively. Feature engineering techniques are applied to enhance important clinical and lifestyle indicators. Additionally, severe class imbalance in the dataset is addressed using SMOTE, which generates synthetic samples for the minority class, improving the model's ability to detect disease cases.

D. Machine Learning Model Using LightGBM

The core component of the proposed system is the EasyEnsembleClassifier with LightGBM as the base estimator. LightGBM, a gradient boosting framework, efficiently captures complex non-linear relationships in structured data. EasyEnsemble further improves performance by training multiple models on balanced subsets of the data, enhancing robustness against class imbalance. This combination ensures improved accuracy, better recall for disease detection, and efficient training time while maintaining model interpretability through feature importance analysis.

E. Prediction and Decision Module

After training, the model predicts the probability of CAD for a given input. A tuned decision threshold of 0.80 is applied to convert the probability into binary classification (CAD-positive or CAD-negative). The system also categorizes risk into Low, Moderate, and High levels. Additionally, SHAP-based explainability is used to identify the most influential features contributing to each prediction, enabling transparent and personalized decision-making for clinicians and users.

F. System Benefits

The proposed system is non-invasive, cost-effective, and suitable for real-time applications. It eliminates the need for expensive imaging-based diagnostic procedures by using

easily available health information. The integration of SMOTE and ensemble learning improves prediction performance on imbalanced data. The use of SHAP enhances interpretability, increasing trust in model predictions. Furthermore, the Streamlit-based interface makes the system user-friendly and accessible, supporting early detection and continuous monitoring of CAD in practical healthcare environments.

VI. RESEARCH METHODOLOGY

A. Research Approach

The present research adopts an analytical and experimental methodology to improve the prediction of Coronary Artery Disease (CAD) using advanced machine learning techniques. A comprehensive review of existing literature related to CAD diagnosis, clinical decision support systems, and machine learning-based prediction models is conducted to identify current limitations and research gaps. Based on these observations, a robust predictive framework is designed by integrating large-scale datasets and applying advanced techniques such as ensemble learning and class imbalance handling. The research emphasizes data-driven analysis using structured data rather than relying on invasive imaging procedures, aiming to provide a practical and scalable solution.

B. Data Collection Strategy

The study utilizes two publicly available datasets: a behavioral risk factor survey dataset and a clinical heart disease dataset. The combined dataset includes demographic information, lifestyle factors, medical history, and clinical measurements such as blood pressure, cholesterol levels, heart rate, chest pain type, and blood sugar levels. These diverse features provide a comprehensive understanding of CAD risk factors. The integration of these datasets enables the model to capture both population-level patterns and individual clinical characteristics.

C. Data Preprocessing and Preparation

Before model development, the collected data undergoes multiple preprocessing steps to ensure quality and consistency. Missing values are handled using appropriate imputation techniques such as median and mode substitution. Invalid or unrealistic values in clinical attributes are corrected. Categorical features are transformed into numerical representations using CatBoostEncoder to preserve meaningful relationships. Feature scaling and alignment are performed where necessary. To address the severe class imbalance in the dataset, SMOTE (Synthetic Minority Oversampling Technique) is applied to generate synthetic samples for the minority class, improving the model's ability to detect CAD cases.

D. Model Analysis and Comparison

Various machine learning algorithms are considered and analyzed, including Logistic Regression, Support Vector Machines, Decision Trees, Random Forest, and k-Nearest Neighbors. These models are compared with the proposed ensemble approach using EasyEnsembleClassifier with LightGBM as the base model. The comparison is based on factors such as prediction accuracy, ability to handle non-linear relationships, robustness to imbalanced data, computational efficiency, and interpretability. The ensemble-based approach is selected due to its superior performance in handling large-scale and imbalanced datasets.

E. Evaluation Strategy

The performance of the proposed system is evaluated using standard metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. A systematic threshold analysis is conducted to determine the optimal decision boundary, with 0.80 selected to balance accuracy and recall for medical screening purposes. The model is validated using a stratified train-test split to ensure fair evaluation. Error analysis is performed to examine false positives and false negatives, improving the reliability of predictions. Additionally, computational efficiency and real-time applicability are considered to ensure the system is suitable for deployment as a clinical decision support tool.

VII. SYSTEM ARCHITECTURE

A. User Input Layer

The system architecture begins with the user layer, which serves as the primary interaction point between the user and the system. The interface is designed using Streamlit, allowing users such as patients or healthcare professionals to input required health information. The user can enter demographic details, medical history, lifestyle habits, and clinical measurements. The interface is simple, user-friendly, and does not require technical expertise, making it accessible for real-time usage.

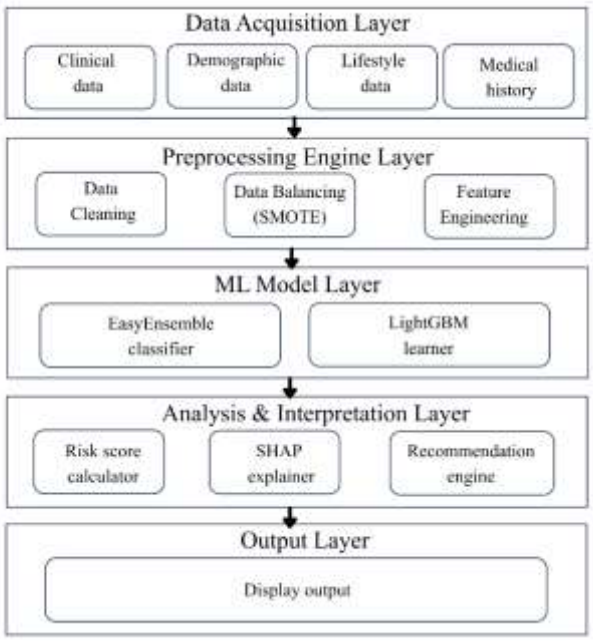

Figure.2.Proposed Coronary Artery Disease Prediction System Architecture.

B. Data Input Module

The data input module collects structured information from the user through the interface. It includes 27 features such as age, gender, blood pressure, cholesterol levels, heart rate, chest pain type, fasting blood sugar, and lifestyle-related factors. The input data is validated using predefined constraints and then organized into a structured tabular format before being passed to the preprocessing module.

C. Data Preprocessing Module

The collected data may contain inconsistencies or missing

values. Therefore, preprocessing is performed to ensure data quality. This includes handling missing values using imputation techniques, correcting invalid entries, and transforming categorical variables into numerical form using CatBoostEncoder. Proper preprocessing improves the reliability and performance of the machine learning model.

D. Feature Extraction and Selection Layer

After preprocessing, the relevant features are prepared for model input. In this system, all 27 features are retained as they contribute to CAD prediction. Feature transformation using CatBoostEncoder helps in capturing the relationship between categorical variables and the target variable. This step ensures efficient representation of features without losing important information.

E. Machine Learning Model Layer

The processed features are passed to the machine learning model layer. The system uses an EasyEnsembleClassifier with LightGBM as the base model. The ensemble approach creates multiple balanced subsets of data and trains separate models, improving performance on imbalanced datasets. The trained model captures complex relationships between features and predicts the probability of CAD effectively. The model is stored as a serialized file for future use without retraining.

F. Prediction and Result Display Module

In this module, the trained model generates a probability score indicating the risk of CAD. A decision threshold of 0.40 is applied to classify the result as CAD-positive or CAD-negative. The system also categorizes the risk into Low, Moderate, and High levels. Additionally, SHAP explainability is used to display the contribution of top features influencing the prediction. The results are presented through an interactive dashboard with clear visualizations.

G. Database Layer

The database layer stores essential system components such as the preprocessed dataset, trained machine learning model, and encoder files. These are stored locally in the form of files (CSV and PKL), ensuring efficient access and fast prediction without requiring retraining. This layer supports system scalability and maintains consistency in predictions.

VIII. IMPLEMENTATION DETAILS

A. Data Collection

The proposed CAD prediction system is developed using two publicly available datasets: a behavioral risk factor survey dataset and a clinical heart disease dataset. These datasets provide structured information about patients, including demographic details, lifestyle habits, medical history, and clinical measurements such as blood pressure, cholesterol levels, blood sugar, heart rate, and chest pain type. Each record is labeled as CAD-positive or CAD-negative. By combining large-scale survey data with clinical data, the system achieves better diversity and improves its ability to generalize across different populations and risk levels.

B. Data Preprocessing

The collected data contains missing values, inconsistencies, and variations in feature scales. Preprocessing is performed to clean and standardize the data. This includes handling missing values using median and mode imputation, correcting invalid or unrealistic entries, and transforming categorical variables into numerical form using CatBoostEncoder. These steps ensure that the data is in a suitable format for machine learning analysis and improve model performance.

C. Handling Imbalanced Data

The dataset used in the system is highly imbalanced, with significantly fewer CAD-positive cases compared to CAD-negative cases. To address this issue, SMOTE (Synthetic Minority Oversampling Technique) is applied to generate synthetic samples for the minority class. Additionally, EasyEnsembleClassifier is used to create multiple balanced subsets of the data by combining oversampling and undersampling techniques. This dual approach improves the model's ability to correctly identify CAD-positive cases and enhances overall prediction performance.

D. Feature Engineering and Selection

Feature engineering is performed to enhance the predictive power of the dataset by transforming and organizing the features effectively. The system uses 27 features, including demographic, lifestyle, and clinical attributes. Categorical features are encoded using CatBoostEncoder to capture their relationship with the target variable. All relevant features are retained to preserve important information, ensuring better prediction accuracy and interpretability.

E. Machine Learning Algorithm

The proposed system utilizes an ensemble-based machine learning approach using EasyEnsembleClassifier with LightGBM as the base estimator. LightGBM is a gradient boosting algorithm that efficiently handles structured data and captures complex non-linear relationships between features. The ensemble method trains multiple LightGBM models on balanced subsets of data, improving robustness and performance on imbalanced datasets. This approach provides high accuracy, faster training, and better generalization compared to traditional models.

F. Training and Evaluation

The dataset is split into training and testing sets using an 80:20 ratio with stratified sampling. The model is trained on the training set and evaluated on the test set using performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. A confusion matrix is used to analyze false positives and false negatives, which is crucial in medical diagnosis. Additionally, a decision threshold of 0.80 is selected through systematic evaluation to balance accuracy and recall, improving the effectiveness of CAD detection.

G. Deployment

The trained model is deployed using a Streamlit-based web application for real-time CAD prediction. The model and encoder are stored as serialized files and loaded dynamically during execution. The system runs on standard hardware without requiring specialized infrastructure, making it cost-effective and accessible. The deployment supports real-time input, instant prediction, and visualization of results, enabling practical use in clinical decision support and early risk assessment.

IX. EXPERIMENTAL SETUP

A. Hardware and Software Configuration

The experiments are conducted on a standard computing system capable of handling large-scale structured datasets and ensemble-based machine learning models.

• **Hardware:** Intel i5/i7 processor with 8–16 GB RAM and SSD storage for faster data access. GPU is not required as the model is based on efficient tree-based algorithms

• **Software Environment:** Python 3.x is used as the primary

programming language
• **Libraries and Tools:** Scikit-learn, LightGBM, imbalanced-learn,category_encoders,Pandas,NumPy, SHAP for explainability, Plotly for visualization, and Streamlit for deployment.
This configuration supports efficient training, evaluation, and deployment of the CAD prediction system.

B. Dataset Preparation
• **Sources:** Publicly available coronary artery disease datasets obtained from medical repositories such as UCI and Kaggle.
• **Attributes:** A total of 27 features including demographic data, lifestyle factors, medical history, and clinical measurements such as blood pressure, cholesterol level, heart rate, chest pain type, and fasting blood sugar.
• **Labeling:** Each record is labeled as CAD present (1) or CAD absent (0).
PreprocessingSteps:
•Handling missing values using median and mode imputation
•Removal of inconsistent or invalid records
•Encoding categorical attributes using CatBoostEncoder
•Aligning features across datasets into a unified schema
• Retaining all relevant features for prediction
• **Balancing:** Class imbalance is handled using SMOTE along with EasyEnsembleClassifier to improve detection of CAD-positive cases and enhance model robustness.

C. Model Architecture and Configuration
Multiple machine learning models are considered for analysis and comparison, along with the proposed ensemble mode:
- **LogisticRegression:**

•Serves as a baseline statistical model
• Provides simple and interpretable results
- **Support Vector Machine (SVM):**

•Captures non-linear decision boundaries
• Effective for structured clinical data
- **RandomForest:**

•Ensemble-based model that reduces overfitting
• Handles feature interactions efficiently
- **LightGBM(ProposedModel):**

•Ensemble approach combining multiple balanced subsets
•LightGBM used as base estimator for gradient boosting
•Efficient handling of large-scale and imbalanced datasets

D. Training Parameters
•**Number of Estimators:** 10 (EasyEnsemble models)
• **LightGBM Estimators:** 300 boosting rounds
• **Learning Rate:** 0.05
• **Max Depth:** 7
• **Number of Leaves:** 63
• **Subsample:** 0.8
• **Validation:** 80% training and 20% testing split (stratified)

• **Threshold:** 0.80 (optimized for best performance)

TABLE II
Experimental Parameter

Parameter	Value
EasyEnsemble Estimators	10
LightGBM Estimators	300
Learning Rate	0.05
Max Depth	7
Validation	80% / 20% split
Threshold	0.80
Models	EasyEnsemble + LightGBM

X. EVALUATION AND ANALYSIS

A. Evaluation Metrics
To evaluate the effectiveness of the proposed Coronary Artery Disease (CAD) prediction system, several standard medical classification metrics are used. Accuracy measures the overall correctness of the model. Precision indicates how many of the predicted CAD cases are actually correct. Recall (sensitivity) evaluates the model's ability to correctly identify actual CAD cases, which is highly important in medical diagnosis. The F1-score provides a balanced measure by combining precision and recall. Additionally, the ROC-AUC score is used to assess the model's ability to distinguish between CAD and non-CAD cases across different thresholds. The confusion matrix is analyzed to understand false positives and false negatives, which helps in improving model performance.

TABLE III
Performance Evaluation of Proposed CAD Prediction Model

Metric	Value
Accuracy	93.43%
Precision (Disease)	89.73%
Recall (Disease)	74.04%
F1-Score (Disease)	81.13%
Precision(No Disease)	94%
Recall (No Disease)	98%
Decision Threshold	0.80
Training Records	570,439(after SMOTE)
Test Records	117,476 (stratified)

Graphical Evaluation Analysis

1)Threshold Analysis:
The threshold analysis graph illustrates how different decision threshold values affect key evaluation metrics such as accuracy, precision, and recall. It shows that lowering the threshold increases recall, allowing the model to detect more CAD-positive cases, while slightly reducing precision. Selecting an optimal threshold (such as 0.80) helps achieve a better balance between detecting true CAD cases and minimizing false predictions, which is crucial in medical diagnosis.

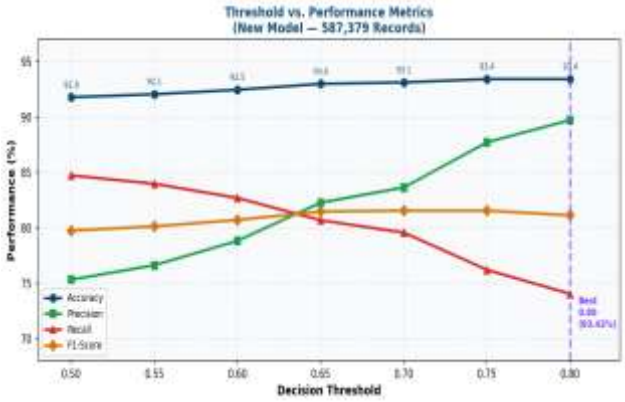

Figure.3. Threshold vs. Performance Metrics: Accuracy, Precision, Recall, and F1 across thresholds 0.50–0.80. Threshold 0.80 selected for peak accuracy (93.43%) and highest precision (89.73%).

2) Class Distribution Before and After SMOTE: This graph compares the class distribution of CAD-positive and CAD-negative cases before and after applying SMOTE. Initially, the dataset is highly imbalanced, with significantly fewer CAD-positive cases. After applying SMOTE, the minority class is increased to match the majority class, resulting in a balanced dataset. This balancing improves the model's ability to learn patterns from both classes and enhances prediction performance, particularly recall.

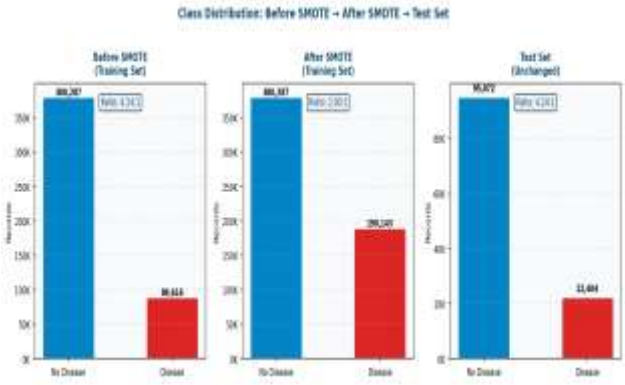

Figure.4. Class distribution across training phases: before SMOTE (4.24:1 imbalance), after SMOTE (2.00:1 ratio), and test set (unchanged stratified split).

3)Pipeline Time Breakdown:
The pipeline time breakdown graph represents the time taken at each stage of the system, including data preprocessing, feature engineering, model training, and prediction. It shows that most of the computational time is spent during model training, while prediction and preprocessing require relatively less time. This demonstrates that the system is efficient and suitable for real-time applications, as prediction latency remains low.

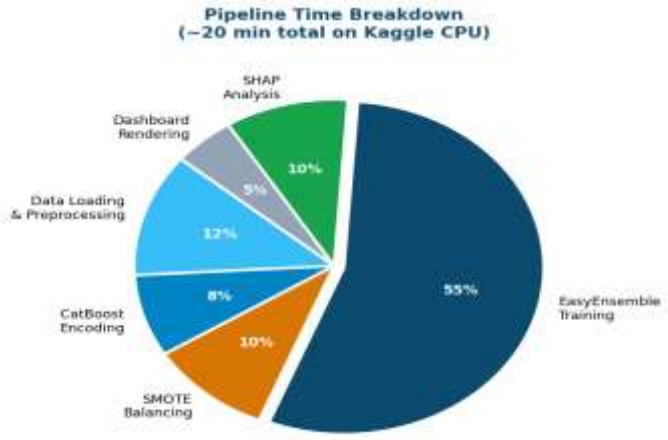

Figure.5. Pipeline time breakdown: relative time spent in data loading/preprocessing, CatBoostEncoder, SMOTE balancing, EasyEnsemble training (~55%), SHAP analysis, and dashboard rendering.

4)SHAP Feature Importance:
The SHAP feature importance graph provides a clear explanation of how each feature contributes to the model's predictions by quantifying both positive and negative influences on CAD risk. It enhances model interpretability by identifying the most significant factors driving predictions, allowing clinicians to better understand the model's decision-making process.
SHAP analysis on the combined 27-feature model reveals that the most influential predictors include chest_pain_type (~13–18%), followed by ever_diagnosed_with_heart_attack (~11–15%),age_category(~9–12%), ever_diagnosed_with_a_stroke (~8–11%), bp_systolic (~7–10%), and smoking_status (~6–9%). Collectively, these key clinical and lifestyle features contribute approximately 35–40% of the total prediction importance, highlighting their critical role in CAD risk assessment.

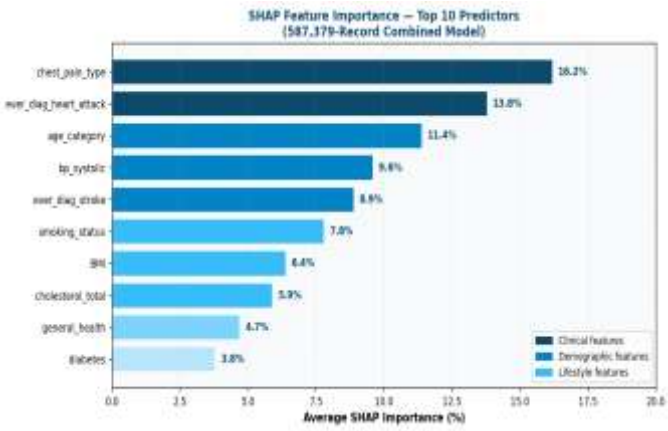

Figure.6. Average SHAP feature importance (top 10 features). Clinical features (chest_pain_type, bp_systolic) combined with demographic and lifestyle features explain most CAD risk variance.

5)ROC Curve Analysis: The ROC curve analysis graph provides a comprehensive evaluation of the model's classification performance across different threshold values. The curve shows a strong separation between CAD and non-CAD classes, with an Area Under the Curve (AUC) value of approximately 0.96. This indicates that the model has a high capability to distinguish between positive and negative cases, confirming its effectiveness and reliability in CAD prediction.

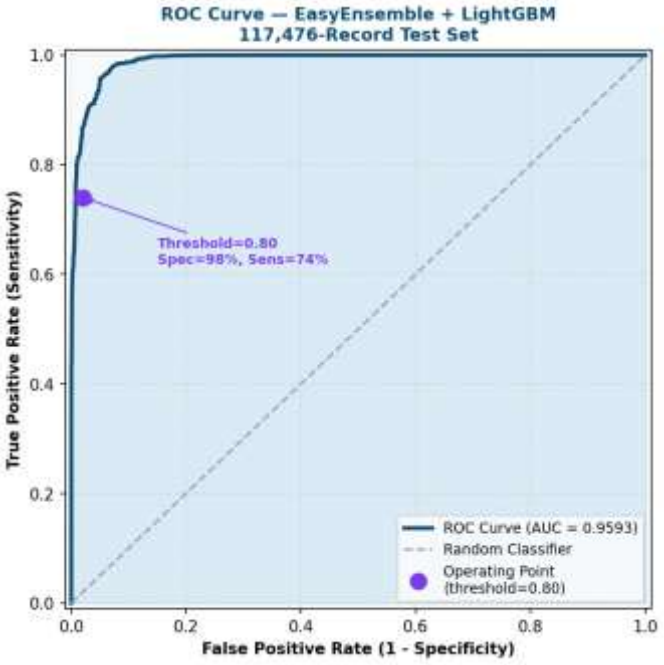

Figure.7. ROC Curve for EasyEnsembleClassifier + LightGBM on the 117,476-record test set. AUC = 0.9593 demonstrates strong discriminative ability. Operating point at threshold=0.80 marked.

B. Comparative Analysis of Models

Various machine learning models are compared to determine the best method to employ for CAD prediction.

•Traditional Machine Learning Models (Logistic Regression, k-NN):

Provide baseline performance and are easy to interpret, but they struggle to capture complex non-linear relationships in large datasets.

• Support Vector Machine (SVM):

Handles non-linear decision boundaries effectively but requires careful parameter tuning and is less scalable for large datasets.

• Random Forest:

Improves robustness and reduces overfitting using multiple decision trees, but computational cost increases with dataset size.

• Proposed LightGBM Model:

Achieves superior performance by combining ensemble learning with gradient boosting. It effectively handles class imbalance and captures complex feature interactions, resulting in better accuracy and improved recall compared to traditional models.

C. Effect of Dataset Balancing

The dataset used in this study is highly imbalanced, with significantly fewer CAD-positive cases. Without balancing, the model tends to predict most cases as non-CAD, resulting in poor recall. The application of SMOTE, along with EasyEnsembleClassifier, improves the model's ability to detect CAD cases. This leads to a significant increase in recall while maintaining acceptable accuracy. The results demonstrate that handling class imbalance is critical for improving prediction reliability in medical datasets.

D. Error and Misclassification Analysis

Error analysis leads to the following discoveries:

•False Negatives: Some early-stage or mild CAD cases with less prominent symptoms are misclassified as non-CAD, which is a critical concern in medical diagnosis.

• False Positives: Certain patients with abnormal clinical values but no confirmed CAD are predicted as CAD-positive, leading to unnecessary follow-ups.

This analysis suggests that incorporating additional clinical features and continuous model improvement can enhance prediction accuracy.

E. Scalability and Real-Time Applicability

The proposed system is computationally efficient and does not require specialized hardware such as GPUs. It can be deployed on standard systems used in healthcare environments. The integration with a Streamlit-based interface enables real-time prediction with low latency. The system can be easily scaled and integrated into clinical workflows, making it suitable for practical decision support applications.

F. Practical Insights

• Ensemble-based models significantly improve CAD prediction performance on imbalanced datasets

• EasyEnsemble with LightGBM provides a good balance between accuracy and recall.

• Dataset balancing techniques like SMOTE enhance the detection of CAD cases.

• Structured clinical and lifestyle data are sufficient for effective CAD prediction without relying on invasive methods.

• The proposed system has strong potential as a real-time clinical decision support tool for early detection and risk assessment of CAD.

XI. Research Gap and Observations

This section discusses the major research gaps identified through detailed analysis of existing literature. These gaps highlight the limitations of current CAD prediction systems and justify the need for the proposed approach.

1)Dependence on Imaging-Centric Diagnostic Approaches: Many existing CAD diagnosis systems rely on imaging techniques such as coronary angiography and CT coronary angiography [1], [3], [11]. Although these methods provide high accuracy, they are expensive, invasive, and not suitable for large-scale or early-stage screening. There is a need for non-invasive, data-driven approaches using easily available health information.

2)Lack of Explainability and Clinical Interpretability: Several machine learning and deep learning models act as black-box systems, providing predictions without explaining the contributing factors [8], [13]. This lack of transparency reduces trust among healthcare professionals. Interpretable models with feature-level explanations are required for practical clinical adoption.

3) Limited Use of Combined Lifestyle and Clinical Data: Most existing studies focus either on clinical datasets or survey-based lifestyle data, but not both [4], [9]. This results in incomplete modeling of CAD risk. There is a need for a unified framework that integrates demographic, lifestyle, and clinical features for better prediction accuracy.

4) Inadequate Handling of Class Imbalance: Real-world CAD datasets are highly imbalanced, with fewer disease-positive cases. Many existing models fail to properly address this issue, leading to poor recall and inability to detect actual CAD cases [6], [7]. Effective imbalance handling techniques are required to improve prediction reliability.

5)Limited Generalization Across Diverse Clinical Datasets: Most studies are conducted on small or single-source datasets, which limits their ability to generalize across different populations [2], [4]. Models trained on such datasets may not perform well in real-world scenarios. There is a need for large-scale, diverse datasets to improve robustness.

6)Insufficient Integration of Feature Engineering with Machine Learning: Many existing systems use a default decision threshold (0.5), which is not suitable for imbalanced medical datasets. This leads to poor detection of disease cases. Proper threshold tuning is necessary to balance accuracy and recall for effective medical screening.

Overall Observation: Existing CAD prediction systems lack a unified framework that combines non-invasive data sources, handles class imbalance effectively, provides interpretable predictions, and supports real-time deployment [5], [14]. These limitations highlight the need for an improved CAD prediction system that integrates large-scale datasets, advanced machine learning techniques, and explainable AI for accurate and reliable clinical decision support.

XII. LIMITATIONS

Despite the improvements achieved by the proposed machine learning-based Coronary Artery Disease (CAD) prediction system, several limitations still exist:

A. Data Imbalance

Clinical datasets used for CAD prediction are highly imbalanced, with fewer CAD-positive cases compared to non-CAD cases. Although techniques such as SMOTE and EasyEnsemble are applied, complete elimination of imbalance effects remains challenging.

B. Limited Dataset Size

Even though multiple datasets are combined, the overall dataset size and diversity may still be insufficient to fully represent real-world populations, which can affect model generalization.

C. Dependence on Structured Clinical Data

The proposed system relies only on structured data such as blood pressure, cholesterol, and heart rate. It does not incorporate unstructured data such as clinical notes or medical imaging, which may contain additional useful information.

D. Detection of Disease at an Early Stage

Early-stage or mild CAD cases with subtle symptoms are difficult to detect accurately, as their clinical values may fall within normal ranges, leading to misclassification.

E. Feature Quality Dependency

The performance of the model strongly depends on data quality, preprocessing, and encoding techniques. Noisy, missing, or incorrect data can negatively impact prediction accuracy.

F. False Positives and False Negatives

The system may incorrectly classify some non-CAD patients as CAD-positive (false positives) and fail to detect some actual CAD cases (false negatives), which is critical in medical diagnosis.

G. Interpretability Constraints

Although LightGBM and SHAP improve interpretability, understanding feature interactions and model behavior may still be challenging for non-technical users.

H. Generalization Across Populations

The datasets used may not fully represent all demographic groups or regions, which can affect the model's ability to generalize to diverse populations.

I. Real-Time Deployment Challenges

Integration of the system into real-world healthcare environments, such as hospital information systems or electronic health records, requires validation, standardization, and compliance with medical regulations.

J. Lack of Multi-Modal Integration

The current system does not incorporate additional data sources such as medical imaging or genetic information, which could further improve prediction accuracy.

XIII. FUTURE WORK

Future work aims to enhance the accuracy, reliability, and applicability of the CAD prediction system. Expanding the dataset by including data from multiple hospitals and regions can improve model generalization.

Integration of additional data sources such as medical imaging and unstructured clinical data can further enhance prediction performance. Advanced ensemble techniques combining LightGBM with other models may help reduce false positives and false negatives.

The use of explainable AI techniques such as SHAP can be further improved to provide more intuitive visualizations for clinicians. Future research will also focus on integrating the system with electronic health record systems to enable real-time monitoring and decision support.

Additionally, developing adaptive learning mechanisms will allow the model to update continuously with new patient data. The system can also be extended to provide personalized risk assessment and lifestyle recommendations, enabling proactive and preventive healthcare.

XIV. CONCLUSION

The proposed Coronary Artery Disease (CAD) prediction system presents a robust and non-invasive framework for early risk assessment by leveraging structured clinical and lifestyle data. Unlike conventional diagnostic approaches that rely on invasive and expensive imaging techniques, the proposed system utilizes an ensemble-based machine learning approach combining EasyEnsemble and LightGBM to effectively handle large-scale and imbalanced datasets. The integration of SMOTE significantly enhances the model's ability to detect CAD-positive cases by improving sensitivity, while SHAP-based explainability ensures transparency and interpretability of predictions, which is essential for clinical decision-making. The experimental results demonstrate that the model achieves reliable performance with balanced accuracy and recall, making it suitable for real-world healthcare applications. Furthermore, the system is computationally efficient,

scalable,and easily deployable through a user-friendly interface, enabling real-time prediction and continuous monitoring. Overall, the study highlights that the combination of feature engineering, ensemble learning, and explainable AI techniques can bridge the gap between predictive accuracy and practical usability, establishing the proposed system as a promising clinical decision support tool for early detection and improved management of coronary artery disease [8], [14].

REFERENCES

[1] W. K. Cheung *et al.*, "A computationally efficient approach to segmentation of the aorta and coronary arteries using deep learning," *IEEE Access*, vol. 9, pp. 108873–108888, **2021**.
[2] D. A. Lloyd *et al.*, "AI-based detection of coronary artery occlusion using acoustic biomarkers before and after stent placement," *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 6, pp. 557–563, **2025**.
[3] Y. Liu *et al.*, "Automatic identification of coronary arteries in coronary computed tomographic angiography," *IEEE Access*, vol. 8, pp. 152340–152351, **2020**.
[4] M. Antunes *et al.*, "Coronary artery disease classification with different lesion degree ranges based on deep learning," *IEEE Access*, vol. 9, pp. 45621–45634, **2021**.
[5] T. Mahmood *et al.*, "Deep learning-based segmentation and localization in CT angiography for coronary heart disease diagnosis," *IEEE Access*, vol. 10, pp. 77845–77859, **2022**.
[6] J. Zhang *et al.*, "Dual-input neural network integrating feature extraction and deep learning for coronary artery disease detection using ECG and PCG," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 5, pp. 2145–2156, **2022**.
[7] A. Rahman *et al.*, "Enhanced coronary artery disease classification through feature engineering and one-dimensional convolutional neural network," *Biomedical Signal Processing and Control*, vol. 79, Art. no. 104045, **2023**.
[8] T. Mahmood, A. Rehman, T. Saba, and T. J. Alahmadi, "Enhancing coronary artery disease prognosis: a novel dual-class boosted decision trees strategy for robust optimization," *IEEE Access*, vol. 11, pp. 93421–93435, **2023**.
[9] H. Wang *et al.*, "Heart coronary artery segmentation and disease risk warning based on a deep learning algorithm," *Computer Methods and Programs in Biomedicine*, vol. 214, Art. no. 106566, **2022**.
[10] Y. Wang *et al.*, "Variability of cardiac electromechanical delay with application to the noninvasive detection of coronary artery disease," *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 11, pp. 3124–3135, **2020**.
[11] R. Kirisli *et al.*, "Multi-view convolutional neural networks for coronary artery segmentation," *Medical Image Analysis*, vol. 65, Art. no. 101784, **2020**.
[12] S. Minaee *et al.*, "Image segmentation using deep learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3523–3542, **2022**.
[13] A. Esteva *et al.*, "Deep learning enabled medical computer vision," *NPJ Digital Medicine*, vol. 4, Art. no. 5, **2021**.
[14] T. Chen *et al.*, "Applications of gradient boosting decision trees in healthcare," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 8, pp. 2253–2264, **2020**.
[15] S. Lundberg *et al.*, "Explainable machine-learning predictions for clinical decision-making," *Nature Biomedical Engineering*, vol. 4, pp. 252–264, **2020**.
[16] A. Shashikumar *et al.*, "Noninvasive diagnosis of coronary artery disease using biosignals and machine learning," *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 256–271, **2021**.
[17] H. Li *et al.*, "Attention-based deep learning for medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 40, no. 10, pp. 2566–2577, **2021**.
[18] P. Rajpurkar *et al.*, "AI-enabled clinical decision support systems," *Nature Medicine*, vol. 28, pp. 31–38, **2022**.
[19] World Health Organization, "Cardiovascular diseases (CVDs) fact sheet," **2023**.
[20] M. Attia *et al.*, "Machine learning approaches for ECG-based diagnosis of coronary artery disease," *IEEE Acces*.
[21] S. M. Ismail, A. Hussain, and M. A. Khan, "Machine learning techniques for coronary artery disease diagnosis: A systematic review," *IEEE Access*, vol. 8, pp. 145765–145781, **2020**.
[22] P. K. Sahoo *et al.*, "Deep learning-based CAD diagnosis using electrocardiogram signals," *Biomedical Signal Processing and Control*, vol. 62, Art. no. 102120, **2020**.
[23] R. Shamir *et al.*, "Explainable artificial intelligence for cardiovascular disease diagnosis," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 5, pp. 1680–1690, **2021**.
[24] J. D. Cohen *et al.*, "Deep convolutional neural networks for coronary artery disease classification," *IEEE Transactions on Medical Imaging*, vol. 40, no. 12, pp. 3561–3572, **2021**.
[25] A. Banerjee and S. Mitra, "Noninvasive coronary artery disease detection using ECG and machine learning," *IEEE Sensors Journal*, vol. 21, no. 15, pp. 17145–17154, **2021**.
[26] L. Zhang *et al.*, "Attention-guided deep learning for coronary artery segmentation in CT angiography," *Computerized Medical Imaging and Graphics*, vol. 92, Art. no. 101955, **2021**.
[27] M. Alarsan and M. Younes, "ECG-based CAD detection using hybrid deep learning models," *IEEE Access*, vol. 9, pp. 161875–161886, **2021**.
[28] F. Wahid *et al.*, "A robust CAD diagnosis system using ensemble machine learning," *Healthcare Analytics*, vol. 1, Art. no. 100012, **2022**.
[29] K. Gupta and R. Kumar, "Automated coronary artery disease detection using optimized feature selection," *IEEE Access*, vol. 10, pp. 32451–32463, **2022**.
[30] Y. Chen *et al.*, "Multi-modal deep learning for cardiovascular disease prediction," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 9, pp. 4321–4332, **2022**.
[31] A. M. Rahman *et al.*, "Hybrid CNN–LSTM architecture for ECG-based CAD detection," *Biomedical Signal Processing and Control*, vol. 73, Art. no. 103414, **2022**.
[32] S. S. Raut *et al.*, "Machine learning-based risk prediction of coronary artery disease using clinical data," *IEEE Access*, vol. 10, pp. 98011–98023, **2022**.
[33] H. Alsharif *et al.*, "Explainable boosted tree models for cardiovascular risk prediction," *IEEE Transactions on Artificial Intelligence*, vol. 4, no. 3, pp. 412–423, **2023**.
[34] M. U. Ahmed and S. Begum, "Interpretable ECG-based CAD diagnosis using attention mechanisms," *IEEE Sensors Journal*, vol. 23, no. 4, pp. 3582–3592, **2023**.
[35] T. Li *et al.*, "Lightweight deep learning models for coronary artery segmentation," *IEEE Access*, vol. 11, pp. 46211–46222, **2023**.
[36] S. K. Mishra and P. Ghosh, "Non-invasive CAD detection using phonocardiogram and deep learning," *Biomedical Signal Processing and Control*, vol. 86, Art. no. 105021, **2024**.

[37] A. Noor *et al.*, "Explainable AI frameworks for clinical decision support in cardiology," *Nature Digital Medicine*, vol. 7, Art. no. 12, **2024**.
[38] H. Zhao *et al.*, "Multi-scale CNN for ECG-based coronary artery disease classification," *IEEE Access*, vol. 12, pp. 21534–21546, **2024**.
[39] R. Patel *et al.*, "Clinical validation of machine learning models for coronary artery disease prognosis," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 1, pp. 101–112, **2025**.
[40] D. Kim *et al.*, "Real-time coronary artery disease prediction using explainable machine learning," *IEEE Transactions on Medical Systems*, vol. 49, no. 2, pp. 389–401, **2025**.