Venkata Krishna Kandalai
HIFX Independent Study

A Review of Coronary Artery Disease Risk Factors and Predictive Modeling Using Machine Learning Approaches

Introduction

With about one-third of all fatalities worldwide, cardiovascular disease (CVD) ranks as the main cause of mortality globally, killing almost 17.9 million persons annually. Four-fifths of these deaths are due to heart attacks and strokes, and one-third occur prematurely, before the age of 70 years (WHO). In the United States alone, 941,652 deaths in 2022 were attributed to CVD, more lethal than cancer and chronic lower respiratory disease combined (AHA 2025). The financial burden is no less overwhelming: direct and indirect expenditures of CVD exceeded $417 billion in 2020‑2021(AHA 2025). Of the several types of CVD, coronary artery disease (CAD), epicardial coronary artery stenosis brought on by atherosclerosis, takes a disproportionate share. CAD by itself was responsible for 371,506 U.S. deaths in 2022 and accounts for 39.5 % of all CVD-attributable deaths (AHA 2025).

Since CAD includes numerous symptoms that are common to other CVD entities (e.g., secondary vascular remodeling from hypertension, lipid buildup, endothelial dysfunction from diabetes), CAD is frequently used as a primary phenotype to investigate, and ultimately prevent, cardiovascular events. Epidemiology, over decades, has also developed a consistent set of modifiable risk factors such as smoking, unhealthy diet, physical inactivity, unhealthy alcohol consumption, obesity, hyperlipidemia, hypertension, and hyperglycemia, along with non-modifiable determinants such as age, sex, race/ethnicity, and socioeconomic status (WHO). These determinants are the foundation of public health frameworks, e.g., the American Heart Association's "Life's Essential 8"; however, they do not account for individual-level heterogeneity in risk for events. Accordingly, both policymakers and clinicians are increasingly demanding data-driven strategies for more detailed stratification.

Machine learning (ML) methodologies have shown considerable promise in this area. Recent studies have reported area under receiver operating characteristic (AUROC) values exceeding 0.95 when ensemble or AutoML pipelines are applied to curated CAD datasets and evaluated using techniques such as SHAP (Wang et al. 2024). These advancements suggest that adaptable algorithms paired with transparent feature attribution might outperform traditional risk-scoring systems without sacrificing clinical interpretability.

This paper aims to (i) synthesize current literature on demographic, lifestyle, clinical, and psychosocial drivers of CVD, and (ii) demonstrate a reproducible ML workflow that predicts binary CAD outcomes as a practical example for the broader CVD domain. By explicitly positioning CAD within the CVD classification system, and by discussing the transferability of model insights, I hope to bridge the gap between focused, high-quality datasets and the broader goal of comprehensive CVD risk prediction.

Background

CVD remains the single largest cause of mortality and is an umbrella term for a spectrum of disorders that affect the heart and blood vessels and remains the world’s leading killer. Within this broad category are several clinically distinct yet pathophysiologically overlapping entities. Coronary‑artery disease (CAD), characterized by atherosclerotic narrowing of the epicardial coronary arteries, accounts for roughly 702,880 U.S. deaths each year and almost 40 % of all CVD mortality nationally (CDC). Stroke, a blocked vessel in the brain, affects 93.8 million people worldwide, with nearly 12 million new events reported in 2021 (Feigin et al. 2025). Heart failure (HF), the failure of the ventricle, affects about 6.7 million U.S. adults and is projected to exceed 8 million by 2030 (AHA). Peripheral-artery disease (PAD), involving the narrowing of lower-extremity arteries, is present in more than 6.5 million Americans aged ≥ 40 years and up to 15 % of adults over 80 worldwide. Although their clinical presentations differ, these conditions share enough biology and risk-factor overlap to be best viewed as interrelated facets of the same public health challenge. Although their clinical presentations differ, these conditions share enough biology and risk-factor overlap that they are best viewed as interrelated facets of the same public health challenge.

The predominant pathway across heart disease types is atherosclerosis, an inflammatory process triggered by damage to artery walls and lipid buildup that gradually narrows blood vessels. Various risk factors worsen this process throughout the body. Hypertension damages the endothelium and destabilizes plaques; these effects get amplified with diabetes, where high blood sugar levels create oxidative stress and vascular inflammation (Mathur 2014). High LDL-cholesterol has a correlation to plaque growth and serves as a major contributor to cardiovascular disease burden globally, alongside high blood pressure, smoking, and elevated fasting glucose, as shown in the Global Burden of Disease 2019 analysis (Mansouri et al. 2019). Importantly, there's no safe level of tobacco exposure: smoking just one cigarette daily carries about half the heart attack and stroke risk seen in pack-a-day smokers. Since these factors act systemically, people often develop multiple cardiovascular conditions over their lifetime, supporting the approach of studying any single subtype, especially coronary artery disease, as a meaningful indicator of overall cardiovascular risk (Gallucci et al. 2020).

CAD works as a key indicator condition for several reasons. First, it's numerically dominant, causing more deaths than any other single cardiovascular disease in wealthy countries, including the US where it led to over 370,000 deaths in 2022 (CDC). Second, CAD diagnosis relies on objective criteria: invasive coronary angiography remains the definitive test for identifying arterial blockages, while coronary CT angiography provides a non-invasive alternative with similar accuracy (Detrano et al. 1989). This diagnostic clarity reduces outcome ambiguity compared to conditions like heart failure with preserved ejection fraction. Third, there's a wealth of publicly accessible CAD datasets like the well-known Cleveland dataset from the UCI Machine Learning Repository that enable reproducible modeling and benchmarking. Recent studies using these resources have developed machine-learning approaches that achieve AUROC values of 0.95 or higher and offer interpretable features of importance through techniques like SHAP analysis. The combination of prevalence, diagnostic certainty, and data availability makes CAD an excellent starting point for developing methods that can later be applied to prediction tasks across the entire cardiovascular disease spectrum.

Related Work

Early machine learning approaches aimed to improve or supplement the Framingham Risk Score. Narain et al. developed a quantum neural network using FRS variables that demonstrated superior discrimination compared to the original score across 5,209 Framingham participants (Narain et al. 2016). Meta-analyses have confirmed that tree-based and boosting algorithms consistently outperform traditional scoring systems, with pooled AUCs of approximately 0.86 versus 0.76 for conventional tools (Krittanawong 2020).

To overcome the racial limitations and restricted applicability of the 2013/2018 Pooled Cohort Equations, Ward et al. implemented gradient-boosting and XGBoost models using 262,923 multi-ethnic electronic health records; their gradient-boosting method increased the AUC from 0.78 (PCE) to 0.84 and enabled risk estimation for patients ineligible under PCE criteria. Similarly, random-forest models applied to a 30,000-patient high-risk Chinese cohort outperformed multivariate regression (AUC 0.79 vs 0.71) and identified previously unrecognized interactions among 30 common risk factors (Yang et al. 2020).

Moving beyond tabular clinical data, imaging, and continuous monitoring data are emerging in cardiovascular risk prediction. A multi-center study from Cedars-Sinai employed deep learning on coronary CT angiography to quantify plaque burden; these automated measurements correlated with intravascular ultrasound findings and predicted future myocardial infarction. Other research groups have developed convolutional neural networks capable of estimating 10-year cardiovascular risk from a single chest radiograph or retinal photograph, often performing comparably to blood-based risk calculators, though most remain in experimental stages (Lin 2022).

Benchmark tabular datasets remain central to early-stage CAD prediction research. The UCI Cleveland Heart Disease dataset (comprising 14 features from 303 patients) has generated hundreds of studies, even basic logistic regression achieves 85-90% accuracy, while optimized KNN and random-forest models surpass 95% accuracy on internal data splits. Recent publications claim accuracy exceeding 98% following class balancing and feature selection procedures, though these impressive metrics depend on limited sample sizes and often reflect overly optimistic validation approaches. Additionally, researchers have developed ML markers directly from ECG recordings, stress testing, and imaging studies. A recent study trained gradient-boosting algorithms on clinical variables combined with automatically extracted ECG features, achieving an AUROC of 0.93 on an external hospital validation cohort (Forrest 2024). Schöbel et al. integrated clinical variables with machine learning-derived stress-ECG parameters, outperforming expert cardiologists in detecting functionally significant CAD (AUROC 0.71 vs 0.65) ((Forrest 2024).

Machine learning has transformed cardiovascular risk prediction by advancing from enhanced traditional scores to sophisticated models integrating diverse clinical data streams. While performance metrics appear promising, particularly in models combining clinical variables with physiological signals, significant challenges remain regarding external validation and clinical implementation.

Methods

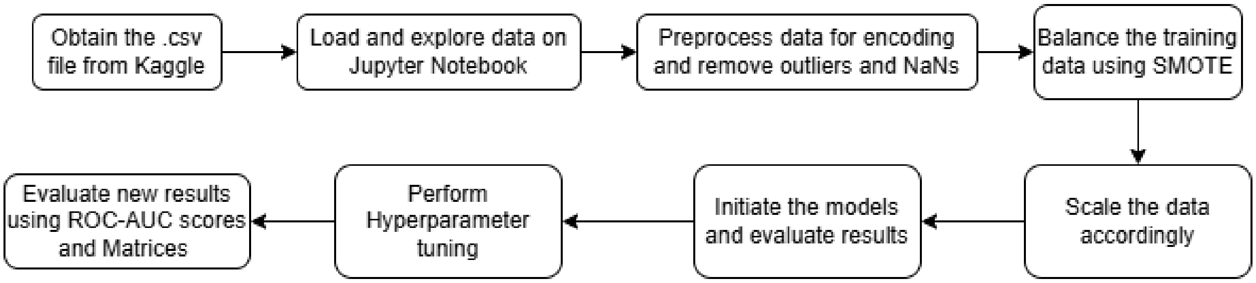

Obtain the .csv file from Kaggle
Load and explore data on Jupyter Notebook
Preprocess data for encoding and remove outliers and NaNs
Balance the training data using SMOTE
Scale the data accordingly
Initiate the models and evaluate results
Perform Hyperparameter tuning
Evaluate new results using ROC-AUC scores and Matrices

Fig 1. Workflow of the project

Figure 1 illustrates the overall workflow for developing my CAD prediction system. The process began with dataset collection and was followed by plotting some graphs to observe the relationship between the target variable and other features in the dataset. Preprocessing steps to ensure consistency and reliability. This included replacing biologically implausible values (e.g., zero cholesterol or resting BP) with statistically appropriate imputations, encoding categorical features, and scaling numerical ones where necessary. I then divided the cleaned dataset into training and testing sets (75:25 split) and applied several machine learning algorithms to evaluate which provided the most effective prediction performance.

The dataset used in this study is a consolidated version of the classic UCI Heart Disease dataset, available from Kaggle. It merges patient records from four locations: the Cleveland Clinic Foundation, the Hungarian Institute of Cardiology, University Hospital Zurich (Switzerland), and Long Beach VAMC (California). While the original dataset includes 76 features, most published work, including this study, focuses on a standard subset of 14 variables selected for their clinical relevance and data availability. The dataset contains 1109 records, and the target variable is a binary indicator: 0 indicating no presence of CAD and 1 indicating the presence of CAD. Although the original data were collected between 1981 and 1984, they remain one of the most widely used benchmarks for machine learning in clinical prediction tasks. The dataset has since been anonymized and cleaned for educational and research use.

The original data collection included patients referred for coronary angiography based on clinical suspicion of CAD. Standard diagnostic evaluations included medical history, physical examination, ECG, lab tests (cholesterol, fasting glucose), stress testing, and imaging (e.g., thallium scans and fluoroscopy). The merged dataset used here does not distinguish records by site but instead pools all 1109 patient entries into a single table. No explicit exclusion criteria are documented in the Kaggle version, but incomplete rows were removed during preprocessing. The dataset comprised 12 features commonly employed in coronary artery disease prediction, including both numerical and categorical variables. Numerical predictors consisted of patient age (in years), resting blood pressure (in mm Hg), serum cholesterol (in mg/dL), maximum heart rate achieved during exercise, and ST depression induced by exercise relative to rest (oldpeak). Categorical variables included sex (binary: 0 = female, 1 = male), chest pain type (ordinal: 1 to 4), fasting blood sugar (binary: 0 = ≤120 mg/dL, 1 = >120 mg/dL), resting electrocardiogram results (nominal: 0 = normal, 1 = ST-T abnormality, 2 = probable left ventricular hypertrophy), exercise-induced angina (binary: 0 = no, 1 = yes), and ST segment slope during exercise (ordinal: 1 = upsloping, 2 = flat, 3 = downsloping).

Records with biologically implausible values (e.g., zero for cholesterol or BP) were imputed using median values. Cholesterol outliers above 500 were capped. Categorical variables were encoded (one-hot for multi-class; binary retained), and numerical features were scaled with StandardScaler. Duplicate rows were removed. Due to moderate class imbalance (~53% CAD), SMOTE was applied to balance the training set. Although the dataset specifically targets coronary artery disease (CAD), this condition shares key risk factors, such as hypertension, diabetes, smoking, and high cholesterol, with other CVD subtypes like stroke or heart failure. As a result, CAD serves as a practical and well-characterized proxy for exploring broader CVD prediction using machine learning.

Given that the dataset had a slight class imbalance (53% positive, 47% negative cases), we implemented SMOTE (Synthetic Minority Oversampling Technique) to rebalance the training set. SMOTE generates synthetic samples of the minority class by interpolating between existing observations, which helps prevent model bias toward the majority class and improves sensitivity in detecting CAD. This strategy was particularly important for ensuring that recall, the ability to correctly identify patients with CAD, remained high.

Model Implementation and Hyperparameter Tuning

A variety of machine learning models were employed in this study to explore different strengths in predictive performance and interpretability. Logistic Regression was used as a linear baseline model, offering high interpretability through its coefficients, which can be directly translated into odds ratios, making it especially useful for clinical interpretation. A single Decision Tree was included to capture non-linear relationships through rule-based splits, though its tendency to overfit necessitates cautious evaluation. To address this limitation, Random Forest aggregates multiple bootstrapped decision trees, enhancing predictive stability and providing reliable feature importance measures. AdaBoost improves classification by sequentially adjusting weights on misclassified samples, creating a streamlined ensemble well-suited to moderately noisy datasets. Gradient Boosting builds on this by iteratively fitting trees to the residuals of previous models, effectively capturing complex feature interactions while requiring thoughtful tuning to avoid overfitting. XGBoost further enhances this framework by incorporating second-order gradient information, regularization techniques, and efficient parallel processing, making it one of the most powerful algorithms for structured tabular data like clinical records.

Hyperparameter tuning was performed using RandomizedSearchCV, a resource-efficient method ideal for the 8 GB RAM environment used in this study. Each model underwent 5-fold cross-validation on the SMOTE-balanced training set, with mean ROC–AUC as the evaluation metric. For Random Forest, key parameters included n_estimators (50–200), max_depth (None, 10, 20, 30), and max_features (sqrt, log2, all). Boosting models (AdaBoost, Gradient Boosting, XGBoost) were tuned over learning rates (0.01–0.2), subsample (0.8–1.0), and max_depth (3–7). XGBoost also included gamma (0–0.2). Logistic Regression was optimized for penalty (l1, l2, elastic-net), regularization strength (C from 0.001–10), and optionally class_weight="balanced". This approach allowed efficient exploration of parameter space while minimizing overfitting and ensuring fair model comparison.

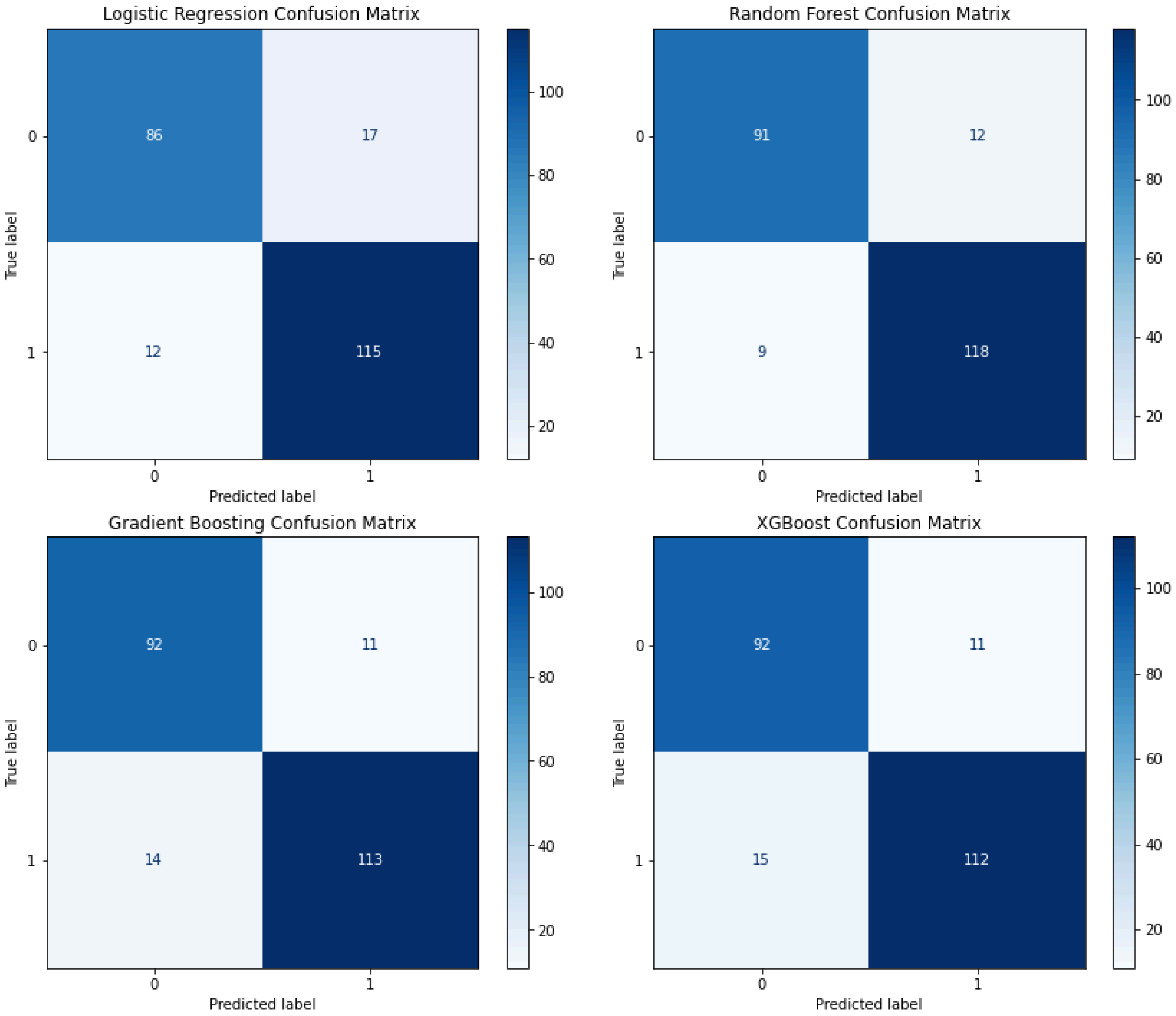
Results

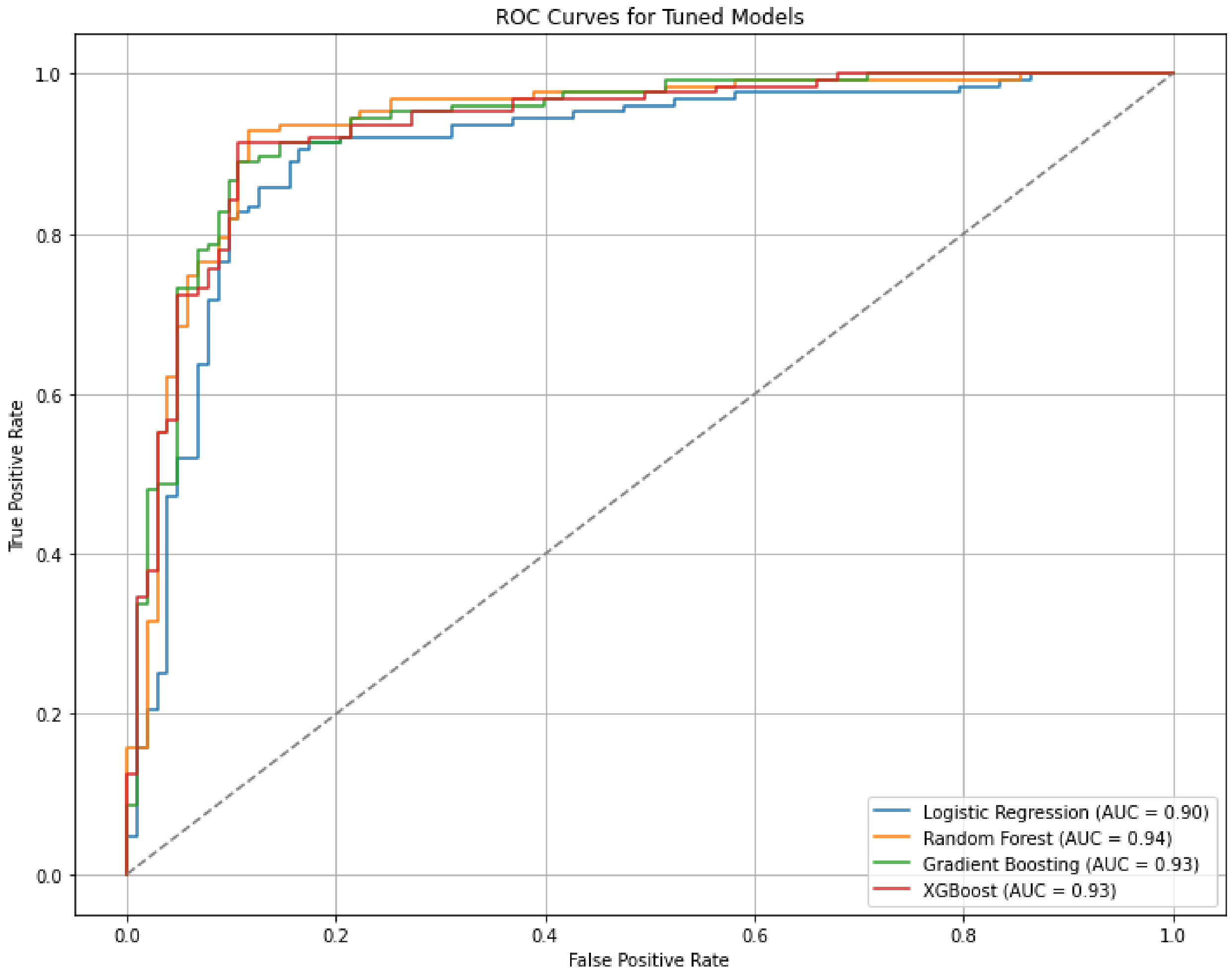
The final dataset utilized in this study comprised 1,109 patient records, with 53% (n = 586) classified as having coronary artery disease (CAD) and 47% (n = 523) without. The mean patient age was approximately 54 years, with a male predominance (~69%). Among CAD-positive cases, elevated cholesterol levels, reduced maximum heart rate, increased ST depression (oldpeak), and higher incidence of exercise-induced angina were frequently observed. These patterns align with established clinical understanding of CAD risk factors, supporting the dataset's validity for model development.

The experiment evaluated six classification models: Logistic Regression, Decision Tree, Random Forest, AdaBoost, Gradient Boosting, and XGBoost. Table 1 summarizes the key performance metrics for all models before and after hyperparameter tuning. Initial testing revealed that Random Forest and Gradient Boosting outperformed simpler models, achieving accuracy and F1-scores approaching 90%. Logistic Regression demonstrated strong recall (0.91), establishing it as a viable baseline for clinical application.

Model	**Accuracy**	**Recall**	**F1 Score**
Logistic Regression	0.88	0.91	0.89
Decision Tree	0.82	0.80	0.83
AdaBoost	0.87	0.87	0.88
Random Forest	0.90	0.92	0.91
Gradient Boosting	0.90	0.90	0.90
XGBoost	0.88	0.86	0.89

Table 1. Model metrics that were calculated before performing hyperparameter tuning.


Logistic Regression Confusion Matrix
True label
Predicted label
0
1
0
1
86
17
12
115
20
40
60
80
100
Random Forest Confusion Matrix
True label
Predicted label
0
1
0
1
91
12
9
118
20
40
60
80
100
Gradient Boosting Confusion Matrix
True label
Predicted label
0
1
0
1
92
11
14
113
20
40
60
80
100
XGBoost Confusion Matrix
True label
Predicted label
0
1
0
1
92
11
15
112
20
40
60
80
100

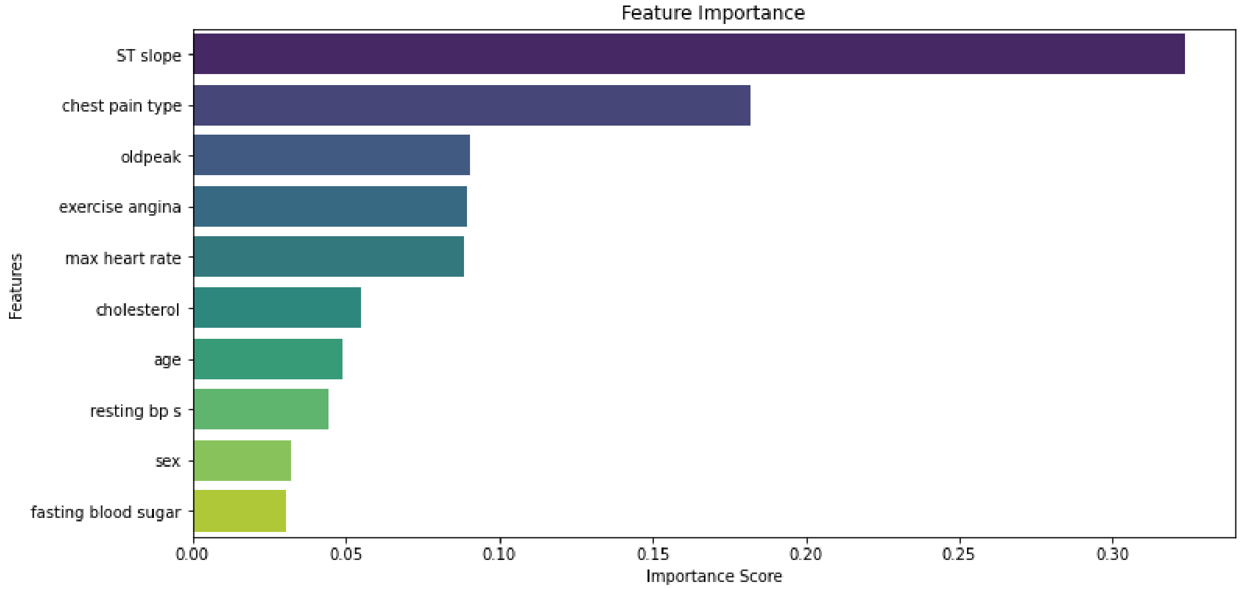


Following hyperparameter optimization, all models demonstrated improved ROC-AUC and recall scores, with Random Forest achieving the highest ROC-AUC (0.9353), followed closely by Gradient Boosting (0.9346) and XGBoost (0.9321). These results underscore the effectiveness of ensemble tree-based classifiers when properly optimized. Logistic Regression also showed improvement, reaching an AUC of 0.9037 while maintaining high recall (0.91), suggesting robust generalization capacity despite its linear architecture.

Model	**Accuracy**	**Recall**	**F1 Score**	**ROC–AUC**
Logistic Regression	0.87	0.91	0.89	0.9037
Random Forest	0.89	0.92	0.90	0.9353
Gradient Boosting	0.87	0.90	0.88	0.9346
XGBoost	0.88	0.89	0.89	0.9321

Table 2. Model metrics after hyper parameter tuning.

All four models demonstrated strong sensitivity toward CAD detection, with Random Forest and Gradient Boosting outputting the fewest false negatives, which is critical in clinical screening to avoid missed outcomes.



Feature importance analysis derived from Random Forest revealed that ST segment slope, chest pain type, and ST depression (oldpeak) were the most influential predictors of CAD. These findings align with established clinical criteria, as ST slope and oldpeak reflect ischemic changes during exercise testing, while chest pain type differentiates typical anginal symptoms from non-specific complaints. Exercise-induced angina, maximum heart rate, and cholesterol levels also contributed meaningfully to prediction performance. Lower-ranked features included resting blood pressure, sex, and fasting blood sugar, consistent with their more modest roles in CAD risk when considered independently. These importance scores highlight the predictive value of stress test-related parameters and symptom classification, validating the models' clinical interpretability.

Discussion

This study highlights the predictive value of both clinical and stress-test-derived features in identifying coronary artery disease (CAD). Variables such as ST segment slope, chest pain type, and ST depression (oldpeak) emerged as the most influential predictors, aligning with established diagnostic markers of myocardial ischemia. Among the evaluated algorithms, ensemble methods, particularly Random Forest and Gradient Boosting, achieved superior performance, with ROC–AUC values exceeding 0.93, reflecting strong discriminative capacity. While Logistic Regression demonstrated slightly lower overall accuracy, its consistently high recall supports its utility in high-sensitivity screening contexts.

However, these performance figures should be interpreted with caution: no formal assessment of multicollinearity was performed, and several predictors (e.g., exercise-related ECG variables) are highly correlated with each other and with the target variable. Consequently, some feature importance estimates and performance metrics may be inflated due to redundant information, a point explored further below.

While this study specifically targeted coronary artery disease (CAD), many of the predictive variables used such as age, blood pressure, cholesterol, fasting blood sugar, and smoking status represent common risk factors across a wide range of cardiovascular conditions, including stroke, peripheral artery disease, and heart failure. This overlap supports the potential for extending the model framework to broader CVD prediction. However, certain variables like ST slope, exercise-induced angina, and ECG abnormalities are more specific to ischemic heart disease and may not generalize well to non-coronary CVD phenotypes. Future work should take these condition-specific differences into account when applying the model to other CVD subtypes.

One of the techniques that was used to reach a high level of accuracy and recall in this experiment was the implementation of SMOTE (Synthetic Minority Oversampling Technique). SMOTE works by generating synthetic samples along the line segments connecting a minority class sample and its nearest neighbors.

$$x(new) \ = xi + \delta\cdot(xneighbor - xi)$$

Above is the formula for SMOTE, and δ is a random number between 0 and 1. This method improves the model's ability to learn from underrepresented CAD cases and helps mitigate bias in classification results. Additionally, all models were tuned via RandomizedSearchCV using 5-fold cross-validation, allowing fair and efficient comparison across classifiers.

Limitations

Despite the strong predictive performance, this study has a few important limitations. First, the dataset focuses exclusively on CAD outcomes and doesn't include other major cardiovascular conditions like stroke or heart failure, which limits how easily the results can be generalized. Second, because the data are cross-sectional, the models can't support time-to-event prediction or survival analysis. Finally, since the data come from patients already referred for cardiac evaluation, there's a risk of selection bias that could reduce how well the findings apply to the general population. One important limitation is that multicollinearity wasn't formally assessed. Several features, especially those related to exercise ECG, like oldpeak, ST slope, and exercise-induced angina, are likely to carry overlapping information. This kind of redundancy can lead to inflated importance scores and potentially boost model performance in ways that don't reflect true predictive value, especially in tree-based models. Future versions of the pipeline should incorporate collinearity checks, like Variance Inflation Factor (VIF) scores or correlation matrices, to get a clearer picture of each feature's individual contribution and to improve the overall reliability of the model.

Even with recent advancements, machine learning approaches for cardiovascular disease prediction encounter several significant limitations affecting clinical implementation and real-world performance. Data quality and representation present fundamental challenges, as many models incorporate subjective or self-reported variables susceptible to recall bias and misreporting. Even objective electronic health record data frequently contains inconsistencies that potentially distort model predictions (Cai et al. 2024). Additionally, the class imbalance prevalent in cardiovascular datasets, where disease-positive cases represent a minority, leads to biased models with poor performance on positive cases unless specialized correction techniques are applied (El-Sofany et al. 2024).

External validity and generalizability are other critical concerns. Many ML models are developed using relatively small or homogeneous datasets, such as the UCI Cleveland Heart Disease dataset, which may not reflect real-world clinical populations. This restricts the model's ability to generalize across diverse demographic or geographic settings (Krittanawong 2020). Additionally, the lack of interpretability remains a concern, particularly for complex ensemble or deep learning models, which are making it difficult for clinicians to trust or understand how predictions are made. While explainable AI tools like SHAP help address this, their use is not yet standardized. Finally, ethical and clinical integration challenges, such as unclear thresholds for clinical decision-making, lack of prospective validation, and difficulty integrating ML outputs into existing workflows, all slow down practical deployment (Cai et al. 2024).

Future Work

Future research should focus on external validation using large, diverse datasets like All of Us, UK Biobank, and MIMIC-IV. These resources offer rich, longitudinal data across different populations, making them ideal for testing how well models generalize to real-world clinical settings. For example, the UK Biobank has already played a key role in validating cardiovascular risk models and identifying new biomarkers across large cohorts (Li et al. 202). Another important direction is shifting from static classification to longitudinal prediction. By incorporating time-based features, models can better track how disease develops over time and potentially flag issues earlier. Studies have shown that machine learning models using longitudinal data often outperform traditional ones in predicting cardiovascular events (Li et al. 2022).

It's also important to handle multicollinearity more effectively. Using techniques like Bayesian feature selection in joint models can help manage correlated variables and improve both model performance and interpretability. Exploring how well models work across different groups, by sex or age, would also help ensure fairness and make the models more broadly applicable. Taken together, these steps are key to moving from proof-of-concept to practical clinical tools for cardiovascular risk prediction.

Conclusion

This study showed that machine learning can be an effective tool for predicting coronary artery disease (CAD), using a widely studied tabular dataset and several classification models. Ensemble approaches like Random Forest and Gradient Boosting performed best, with ROC–AUC scores above 0.93 after tuning, while Logistic Regression stood out for its strong recall and interpretability, making it a solid option for clinical screening. Feature importance analysis pointed to key predictors like ST segment slope, chest pain type, and ST depression, which are consistent with established markers of ischemia. While the focus was on CAD, many of the input variables are also relevant to other cardiovascular conditions, making CAD a reasonable starting point for broader CVD risk modeling. Still, the study has some important limitations, including dataset bias, potential multicollinearity, and the absence of longitudinal data. Future research should aim to validate these models on larger and more diverse datasets, integrate time-based predictions, and handle redundant features more effectively to create models that are both accurate and practical for real-world clinical use.

References

2025 heart disease and stroke statistics update fact sheet. (n.d.). https://www.heart.org/en/-/media/PHD-Files-2/Science-News/2/2025-Heart-and-Stroke-Stat-Update/2025-Statistics-At-A-Glance.pdf?sc_lang=en

Cai, Y.-Q., Gong, D.-X., Tang, L.-Y., Cai, Y., Li, H.-J., Jing, T.-C., Gong, M., Hu, W., Zhang, Z.-W., Zhang, X., & Zhang, G.-W. (2024, July 26). Pitfalls in developing machine learning models for predicting cardiovascular diseases: Challenge and solutions. Journal of medical Internet research. https://pmc.ncbi.nlm.nih.gov/articles/PMC11316160/

Centers for Disease Control and Prevention. (n.d.). Heart disease facts. Centers for Disease Control and Prevention. https://www.cdc.gov/heart-disease/data-research/facts-stats/index.html?utm_source=chatgpt.com

Detrano R;Janosi A;Steinbrunn W;Pfisterer M;Schmid JJ;Sandhu S;Guppy KH;Lee S;Froelicher V; (1989, August 1). International application of a new probability algorithm for the diagnosis of coronary artery disease. The American journal of cardiology. https://pubmed.ncbi.nlm.nih.gov/2756873/

El-Sofany, H., Bouallegue, B., & Abd El-Latif , Y. (2024). A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method. Scientific reports. https://pubmed.ncbi.nlm.nih.gov/39375427/

Feigin, V. L., Brainin, M., Norrving, B., Martins, S. O., Pandian, J., Lindsay, P., F Grupper, M., & Rautalin, I. (2025, February). World Stroke Organization: Global Stroke Fact Sheet 2025. International journal of stroke : official journal of the International Stroke Society. https://pmc.ncbi.nlm.nih.gov/articles/PMC11786524/?utm_source=chatgpt.com

Forrest, I. S., Petrazzini, B. O., Duffy, Á., Park, J. K., Marquez-Luna, C., Jordan, D. M., Rocheleau, G., Cho, J. H., Rosenson, R. S., Narula, J., Nadkarni, G. N., & Do, R. (2023, January 21). Machine learning-based marker for coronary artery disease: Derivation and validation in two longitudinal cohorts. Lancet (London, England). https://pmc.ncbi.nlm.nih.gov/articles/PMC10069625/?utm_source=chatgpt.com

Gallucci, G., Tartarone, A., Lerose, R., Lalinga, A. V., & Capobianco, A. M. (2020, July). Cardiovascular risk of smoking and benefits of smoking cessation. Journal of thoracic disease. https://pmc.ncbi.nlm.nih.gov/articles/PMC7399440/

Krittanawong, C., Virk, H. U. H., Bangalore, S., Wang, Z., Johnson, K. W., Pinotti, R., Zhang, H., Kaplin, S., Narasimhan, B., Kitai, T., Baber, U., Halperin, J. L., & Tang, W. H. W. (2020, September 29). Machine learning prediction in cardiovascular diseases: A meta-analysis. Nature News. https://www.nature.com/articles/s41598-020-72685-1?utm_source=chatgpt.com

Li, Y., Salimi-Khorshidi, G., Rao, S., Canoy, D., Hassaine, A., Lukasiewicz, T., Rahimi, K., & Mamouei, M. (2022, October 21). Validation of risk prediction models applied to longitudinal electronic health record data for the prediction of major cardiovascular events in the presence of data shifts. European heart journal. Digital health. https://pmc.ncbi.nlm.nih.gov/articles/PMC9779795/?utm_source=chatgpt.com

Lin, A., Manral, N., McElhinney, P., Killekar, A., Matsumoto, H., Kwiecinski, J., Pieszko, K., Razipour, A., Grodecki, K., Park, C., Otaki, Y., Doris, M., Kwan, A. C., Han, D., Kuronuma, K., Flores Tomasino, G., Tzolos, E., Shanbhag, A., Goeller, M., … Dey, D. (2022, April). Deep learning-enabled coronary CT angiography for plaque and stenosis quantification and cardiac risk prediction: An International Multicentre Study. The Lancet. Digital health. https://pmc.ncbi.nlm.nih.gov/articles/PMC9047317/

Mansouri, A., Khosravi, A., Mehrabani-Zeinabad, K., Kopec, J., Adawi, K., & Lui, M. (2023, June 23). Trends in the burden and determinants of hypertensive heart disease in the Eastern Mediterranean region, 1990–2019: an analysis of the Global Burden of Disease Study 2019. eClinicalMedicine. https://www.thelancet.com/journals/eclinm/article/PIIS2589-5370%2823%2900211-0/fulltext?utm_source=chatgpt.com

Mathur, R. K. (2010, April). Role of diabetes, hypertension, and cigarette smoking on Atherosclerosis. Journal of cardiovascular disease research. https://pmc.ncbi.nlm.nih.gov/articles/PMC2945206/?utm_source=chatgpt.com

Narain, R. (2016). Cardiovascular risk prediction: A comparative study of framingham and Quantum Neural Network based approach. Patient preference and adherence. https://pubmed.ncbi.nlm.nih.gov/27486312/

Wang, J., Xue, Q., Zhang, C. W. J., Wong, K. K. L., & Liu, Z. (2024, July 1). Explainable coronary artery disease prediction model based on AutoGluon from AutoML framework. Frontiers in cardiovascular medicine. https://pmc.ncbi.nlm.nih.gov/articles/PMC11246996/

World Health Organization. (n.d.). Cardiovascular diseases. World Health Organization. https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

Yang, L., Wu, H., Jin, X., Zheng, P., Hu, S., Xu, X., Yu, W., & Yan, J. (2020, March 23). Study of cardiovascular disease prediction model based on Random Forest in eastern China. Nature News. https://www.nature.com/articles/s41598-020-62133-5