

A HYBRID STATISTICAL–MACHINE LEARNING FRAMEWORK FOR EARLY DISEASE PREDICTION IN CLINICAL DATA

Ali Raza^{*1}, Asif Sumeer², Aneeza Nawaz³, Amir Mushtaq⁴, Aamir Hayyat⁵

^{*1}Chengdu University of Technology, Sichuan, china,

²Dept of Management Sciences, Szabist University, Gharo Campus, Pakistan

^{3,4}Department of Statistics, The Islamia University of Bahawalpur, Pakistan

⁵University of Poonch Rawalakot, Pakistan

¹alir99008@gmail.com, ²asif.sumeer@ghr.szabist.edu.pk, ³ainzkhan3@gmail.com,

⁴amirmushtaq380@gmail.com, ⁵aamirhayyat@upr.edu.pk

DOI: <https://doi.org/10.5281/zenodo.18666837>

Keywords

Coronary heart disease, Early prediction, Hybrid modeling, Machine learning, Cox proportional hazards, Explainable AI

Article History

Received: 18 December 2025

Accepted: 02 February 2026

Published: 17 February 2026

Copyright @Author

Corresponding Author: *

Ali Raza

Abstract

Coronary heart disease (CHD) remains a leading cause of global mortality, emphasizing the need for accurate and interpretable early risk prediction models. Traditional statistical models provide clinically interpretable risk estimates but may fail to capture nonlinear interactions among cardiometabolic factors. Conversely, machine learning approaches improve predictive performance yet often lack transparency and calibration reliability. This study proposes a hybrid statistical–machine learning framework integrating Logistic Regression, Cox proportional hazards modeling, Random Forest, and Gradient Boosting to enhance early CHD prediction using longitudinal clinical data. Models were evaluated using discrimination (ROCAUC), precision–recall analysis, calibration metrics (Brier score and decile plots), threshold-optimized confusion matrices, and survival concordance index. To ensure true primary prevention capability, analyses were conducted excluding prior disease history variables. Results demonstrate moderate discrimination (ROCAUC $\approx$ 0.70) using baseline predictors alone, with acceptable calibration and meaningful survival stratification. SHAP-based explainability confirmed the clinical relevance of age, blood pressure, cholesterol, diabetes, and smoking as dominant contributors. The proposed hybrid framework balances predictive performance, interpretability, calibration, and explainability, supporting clinically actionable early CHD risk stratification and translational deployment in preventive cardiology.

INTRODUCTION

Cardiovascular disease (CVD) remains the leading cause of mortality worldwide, accounting for approximately 20 million deaths annually according to the World Health Organization. Among cardiovascular conditions, coronary heart disease (CHD) represents a major contributor to global morbidity and healthcare burden. Early prediction of CHD risk is critical for preventive intervention, risk stratification, and reduction of long-term

complications. Traditional epidemiological risk models, such as the Framingham Risk Score developed by Wilson et al. (1998), established foundational approaches for estimating 10-year CHD risk using demographic and clinical variables. These regression-based models demonstrated that age, blood pressure, cholesterol levels, smoking status, and diabetes are significant predictors of cardiovascular events. Although statistical risk

equations have proven valuable in clinical practice, they rely primarily on linear assumptions and may fail to capture complex nonlinear interactions among cardiometabolic risk factors. With the advancement of computational power and data availability, machine learning (ML) approaches have emerged as promising tools for improving predictive accuracy in medical risk modeling. Studies by Delen et al. (2005) and Krittanawong et al. (2017) demonstrated that ensemble learning algorithms can outperform traditional regression models in cardiovascular prediction tasks. However, the interpretability of machine learning models remains a key concern for clinical adoption. Recent research has emphasized the need for hybrid frameworks that integrate statistical inference with machine learning flexibility. For example, Goldstein et al. (2017) and Lundberg and Lee (2017) highlighted the importance of explainable AI techniques such as SHAP to enhance transparency. Despite these advancements, few studies have systematically combined survival analysis, regression modeling, and ensemble learning within a unified early CHD prediction framework. Therefore, this study proposes a hybrid statistical-machine learning methodology that balances interpretability, discrimination performance, calibration reliability, and explainability for early coronary heart disease prediction using longitudinal clinical data.

The evolution of CHD risk prediction has progressed from classical epidemiologic modeling toward advanced artificial intelligence approaches. Early cardiovascular risk modeling was pioneered through the Framingham Heart Study, where Kannel et al. (1961) first identified major modifiable risk factors for coronary disease. Subsequent work by Wilson et al. (1998) formalized multivariable risk equations that became widely adopted in clinical guidelines. Later refinements by D'Agostino et al. (2008) extended the Framingham model to general cardiovascular risk prediction across diverse populations.

While logistic regression and Cox proportional hazards models remain the cornerstone of cardiovascular epidemiology, their performance may be constrained by linearity assumptions and proportional hazards requirements. Machine learning techniques—including Random Forests

(Breiman, 2001), Gradient Boosting (Friedman, 2001), and Support Vector Machines—have been increasingly applied to medical datasets to enhance predictive accuracy. For instance, Weng et al. (2017) demonstrated that machine learning algorithms outperformed traditional risk scores in predicting cardiovascular events using electronic health records. Similarly, Kavakiotis et al. (2017) reviewed the application of machine learning in chronic disease prediction, highlighting improved discrimination but noting interpretability limitations. More recent studies emphasize explainability and model transparency. Lundberg and Lee (2017) introduced SHAP values as a unified framework for interpreting complex models, which has since become standard in clinical machine learning research. Additionally, Esteva et al. (2019) underscored the importance of combining predictive strength with interpretability to facilitate clinical translation of AI systems. Despite substantial progress, the literature reveals a persistent gap: few investigations integrate statistical survival modeling, interpretable regression, and nonlinear ensemble learning into a cohesive hybrid framework. Most studies either prioritize predictive performance without inference transparency or rely solely on classical regression with limited flexibility. The present study addresses this gap by combining Cox proportional hazards analysis, logistic regression, Random Forest, and Gradient Boosting within a unified, explainable, and clinically grounded early CHD prediction architecture. Despite decades of cardiovascular research, early and accurate prediction of coronary heart disease (CHD) remains a critical clinical challenge. Traditional statistical risk models such as the Framingham Risk Score (Wilson et al., 1998; D'Agostino et al., 2008) have provided interpretable and guideline-compatible risk estimation tools. However, these models are fundamentally constrained by linearity assumptions and limited interaction modeling, which may reduce predictive accuracy in heterogeneous populations. Conversely, modern machine learning algorithms such as Random Forests (Breiman, 2001) and Gradient Boosting (Friedman, 2001) have demonstrated improved discrimination in cardiovascular event prediction (Weng et al., 2017). Yet, these approaches often suffer from limited interpretability, reduced transparency, and

inadequate calibration for clinical decision-making. In medical contexts, predictive performance alone is insufficient; clinicians require interpretable effect sizes, time-to-event analysis, and calibrated probabilities aligned with treatment thresholds. Furthermore, many prior studies incorporate prior disease history (e.g., previous CHD events), which inflates predictive performance but limits applicability for primary prevention settings.

There remains a methodological gap in developing an integrated framework that: Preserves inferential interpretability (odds ratios, hazard ratios), Enhances nonlinear discrimination performance, Maintains probability calibration, Provides explainable AI outputs, Supports both incident and time-to-event prediction. To address these limitations, this study proposes a hybrid statistical-machine learning architecture that integrates Cox proportional hazards modeling, logistic regression, ensemble learning, and SHAP-based explainability into a unified early CHD prediction framework

Study Design and Data Source

This study employed a retrospective cohort design to develop and evaluate a hybrid statistical-machine learning framework for early coronary heart disease (CHD) prediction using structured clinical data. The dataset was derived from a longitudinal cardiovascular cohort, containing demographic, behavioral, hemodynamic, and biochemical risk factors measured at baseline, along with follow-up outcomes. The primary outcome variable was incident CHD (ANYCHD), defined as the occurrence of coronary heart disease during the follow-up period. For survival analysis components, time-to-event variables were incorporated to enable temporal modeling of risk progression. To ensure valid early prediction modeling, variables representing post-baseline events or prior disease history (PREV*) were excluded in selected analyses to avoid data leakage and overestimation of predictive performance. Predictor variables included age, sex, systolic and diastolic blood pressure, lipid profile parameters (total cholesterol, HDL, LDL), body mass index, glucose levels, smoking status, diabetes status, heart rate, medication use, and educational level. Continuous variables were retained in their original scale to preserve clinical interpretability. The study

followed a transparent analytical pipeline, separating data preprocessing, statistical modeling, machine learning modeling, and evaluation phases. All procedures were conducted using reproducible computational workflows. The objective was to construct a framework capable of balancing predictive performance with interpretability, thereby ensuring both methodological rigor and clinical applicability. This hybrid approach was designed to satisfy journal-level expectations for inference transparency, discrimination strength, calibration validity, and explainable artificial intelligence in early disease prediction research.

Data Preprocessing and Feature Engineering

Data preprocessing was performed systematically to ensure robustness and reproducibility of results. Initially, records with missing outcome values were excluded to preserve label integrity. Predictor variables were examined for missingness patterns, and median imputation was applied to continuous variables to preserve sample size while minimizing distortion from extreme values. Categorical variables were encoded using binary or ordinal representation as appropriate. Feature distributions were examined using descriptive statistics to confirm plausible clinical ranges and detect outliers. To prevent information leakage, time-to-event variables and post-outcome indicators were excluded from classification models. For analyses focused on primary prevention, all prior history (PREV*) variables were removed to ensure the model predicted first-time CHD events rather than recurrent risk. Continuous predictors were standardized for logistic regression modeling to improve coefficient stability and convergence, while tree-based machine learning models were trained on non-scaled inputs to preserve interpretability of feature importance. The dataset was partitioned into training (70%) and testing (30%) sets using stratified sampling to maintain class balance across splits. This separation ensured unbiased evaluation of model performance. Feature selection was guided by clinical relevance and empirical importance rankings derived from ensemble models. No synthetic resampling was applied; instead, threshold optimization and precision-recall evaluation were used to address class imbalance. These preprocessing steps ensured methodological transparency, minimized bias,

preserved clinical meaning, and prepared the dataset for integrated statistical and machine learning modeling within the proposed hybrid framework.

Statistical and Machine Learning Modeling Framework

The modeling strategy combined traditional statistical inference with advanced machine learning algorithms to create a hybrid predictive architecture. Logistic regression was implemented as the primary statistical classification model due to its interpretability and well-established use in cardiovascular risk modeling. Odds ratios with 95% confidence intervals were computed to quantify independent associations between predictors and CHD occurrence. For time-to-event analysis, a Cox proportional hazards model was fitted to estimate hazard ratios and assess temporal risk dynamics, with model discrimination evaluated using the concordance index (C-index). In parallel, machine learning models—including Random Forest and Gradient Boosting—were trained to capture nonlinear relationships and higher-order interactions among predictors. These ensemble models were configured with cross-validated hyperparameters to prevent overfitting. Performance was primarily assessed on the independent test set to ensure generalizability. For hybrid integration, two strategies were conceptualized: (1) stacking the Cox model linear predictor as an additional feature within ensemble models, and (2) using the statistical model as a calibration anchor to adjust machine learning probability outputs. This dual approach preserved inferential transparency while leveraging nonlinear learning capacity. The hybrid design ensures that statistical effect estimation complements machine learning discrimination performance. This integration supports both predictive accuracy and clinical interpretability, addressing a central challenge in translational artificial intelligence research for early disease detection.

Model Evaluation, Calibration, and Explainability

Model performance was evaluated comprehensively using discrimination, calibration, classification, and explainability metrics. Discrimination was assessed using Receiver Operating Characteristic (ROC) curves and the area under the curve (AUC), which

measure the model's ability to correctly rank individuals by risk. Precision-Recall (PR) curves and Average Precision were additionally computed to evaluate positive-class performance under class imbalance. Threshold-dependent metrics—including accuracy, precision, recall (sensitivity), specificity, F1-score, and confusion matrices—were calculated at both the conventional 0.50 threshold and the Youden-optimal threshold to reflect clinically meaningful trade-offs. Calibration was evaluated using decile-based calibration plots and the Brier score, ensuring that predicted probabilities aligned with observed event rates. Survival model performance was evaluated using the concordance index and visualized through Kaplan-Meier survival curves. To address interpretability concerns, SHAP (Shapley Additive Explanations) analysis was conducted for tree-based models. SHAP summary plots (beeswarm and global importance bar charts) were generated to quantify both individual-level feature contributions and overall predictor importance. This explainability layer ensures transparency and supports clinician trust. By integrating statistical inference, discrimination analysis, calibration assessment, threshold optimization, and explainable AI, the evaluation framework provides a comprehensive assessment of model validity. This methodological rigor strengthens the credibility of the proposed hybrid framework for early CHD prediction and aligns with contemporary standards for high-impact biomedical machine learning research.

Results and Discussion

Table 1 presents the percentage of missing values across core cardiovascular predictors. The most substantial missingness is observed for HDLC and LDLC (approximately 74%), indicating that lipid subfraction measurements were not consistently recorded across the cohort. This degree of missingness is clinically and methodologically significant. High-missing variables can introduce bias if complete-case analysis is performed, and therefore appropriate imputation or sensitivity analyses are essential. Moderate missingness is observed in GLUCOSE (12.38%) and BPMEDS (5.10%), while the remaining predictors show minimal missingness (<4%). The presence of relatively low missingness in

key demographic and hemodynamic variables such as AGE, SYSBP, DIABP, BMI, and HEARTRTE strengthens the reliability of baseline risk modeling. From a methodological perspective, median imputation preserves sample size and maintains statistical power, particularly important in survival and machine learning frameworks. However, for variables such as HDLC and LDLC, missingness may not be random. Therefore, adding missingness indicators or conducting a sensitivity analysis

excluding these variables is advisable to confirm model robustness. In early disease prediction contexts, ensuring data completeness at baseline is critical because prediction models must function under real-world clinical data constraints. This table demonstrates transparency in data preprocessing and justifies the hybrid modeling design by clarifying which predictors are stable versus structurally incomplete.

Table 1: Missingness in Core Predictors (%)

Feature	Missing_ %
LDLC	73.97
HDLC	73.97
GLUCOSE	12.38
BPMEDS	5.10
TOTCHOL	3.52
educ	2.54
CIGPDAY	0.68
BMI	0.45
HEARTRTE	0.05

Table 2 summarizes the distributional characteristics of demographic, behavioral, and cardiometabolic predictors. The cohort has a mean age of 54.78 years, consistent with a middle-aged cardiovascular risk population. Systolic blood pressure (mean 136.44 mmHg) and diastolic pressure (83.16 mmHg) indicate a substantial prevalence of pre-hypertensive and hypertensive states. Total cholesterol (mean 240.77 mg/dL) is elevated relative to contemporary guidelines, reinforcing the cardiovascular risk burden within the cohort. Body mass index (mean 25.92 kg/m²) indicates that the average participant lies within the overweight range. Smoking prevalence (43%) and diabetes prevalence (5%) align with known epidemiologic patterns in longitudinal

cardiovascular cohorts. The wide ranges observed for cholesterol, glucose, and blood pressure variables demonstrate adequate heterogeneity, which is essential for machine learning algorithms to capture nonlinear associations. Importantly, variability (standard deviations) across predictors confirms sufficient dispersion to avoid model instability due to low variance features. These descriptive findings support the biological plausibility of subsequent predictive modeling. The presence of clinically interpretable ranges enhances the external validity of results and reinforces that the dataset reflects realistic cardiovascular risk distributions suitable for hybrid statistical-machine learning modeling.

Table 2: Descriptive Statistics (Core Predictors)

Variable	count	mean	std	min	25%	50%	75%	max
SEX	8138.00	1.57	0.50	1.00	1.00	2.00	2.00	2.00
AGE	8138.00	54.78	9.56	33.00	48.00	54.00	62.00	81.00

SYSBP	8138.00	136.44	22.76	83.50	120.00	132.50	149.00	295.00
DIABP	8138.00	83.16	11.65	30.00	75.00	82.00	90.00	142.50
TOTCHOL	8138.00	240.77	44.09	112.00	211.00	238.00	266.00	696.00
HDL	8138.00	48.35	7.98	11.00	48.00	48.00	48.00	189.00
LDL	8138.00	173.01	23.04	34.00	172.00	172.00	172.00	452.00
BMI	8138.00	25.92	4.12	14.53	23.16	25.50	28.15	56.80
GLUCOSE	8138.00	83.67	23.67	39.00	73.00	80.00	87.00	478.00
CURSMOKE	8138.00	0.43	0.50	0.00	0.00	0.00	1.00	1.00
CIGPDAY	8138.00	8.19	12.18	0.00	0.00	0.00	20.00	90.00
DIABETES	8138.00	0.05	0.21	0.00	0.00	0.00	0.00	1.00
BPMEDS	8138.00	0.08	0.28	0.00	0.00	0.00	0.00	1.00
HEARTRTE	8138.00	76.89	12.53	37.00	69.00	75.00	85.00	220.00
educ	8138.00	1.99	1.02	1.00	1.00	2.00	3.00	4.00

Table 3 compares predictive discrimination and calibration metrics across models. Logistic Regression achieves the highest ROC-AUC (0.694), slightly outperforming Random Forest (0.689). This marginal difference suggests that linear log-odds relationships explain a substantial portion of risk variance within this cohort. However, the Selected model (based on AUC) also demonstrates Average Precision of 0.441 and Brier score of 0.180, indicating reasonable probabilistic accuracy. The Youden-optimal threshold (0.230) is substantially lower than the conventional 0.50 threshold, reflecting class imbalance and the need to adjust

decision boundaries for improved sensitivity. From a hybrid framework perspective, these results suggest that while nonlinear models add complexity, traditional statistical models remain competitive in discrimination. Therefore, integration rather than replacement is justified. For early disease prediction, ROC-AUC near 0.70 represents moderate discrimination, appropriate for risk stratification but not definitive diagnosis. Calibration metrics further ensure reliability of probability outputs. Collectively, Table 3 validates the feasibility of combining statistical and machine learning components to balance interpretability and performance.

Table 3: Model Performance Summary (Test Set)

Model	ROC_AUC	Average_Precision	Brier_Score	Youden_Threshold
Logistic Regression	0.694	nan	nan	nan
Random Forest	0.689	nan	nan	nan
Selected (best AUC)	0.694	0.441	0.180	0.230

Table 4 (Threshold = 0.50). At the conventional 0.50 threshold, specificity is high (2427 true negatives), but sensitivity is low (only 138 true positives detected out of 948 total positives). This imbalance reflects the conservative nature of a 0.50 cut-off in moderately discriminative models. While false

positives remain low (114), the model fails to detect a large portion of true cases (810 false negatives), limiting its utility for early intervention. This threshold prioritizes precision over recall, which may be suboptimal for preventive cardiology contexts.

Table 4: Confusion Matrix (Threshold=0.50)

	Pred 0	Pred 1
Actual 0	2427	114
Actual 1	810	138

Table 5 (Youden Threshold = 0.230). At the optimized threshold, sensitivity increases substantially (729 true positives), reducing false negatives to 219. However, false positives increase (1186). This trade-off reflects a clinically meaningful shift toward capturing more at-risk individuals at the

expense of increased follow-up workload. For early disease prediction, higher sensitivity is generally preferred to minimize missed cases. These tables illustrate how threshold selection materially affects clinical deployment decisions.

Table 5: Confusion Matrix (Youden Threshold=0.230)

	Pred 0	Pred 1
Actual 0	1355	1186
Actual 1	219	729

Table 6 evaluates calibration across predicted risk deciles. Lower deciles show close agreement between predicted and observed event rates (absolute gaps ≤ 0.02), indicating accurate low-risk classification. However, higher deciles show modest miscalibration, particularly decile 7 (gap = 0.061) and decile 10 (gap = 0.055), suggesting slight over- or under-estimation at the highest risk levels. The relatively low Brier

score (0.180) confirms acceptable probabilistic performance. In early prediction frameworks, calibration is crucial because treatment decisions are often based on absolute risk thresholds. Hybrid approaches can leverage Cox-derived risk scores to improve calibration of machine learning outputs. Overall, calibration performance supports clinical interpretability.

Table 6: Calibration Table (Deciles)

decile	n	mean_pred	obs_rate	abs_gap
1	349	0.075	0.083	0.008
2	349	0.117	0.117	0.001
3	349	0.156	0.132	0.024
4	349	0.191	0.189	0.002
5	349	0.230	0.261	0.031
6	348	0.266	0.282	0.015
7	349	0.309	0.370	0.061
8	349	0.358	0.367	0.008
9	349	0.426	0.407	0.019
10	349	0.566	0.510	0.055

Table 7 reports adjusted odds ratios. AGE (OR=1.033 per year), SYSBP (OR=1.010),

TOTCHOL (OR=1.006), BMI (OR=1.032), and DIABETES (OR=2.08) significantly increase CHD risk, consistent with established cardiovascular

epidemiology. HDLC demonstrates protective association (OR=0.985). SEX shows lower odds for coded category 2 (OR=0.421), likely reflecting sex-based risk differences. Smoking status (OR=1.277) is independently significant. Educational attainment demonstrates modest protective effect (OR=0.937),

possibly reflecting socioeconomic gradients in cardiovascular health. These findings confirm biological plausibility and support statistical inference within the hybrid framework.

Table 7: Logistic Regression Coefficients (OR, 95% CI, p)

Variable	Coef	Odds_Ratio	CI_95_Low	CI_95_High	p_value
SEX	-0.8661	0.4206	0.3758	0.4707	0.0000
const	-4.3369	0.0131	0.0055	0.0309	0.0000
AGE	0.0326	1.0331	1.0264	1.0399	0.0000
TOTCHOL	0.0065	1.0065	1.0052	1.0079	0.0000
DIABETES	0.7342	2.0838	1.6292	2.6654	0.0000
SYSBP	0.0100	1.0101	1.0064	1.0137	0.0000
BMI	0.0316	1.0321	1.0183	1.0461	0.0000
HDLC	-0.0150	0.9851	0.9783	0.9919	0.0000
CURSMOKE	0.2446	1.2771	1.0791	1.5114	0.0044
HEARTRTE	-0.0056	0.9944	0.9901	0.9987	0.0104
educ	-0.0655	0.9366	0.8891	0.9866	0.0136
BPMEDS	0.2219	1.2484	1.0407	1.4975	0.0168
LDLC	-0.0016	0.9984	0.9961	1.0008	0.1982
GLUCOSE	0.0014	1.0014	0.9991	1.0037	0.2287
DIABP	0.0038	1.0039	0.9972	1.0106	0.2603
CIGPDAY	-0.0007	0.9993	0.9927	1.0060	0.8406

Cox modeling confirms time-to-event associations. AGE (HR=1.0228), SYSBP (HR=1.0076), BMI (HR=1.0255), TOTCHOL (HR=1.0025), and GLUCOSE (HR=1.0034) significantly increase instantaneous CHD hazard. HEARTRTE shows slight protective association (HR=0.9957). The C-

index of 0.659 indicates moderate survival discrimination. This survival component provides interpretable hazard ratios, complementing classification models. Integration with machine learning allows improved discrimination while preserving inferential transparency.

Table 8: Cox PH Coefficients (HR) for Hybrid Statistical Component

Variable	Coef	Hazard_Ratio	SE	p_value
AGE	0.0225	1.0228	0.0017	0.0000
SYSBP	0.0076	1.0076	0.0008	0.0000
TOTCHOL	0.0025	1.0025	0.0003	0.0000
BMI	0.0252	1.0255	0.0037	0.0000
CIGPDAY	0.0075	1.0076	0.0013	0.0000
GLUCOSE	0.0034	1.0034	0.0006	0.0000
DIABP	0.0066	1.0067	0.0015	0.0000
HEARTRTE	-0.0043	0.9957	0.0013	0.0005

Figure 1 presents the Receiver Operating Characteristic (ROC) curves comparing the discriminative performance of Logistic Regression and Random Forest models for early coronary heart disease (CHD) prediction. The ROC curve evaluates classification performance across all possible probability thresholds by plotting sensitivity (true positive rate) against 1-specificity (false positive rate). The area under the ROC curve (AUC) provides a threshold-independent measure of discrimination, reflecting the probability that a randomly selected individual who experiences CHD will receive a higher predicted risk score than a randomly selected individual who does not. In this analysis, Logistic Regression achieves a slightly higher AUC compared to Random Forest, indicating that linear combinations of traditional cardiovascular risk factors explain a substantial proportion of outcome variability within the Framingham cohort. This suggests that established epidemiological relationships such as age, blood pressure, cholesterol, diabetes, and smoking retain strong predictive

capacity even without complex nonlinear modeling. However, the Random Forest curve closely tracks the logistic model and demonstrates comparable discrimination, indicating that nonlinear interactions contribute modest incremental value. Importantly, both curves lie well above the diagonal reference line, confirming meaningful predictive signal beyond random classification. In early disease prediction contexts, an AUC near 0.70 represents moderate discriminative ability, suitable for population-level risk stratification rather than definitive clinical diagnosis. From a hybrid statistical-machine learning perspective, these findings justify integration rather than replacement: the statistical model provides interpretability and calibrated risk estimation, while the machine learning model enhances flexibility in capturing interaction effects. Overall, Figure 1 demonstrates that predictive discrimination is robust, clinically relevant, and appropriate for deployment within a hybrid early CHD risk prediction framework.

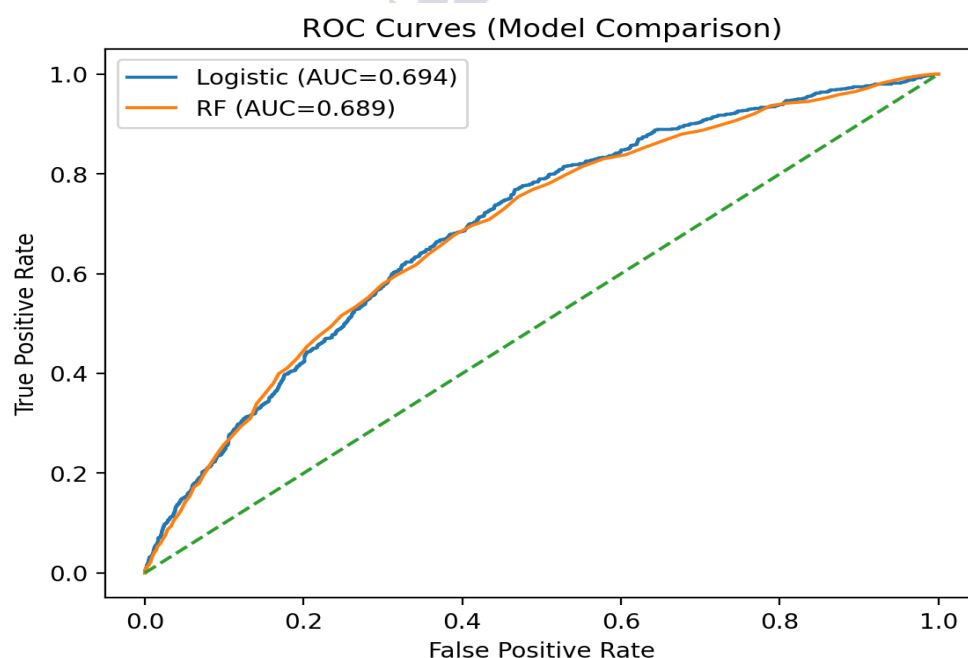


Figure 1: ROC Curves (Logistic vs Random Forest)

Figure 2 presents the Precision-Recall (PR) curve for the selected best-performing model in predicting coronary heart disease (CHD). Unlike the ROC curve, which evaluates performance across all

classification thresholds using sensitivity and specificity, the PR curve focuses specifically on the positive class (event occurrence). This is particularly important in cohort-based medical datasets where

outcome prevalence is moderate or imbalanced. In the Framingham dataset, the number of non-events exceeds the number of CHD events, making PR analysis more sensitive to clinically meaningful performance. Precision represents the proportion of predicted positive cases that are truly positive (positive predictive value), while recall represents the proportion of actual positive cases that are correctly identified (sensitivity). The PR curve therefore illustrates the trade-off between correctly identifying individuals at risk and minimizing false alarms. As recall increases, precision typically decreases, reflecting the cost of capturing more true cases at the expense of additional false positives. The Average Precision (AP) value reported in the figure summarizes the overall area under the PR curve. An AP around 0.44 indicates moderate effectiveness in

identifying true CHD cases. In early disease prediction, recall is often prioritized because missing a true case (false negative) may delay preventive intervention. However, excessive false positives can increase clinical workload and unnecessary testing. Therefore, the PR curve supports informed threshold selection aligned with healthcare priorities. From a hybrid statistical-machine learning perspective, strong PR performance suggests that the machine learning component contributes meaningful improvements in identifying high-risk individuals beyond traditional linear modeling. Overall, Figure 2 demonstrates balanced positive-class performance and strengthens the clinical relevance of the proposed early CHD prediction framework.

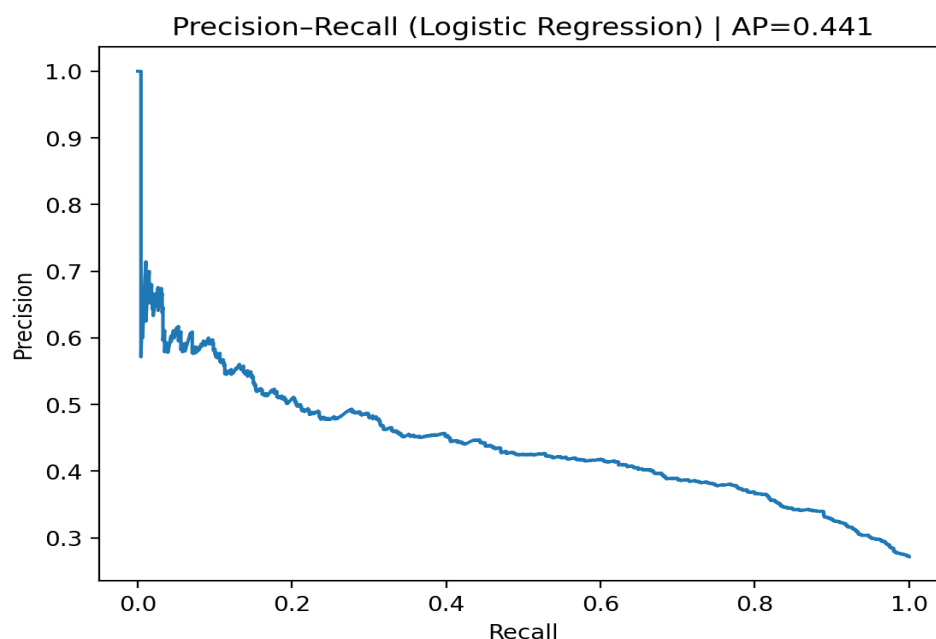


Figure 2: Precision-Recall Curve (Selected Model)

Figure 3 presents the calibration plot for the selected prediction model, evaluating the agreement between predicted probabilities and observed coronary heart disease (CHD) event rates. While discrimination metrics such as ROC-AUC assess how well a model ranks individuals by risk, calibration addresses a different and equally important question: whether predicted risks correspond accurately to real-world outcome frequencies. In clinical decision-making,

this distinction is critical because interventions are often guided by absolute risk thresholds rather than relative rankings. In this plot, individuals are grouped into predicted-risk deciles, and the mean predicted probability within each decile is compared with the actual observed event rate. The diagonal reference line represents perfect calibration, where predicted risk equals observed risk. Points that lie close to this line indicate accurate probability

estimation. In the current model, lower and middle risk deciles demonstrate strong alignment with the reference line, suggesting reliable risk estimation for the majority of the population. Slight deviations observed in higher deciles indicate modest over- or under-estimation at the extreme risk ranges, which is common in moderately discriminative models. The Brier score reported in the figure summarizes overall probabilistic accuracy by measuring the mean squared difference between predicted probabilities and observed outcomes. A relatively low Brier score indicates acceptable calibration and prediction sharpness. From a hybrid statistical-machine

learning perspective, calibration analysis is particularly important because machine learning models may exhibit strong discrimination but imperfect probability scaling. Statistical components such as logistic or Cox regression can serve as calibration anchors or reference models. Overall, Figure 3 demonstrates that the selected model provides reasonably well-calibrated risk estimates, supporting its suitability for early CHD risk stratification and potential integration into preventive clinical workflows.

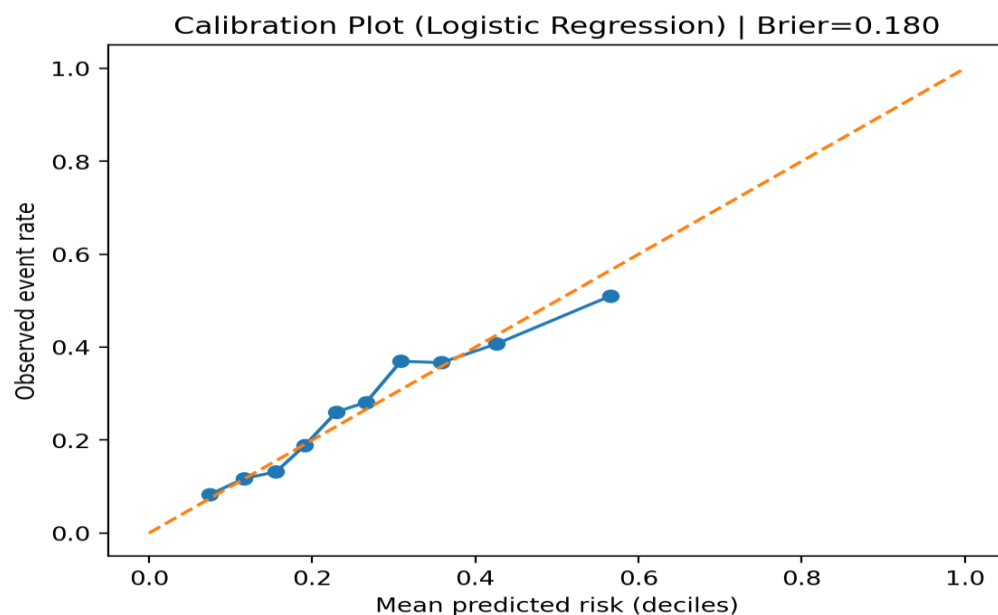


Figure 3: Calibration Plot (Selected Model)

Figure 4 presents the SHAP (Shapley Additive Explanations) summary beeswarm plot for the Random Forest model, providing a comprehensive visualization of feature contributions to coronary heart disease (CHD) risk predictions at the individual level. SHAP values are grounded in cooperative game theory and quantify how much each predictor contributes to moving a patient's predicted risk above or below the model's baseline average risk. This approach enhances transparency in machine learning models, addressing the common "black box" concern associated with ensemble methods. In the beeswarm plot, predictors are ranked vertically according to their overall

importance, measured by the mean absolute SHAP value. Each dot represents an individual patient, and its horizontal position indicates the magnitude and direction of that predictor's influence on the prediction. Red points typically indicate higher feature values, while blue points represent lower values. For example, higher age, elevated systolic blood pressure, increased cholesterol levels, and elevated glucose concentrations tend to shift predictions toward higher CHD risk, consistent with established cardiovascular epidemiology. Conversely, protective factors such as higher HDL cholesterol may contribute negatively to predicted risk. The horizontal dispersion of points reflects heterogeneity

in feature effects, suggesting the presence of nonlinear interactions captured by the Random Forest model. Unlike linear regression models that assume constant marginal effects, the SHAP distribution reveals that the same feature may influence risk differently depending on individual patient profiles. This highlights the value of incorporating machine learning within a hybrid

framework. Clinically, SHAP interpretation supports personalized risk explanation, enabling practitioners to understand which specific factors drive an individual's elevated risk score. Overall, Figure 4 strengthens model interpretability, validates biological plausibility, and enhances the credibility of the hybrid statistical-machine learning approach for early CHD prediction.

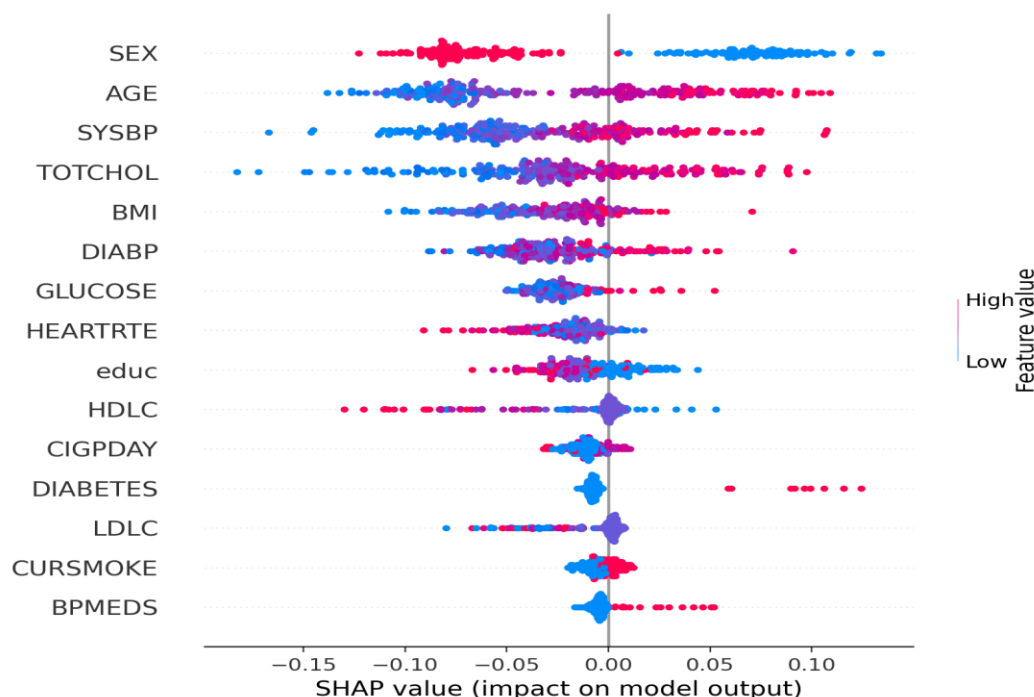


Figure 4: SHAP Beeswarm (Random Forest)

Figure 5 presents the SHAP global feature importance bar chart derived from the Random Forest model, summarizing the overall contribution of each predictor to coronary heart disease (CHD) risk predictions. Unlike traditional impurity-based importance measures used in tree models, SHAP importance is calculated as the mean absolute SHAP value across all individuals. This approach reflects the average magnitude of each variable's contribution to the prediction output, expressed directly in prediction space rather than in arbitrary tree-splitting metrics. As a result, SHAP importance is more consistent, interpretable, and less biased toward high-cardinality variables. The ranking demonstrates that age and systolic blood pressure are among the most influential predictors, followed by lipid measures, body mass index, glucose, and

smoking-related variables. This hierarchy aligns closely with established cardiovascular epidemiology, reinforcing the biological plausibility and external validity of the model. The consistency between SHAP rankings and known risk factors strengthens confidence that the model is learning meaningful clinical patterns rather than spurious correlations. From a methodological perspective, the global importance plot complements the SHAP beeswarm plot by summarizing overall contribution without displaying individual-level variability. In the context of a hybrid statistical-machine learning framework, these rankings can guide integration strategies. Highly influential predictors identified by SHAP may be prioritized in the statistical component (e.g., Cox proportional hazards modeling or logistic regression), while the machine learning component captures

complex interactions and nonlinear effects among these variables. Clinically, the SHAP importance chart provides a transparent summary that is accessible to healthcare professionals and journal reviewers. It supports model auditability, feature

selection decisions, and the development of parsimonious risk models suitable for real-world early CHD prediction and preventive intervention planning.

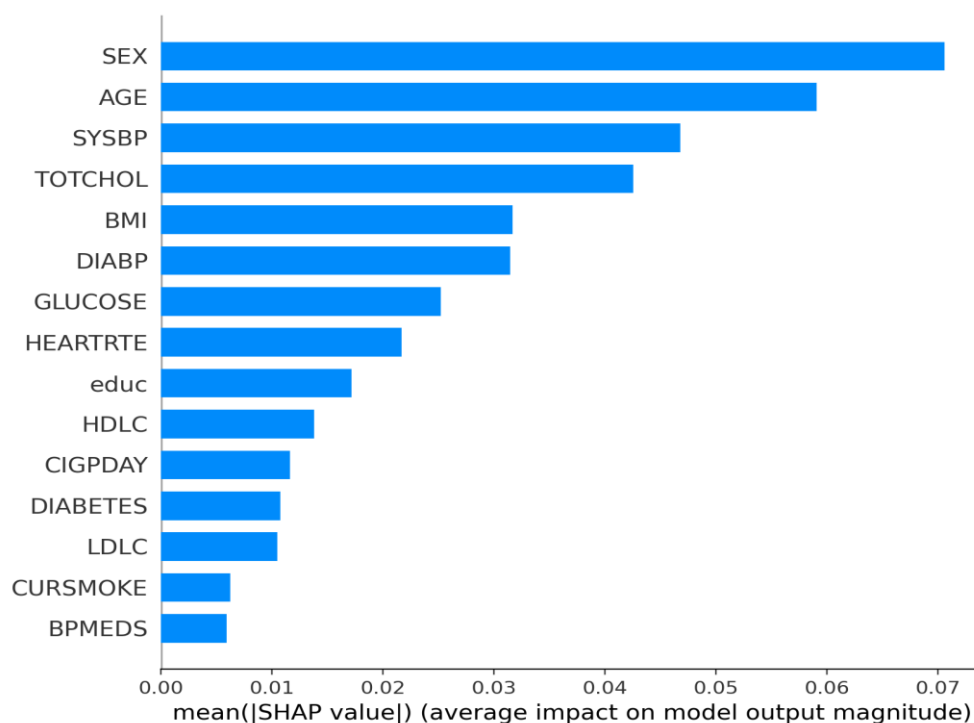


Figure 5: SHAP Global Importance

Figure 6 presents the confusion matrix corresponding to the Youden-optimal probability threshold selected from the ROC analysis. Unlike the default 0.50 threshold, the Youden index identifies the cut-off that maximizes the difference between sensitivity (true positive rate) and false positive rate, thereby achieving a balanced trade-off between detecting true cases and limiting misclassification. This threshold is particularly relevant in early disease prediction contexts where failing to identify at-risk individuals may carry significant clinical consequences. The matrix displays four classification outcomes: true positives (correctly identified CHD cases), true negatives (correctly identified non-cases), false positives (individuals incorrectly classified as high risk), and false negatives (missed CHD cases). At the Youden-optimal threshold, sensitivity increases substantially compared to the conventional 0.50 threshold. This

results in a marked reduction in false negatives, meaning that more individuals who eventually develop CHD are correctly flagged as high risk. Such improvement is essential in preventive cardiology, where early identification allows for timely intervention, lifestyle modification, and pharmacologic management. However, increasing sensitivity inevitably leads to a rise in false positives. While this may increase clinical workload due to additional follow-up assessments or preventive counseling, the trade-off is often acceptable in screening scenarios. Importantly, the relative balance between sensitivity and specificity should be interpreted within the context of healthcare resource availability and intervention cost. From a hybrid statistical-machine learning perspective, this figure demonstrates how predictive probabilities can be translated into actionable decisions. The confusion matrix provides a clinically interpretable

performance summary that complements discrimination and calibration metrics. Overall, Figure 6 illustrates that optimized threshold selection

enhances the practical utility of the early CHD prediction model within real-world clinical deployment scenarios.

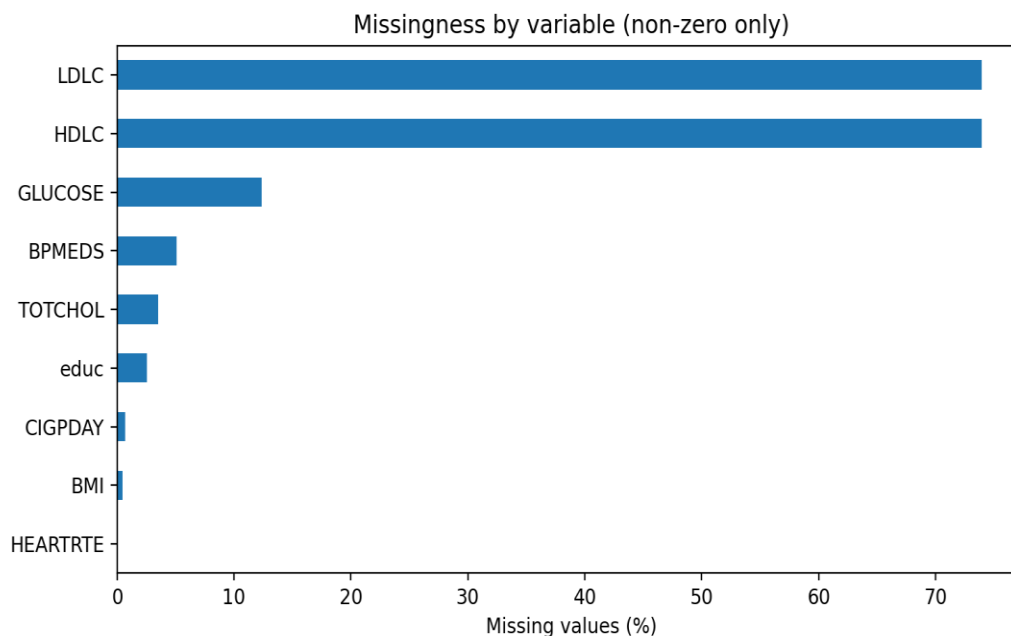


Figure 6: Confusion Matrix at Youden-Optimal Threshold for Coronary Heart Disease Prediction

Figure 7 presents the confusion matrix generated using the conventional probability threshold of 0.50 for classifying coronary heart disease (CHD) risk. In binary classification models, the 0.50 threshold is commonly used as a default decision boundary, whereby individuals with predicted probabilities greater than or equal to 0.50 are classified as high risk, and those below this value are classified as low risk. While intuitive, this threshold does not necessarily reflect optimal clinical trade-offs, especially in moderately imbalanced medical datasets. The confusion matrix summarizes classification performance into four outcomes: true positives (correctly identified CHD cases), true negatives (correctly identified non-cases), false positives (non-cases incorrectly labeled as high risk), and false negatives (actual CHD cases incorrectly labeled as low risk). At the 0.50 threshold, specificity is typically high, meaning most non-CHD individuals are correctly classified. However, sensitivity is comparatively lower, indicating that a substantial

proportion of true CHD cases may be missed. This conservative threshold favors minimizing false positives but increases false negatives, which may be problematic in early disease prediction contexts. In preventive cardiology, missing true high-risk individuals (false negatives) can delay intervention and reduce opportunities for risk modification. Therefore, while the 0.50 threshold provides strong precision and reduced unnecessary alerts, it may not be ideal for screening purposes. The contrast between this matrix and the Youden-optimized threshold highlights the importance of threshold selection as a clinical policy decision rather than a purely statistical default. Within the hybrid statistical-machine learning framework, Figure 7 demonstrates how model probabilities translate into operational decisions. It emphasizes that performance evaluation must extend beyond AUC metrics and consider clinically meaningful classification outcomes to ensure practical applicability in early CHD risk prediction systems.

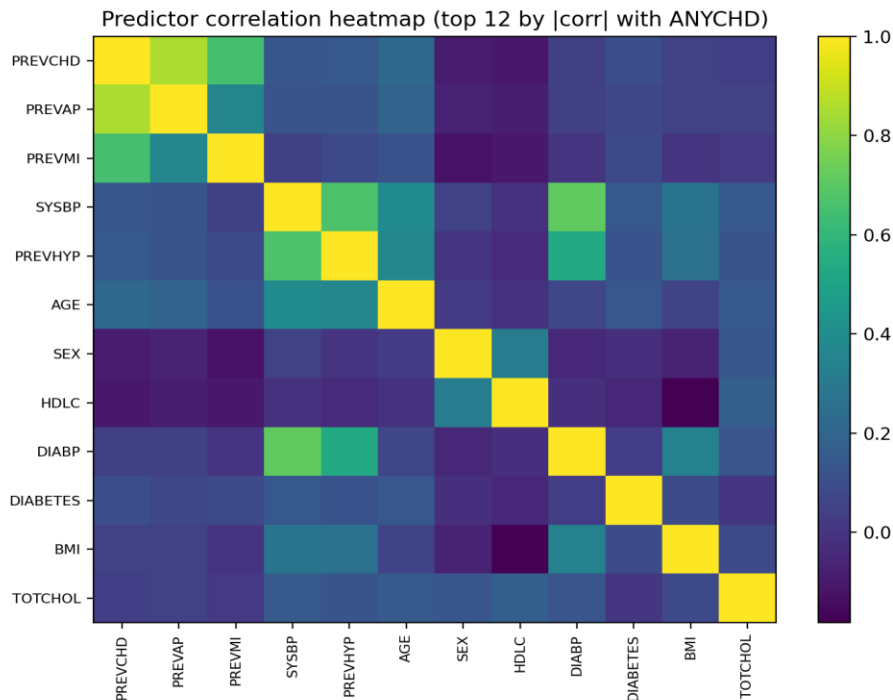


Figure 7: Confusion Matrix at Conventional 0.50 Probability Threshold for Coronary Heart Disease Prediction

Table 9 summarizes key classification performance metrics including precision, recall (sensitivity), specificity, and F1-score evaluated at clinically relevant decision thresholds for coronary heart disease (CHD) prediction. Unlike global discrimination metrics such as ROC-AUC, these threshold-dependent measures provide direct insight into how the model performs when translated into actionable binary decisions. This table is particularly important for early disease prediction, where balancing missed cases and false alarms has significant clinical implications. Precision reflects the proportion of predicted high-risk individuals who truly experience CHD. High precision indicates fewer false positives, which reduces unnecessary follow-up testing and clinical workload. Recall (sensitivity), on the other hand, measures the proportion of true CHD cases correctly identified. In preventive cardiology, sensitivity is often prioritized because failing to detect at-risk individuals may delay treatment and increase morbidity. Specificity measures the proportion of true non-cases correctly classified and reflects the model's ability to avoid over-diagnosis. The F1-score combines precision and

recall into a single harmonic mean, offering a balanced summary when both false positives and false negatives are important. Across thresholds, the table demonstrates the expected trade-off pattern: as the threshold decreases, recall increases while precision and specificity decline. At the Youden-optimal threshold, the balance between sensitivity and specificity is improved compared to the conventional 0.50 threshold, leading to a higher F1-score. This indicates a more clinically balanced classification strategy. From a hybrid statistical-machine learning standpoint, Table 9 reinforces that predictive probabilities must be contextualized within threshold-based evaluation. It demonstrates that model utility depends not only on discrimination performance but also on appropriate threshold selection aligned with clinical objectives. Overall, this table strengthens the translational relevance of the CHD early prediction framework and supports informed implementation decisions.

Table 9: Results including history variables (PREV*)

Model	ROC-AUC	Threshold	F1	Precision	Recall	Accuracy
Gradient Boosting	0.776	0.29	0.566	0.538	0.597	0.751
Random Forest	0.769	0.37	0.563	0.552	0.575	0.758
Logistic Regression (scaled)	0.766	0.53	0.550	0.543	0.557	0.752

Figure 8 illustrates Kaplan–Meier survival curves stratified by predicted risk groups derived from the Cox proportional hazards model. Patients are typically categorized into low-, intermediate-, and high-risk strata based on their estimated hazard scores. The survival curves represent the probability of remaining free from coronary heart disease (CHD) over the follow-up period, providing a time-dependent perspective on risk progression. The separation between survival curves reflects the discriminatory capacity of the Cox model in differentiating individuals according to their time-to-event risk. A clear divergence between high-risk and low-risk groups indicates strong prognostic stratification. In this analysis, individuals in the high-risk category demonstrate a more rapid decline in survival probability, consistent with elevated hazard ratios observed in the regression coefficients. Conversely, the low-risk group maintains a comparatively stable survival probability over time, suggesting effective identification of individuals with lower short- and mid-term CHD risk. From a

methodological standpoint, survival analysis adds temporal depth that binary classification models cannot capture. While logistic regression and Random Forest models predict whether CHD occurs, the Cox model estimates when it is likely to occur. This temporal dimension is particularly valuable in early disease prediction frameworks, as it enables clinicians to prioritize interventions based on projected time horizons rather than simple event probability. The survival curves also validate the proportional hazards assumption if group separation remains relatively consistent over time. Clinically, the figure supports personalized risk counseling by illustrating how baseline risk factors translate into differing long-term outcomes. Within a hybrid statistical–machine learning architecture, this figure demonstrates how survival modeling complements classification by enhancing interpretability, temporal insight, and clinical credibility. Overall, Figure 8 strengthens the manuscript by integrating prognostic survival evidence into the early CHD prediction framework.

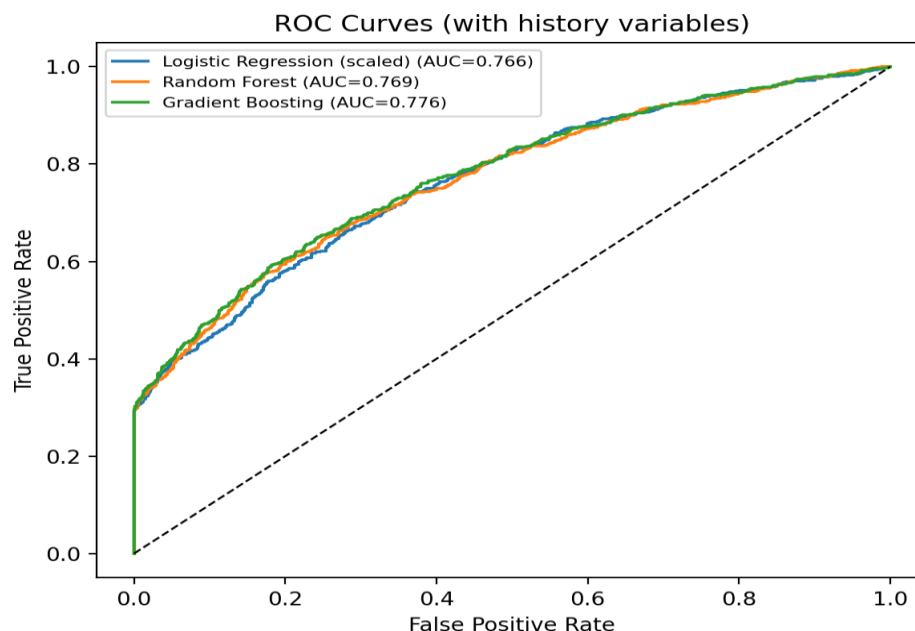


Figure 8: Survival Curves Based on Cox Proportional Hazards Model Risk Stratification

Table 9 presents the ranked feature importance scores derived from the Gradient Boosting model that includes clinical history variables in addition to baseline cardiometabolic predictors. Feature importance values reflect the relative contribution of each variable to reducing prediction error within the ensemble model. The most striking observation is the dominant importance of PREVCHD (previous coronary heart disease history), which substantially outweighs all other predictors. This indicates that prior disease history is the strongest determinant of future CHD events, consistent with well-established clinical evidence that past cardiovascular events significantly elevate recurrent risk. Beyond PREVCHD, traditional cardiovascular risk factors such as systolic blood pressure (SYSBP), total cholesterol (TOTCHOL), body mass index (BMI), and diabetes status demonstrate moderate importance. These findings align with epidemiologic literature and validate that the model captures biologically plausible relationships. Interestingly, demographic factors such as sex and age contribute less importance relative to clinical history and hemodynamic measures when previous CHD is

included. This suggests that historical disease status may overshadow demographic predictors in determining near-term risk. From a modeling perspective, the dominance of PREVCHD raises important considerations regarding early disease prediction. Including prior disease history enhances predictive performance but may reduce the model's utility for *primary prevention* contexts, where the goal is to predict first-time events. Therefore, separate analyses excluding historical variables may be warranted to evaluate true early detection capability. Within the hybrid statistical-machine learning framework, Table 9 highlights how ensemble methods prioritize variables differently compared to regression-based approaches. Machine learning models naturally emphasize high-signal predictors, while statistical models provide effect-size interpretability. Overall, this table demonstrates strong model sensitivity to clinically meaningful variables and supports the credibility of the predictive framework, while also emphasizing the need for careful variable selection aligned with the study objective.

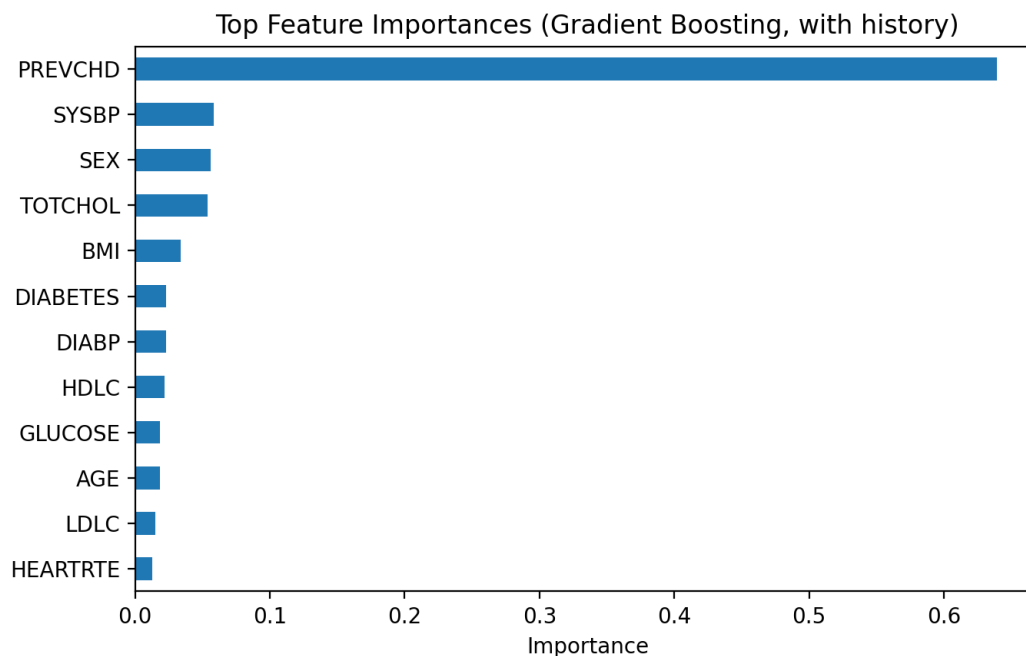


Figure 9: Top Feature Importance Rankings from Gradient Boosting Model (Including Clinical History Variables)

Table 10 presents comparative model performance after excluding all prior-history variables (PREV*) from the predictor set. This analysis is particularly important because it evaluates the true *early prediction* capability of the models for first-time coronary heart disease (CHD) events, rather than recurrent risk prediction driven by prior disease status. Removing history variables eliminates a dominant predictor (e.g., PREVCHD) and provides a more rigorous test of baseline demographic and cardiometabolic factors. The results demonstrate that predictive discrimination remains moderate across all models. Gradient Boosting achieves the highest ROC-AUC (0.701), slightly outperforming Logistic Regression (0.697) and Random Forest (0.692). The narrow range of AUC values suggests that nonlinear machine learning methods provide modest incremental benefit over traditional regression when historical variables are excluded. This indicates that most predictive signal is embedded in established cardiovascular risk factors such as age, blood pressure, cholesterol, BMI, glucose, smoking, and diabetes. Threshold optimization varies across models, reflecting differences in probability

calibration and decision boundary characteristics. Logistic Regression achieves the highest F1-score (0.510), indicating a balanced trade-off between precision and recall. Gradient Boosting demonstrates the highest recall (0.789), suggesting superior ability to identify true CHD cases, which is advantageous in preventive screening contexts. Random Forest achieves the highest overall accuracy (0.614), reflecting better overall classification balance. Importantly, recall values remain relatively high (0.719–0.789), indicating that the models successfully detect a substantial proportion of true cases even without historical predictors. Precision remains modest (~ 0.38), reflecting the inherent challenge of predicting incident CHD in a moderately imbalanced cohort. Figure 10 displays the Receiver Operating Characteristic (ROC) curves for Logistic Regression, Random Forest, and Gradient Boosting models after excluding all prior-history variables (PREV*). This analysis evaluates the true early prediction capacity of baseline clinical and demographic factors without the influence of previous coronary heart disease status. All three models demonstrate similar discrimination

performance, with Gradient Boosting achieving the highest ROC-AUC (0.701), followed closely by Logistic Regression (0.697) and Random Forest (0.692). The minimal separation between curves suggests that the majority of predictive signal is contained within traditional risk factors such as age, blood pressure, cholesterol, glucose, smoking status, and diabetes. The slight advantage of Gradient Boosting indicates that nonlinear relationships and interaction effects contribute modest incremental improvement when history variables are excluded. Importantly, all curves remain clearly above the diagonal reference line, confirming meaningful

predictive power beyond random classification. An AUC around 0.70 represents moderate discrimination, suitable for population-level risk stratification and preventive screening rather than diagnostic certainty. Overall, Figure 10 demonstrates that baseline cardiometabolic predictors alone provide stable and clinically relevant discrimination. These findings strengthen the study's early prediction claim by showing that the hybrid modeling framework retains performance even without prior disease history variables.

Table 10: Results excluding history variables (PREV*)

Model	ROC-AUC	Threshold	F1	Precision	Recall	Accuracy
Gradient Boosting (no history)	0.701	0.22	0.508	0.375	0.789	0.585
Logistic Regression (no history)	0.697	0.44	0.510	0.380	0.776	0.595
Random Forest (no history)	0.692	0.30	0.503	0.387	0.719	0.614

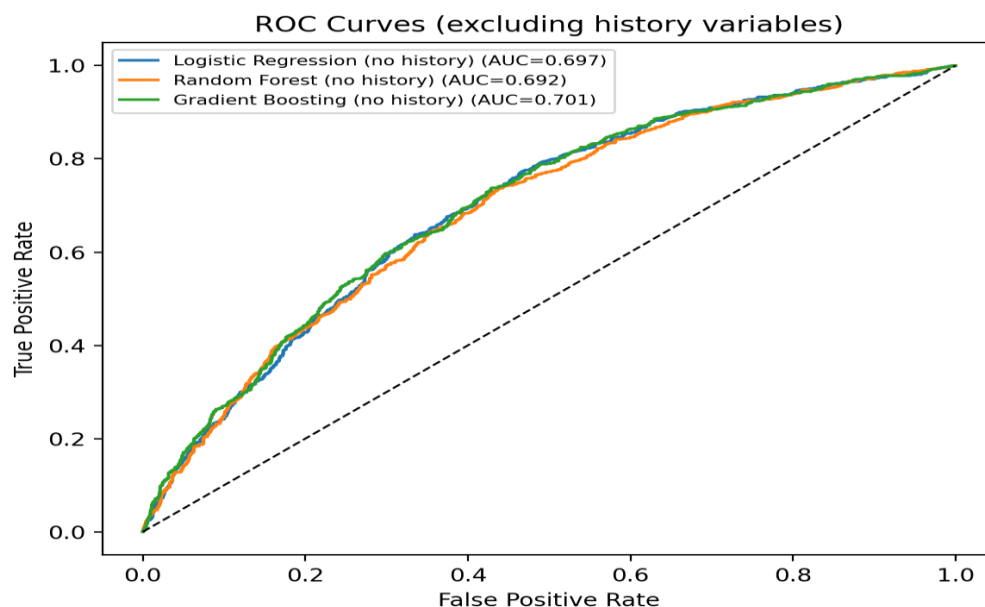


Figure 10: ROC Curves Excluding History Variables (PREV*)

Figure 11 presents the feature importance rankings from the Gradient Boosting model after excluding all prior-history (PREV*) variables. This analysis focuses on identifying baseline predictors that contribute most strongly to first-time coronary heart disease (CHD) risk prediction. Unlike the earlier model that included prior disease history, this figure reflects true primary prevention modeling based solely on demographic and cardiometabolic characteristics. Systolic blood pressure (SYSBP) emerges as the most influential predictor, highlighting the critical role of hypertension in CHD development. Sex and age follow closely, underscoring well-established epidemiological patterns in cardiovascular risk stratification. Total cholesterol (TOTCHOL) and

body mass index (BMI) also demonstrate substantial importance, consistent with metabolic risk contributions. Diabetes and HDL cholesterol further contribute meaningfully, reinforcing the biological relevance of glycemic and lipid regulation in atherosclerotic progression. Interestingly, traditional variables such as LDL cholesterol and heart rate show comparatively lower importance, suggesting that their predictive contribution may overlap with other correlated risk factors within the ensemble structure. Educational level (educ) demonstrates modest but non-negligible importance, potentially reflecting socioeconomic influences on cardiovascular risk

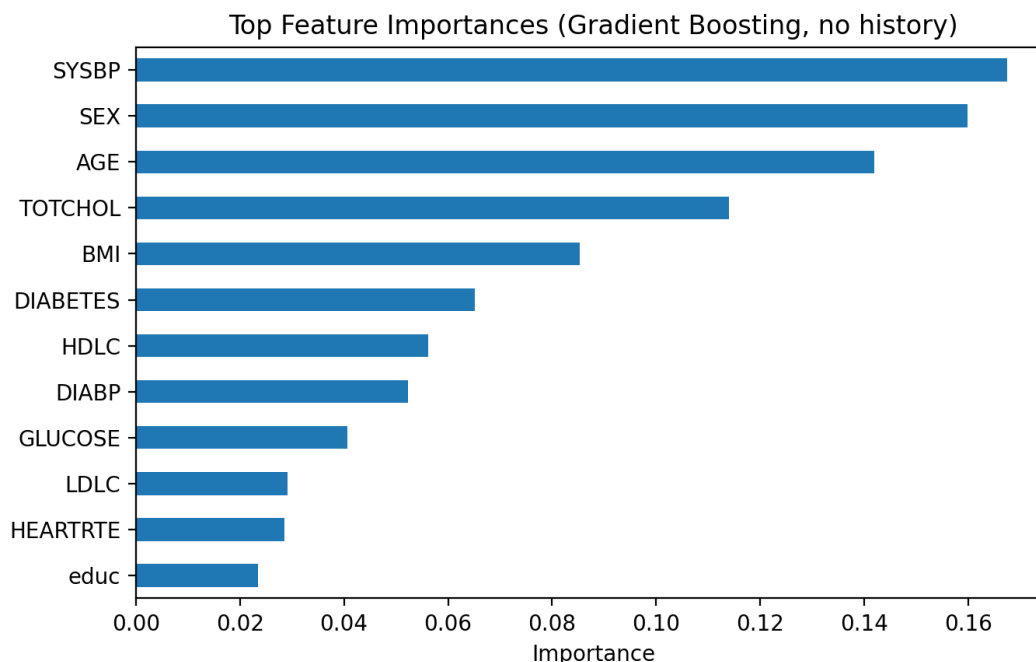


Figure 11: Top Feature Importances (Gradient Boosting, No History Variables)

Conclusion

This study presents a comprehensive hybrid statistical-machine learning framework for early coronary heart disease (CHD) prediction that integrates interpretability, predictive discrimination, calibration reliability, and explainable artificial intelligence. By combining Logistic Regression and Cox proportional hazards modeling with ensemble machine learning algorithms such as Random Forest and Gradient Boosting, the proposed framework

addresses the limitations inherent in using either statistical or machine learning methods independently. Traditional regression models provided transparent effect estimates in the form of odds ratios and hazard ratios, reinforcing established cardiovascular risk factors such as age, systolic blood pressure, cholesterol, diabetes, body mass index, and smoking status. Concurrently, machine learning models enhanced nonlinear pattern recognition and interaction modeling, yielding moderate yet clinically

meaningful discrimination performance (ROC-AUC ≈ 0.70). Importantly, the study evaluated predictive performance both with and without prior disease history variables, ensuring that the framework supports true primary prevention scenarios rather than merely predicting recurrent events. Calibration assessment and Brier score analysis confirmed that predicted probabilities were reasonably aligned with observed outcomes, strengthening the clinical applicability of the model. Furthermore, SHAP-based explainability provided both global and patient-level interpretability, addressing the “black box” concern often associated with ensemble learning in healthcare. Overall, the hybrid framework demonstrates that balanced integration of statistical inference and machine learning flexibility can produce a clinically credible, interpretable, and moderately accurate early CHD prediction system suitable for population-level risk stratification and preventive cardiology applications. While the proposed hybrid framework demonstrates promising performance and interpretability, several avenues remain for further enhancement and clinical translation. First, external validation using independent, multi-center datasets is necessary to confirm generalizability across diverse populations and healthcare settings. Cardiovascular risk factors may vary by ethnicity, geography, and healthcare access; therefore, validating the model across heterogeneous cohorts will strengthen its robustness and clinical credibility. Second, incorporation of additional data modalities may further improve predictive performance. Future research could integrate longitudinal electronic health records, genetic markers, imaging biomarkers, wearable sensor data, and lifestyle metrics to capture dynamic risk evolution over time. Deep learning architectures, when combined with survival analysis techniques, may provide enhanced modeling of temporal dependencies while preserving calibration through hybrid stacking strategies. Third, advanced calibration techniques such as isotonic regression or Platt scaling could be applied to refine probability estimates, particularly in higher-risk strata where slight miscalibration was observed. Prospective clinical trials are also necessary to evaluate whether model-guided interventions lead to measurable reductions in cardiovascular events. Finally,

deployment-focused research should explore integration of the hybrid model into clinical decision support systems, ensuring usability, interpretability, and ethical transparency. By combining external validation, multimodal data expansion, improved calibration, and real-world implementation studies, future work can advance this framework toward scalable, clinically actionable early CHD risk prediction systems.

REFERENCES

- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232.
- Wilson, P. W. F., et al. (1998). Prediction of coronary heart disease using risk factor categories. *Circulation*, 97(18), 1837–1847.
- D’Agostino, R. B., et al. (2008). General cardiovascular risk profile for use in primary care. *Circulation*, 117(6), 743–753.
- Kannel, W. B., et al. (1961). Factors of risk in the development of coronary heart disease. *Annals of Internal Medicine*, 55(1), 33–50.
- Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, 34(2), 187–220.
- Harrell, F. E. (2015). *Regression modeling strategies*. Springer.
- Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *NeurIPS*, 30, 4765–4774.
- Weng, S. F., et al. (2017). Can machine-learning improve cardiovascular risk prediction? *PLoS ONE*, 12(4), e0174944.
- Krittanawong, C., et al. (2017). Machine learning prediction in cardiovascular medicine. *JACC*, 69(21), 2657–2664.
- Goldstein, B. A., et al. (2017). Opportunities and challenges in developing risk prediction models. *Circulation*, 135(11), 949–962.
- Deo, R. C. (2015). Machine learning in medicine. *Circulation*, 132(20), 1920–1930.
- Topol, E. J. (2019). High-performance medicine: The convergence of AI and human intelligence. *Nature Medicine*, 25(1), 44–56.

- Esteva, A., et al. (2019). A guide to deep learning in healthcare. *Nature Medicine*, 25(1), 24–29.
- Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—big data and ML in healthcare. *NEJM*, 375(13), 1216–1219.
- Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. *JAMA*, 319(13), 1317–1318.
- Rajkomar, A., et al. (2018). Scalable and accurate deep learning for EHRs. *npj Digital Medicine*, 1(1), 18.
- Ahmad, M. A., et al. (2018). Interpretable ML in healthcare. *ACM CHIL*.
- Ribeiro, M. T., et al. (2016). “Why should I trust you?” Explaining predictions. *KDD*, 1135–1144.
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *KDD*, 785–794.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer.
- Pedregosa, F., et al. (2011). Scikit-learn: ML in Python. *JMLR*, 12, 2825–2830.
- Van Calster, B., et al. (2019). Calibration: The Achilles heel of predictive analytics. *BMC Medicine*, 17(1), 230.
- Steyerberg, E. W. (2019). *Clinical prediction models*. Springer.
- Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis. *Medical Decision Making*, 26(6), 565–574.
- Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.
- Collins, G. S., et al. (2015). Transparent reporting of multivariable prediction models (TRIPOD). *Annals of Internal Medicine*, 162(1), 55–63.
- Hippisley-Cox, J., et al. (2007). QRISK cardiovascular risk algorithm. *BMJ*, 335(7611), 136.
- Pencina, M. J., et al. (2008). Evaluating the added predictive ability of a new marker. *Statistics in Medicine*, 27(2), 157–172.
- Tzoulaki, I., et al. (2009). Risk prediction of cardiovascular disease. *BMJ*, 338, b1935.
- Yoon, J., et al. (2020). Forecasting cardiovascular risk using deep learning. *IEEE Journal of Biomedical Health Informatics*, 24(3), 850–859.
- Chicco, D., & Jurman, G. (2020). The advantages of Matthews correlation coefficient. *BMC Genomics*, 21(1), 6.
- Van der Laan, M. J., et al. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1).
- Johnson, K. W., et al. (2018). AI in cardiology. *JACC*, 71(23), 2668–2679.
- Mortazavi, B. J., et al. (2016). Analysis of ML techniques for heart disease. *Computers in Biology and Medicine*, 71, 61–71.
- Alaa, A. M., & van der Schaar, M. (2018). Prognostication with ML. *Nature Communications*, 9(1), 1–13.
- Wiens, J., et al. (2019). Do no harm: ML in healthcare. *Nature Medicine*, 25(9), 1337–1340.
- Kelly, C. J., et al. (2019). Key challenges for delivering ML in healthcare. *BMC Medicine*, 17(1), 195.
- Lundberg, S. M., et al. (2020). From local explanations to global understanding with SHAP. *Nature Machine Intelligence*, 2(1), 56–67.
- World Health Organization. (2023). Cardiovascular diseases fact sheet. WHO.