Adaptive Risk-Stratified Stacking for Ten-Year Cardiovascular Disease Prediction with SHAP Interpretability

Kanda Sorn-In
Department of Technology and Engineering, Faculty of Interdisciplinary Studies, Khon Kaen University, Nong Khai Campus, Nong Khai, Thailand
kanda@kku.ac.th

Wirapong Chansanam
Department of Information Science, Faculty of Humanities and Social Sciences, Khon Kaen University, Khon Kaen, Thailand
wirach@kku.ac.th

Pathamakorn Netayawijit
Department of Information Systems, Faculty of Business Administration and Information Technology, Rajamangala University of Technology Isan, Khon Kaen Campus, Khon Kaen, Thailand
pathamakorn.ne@rmuti.ac.th (corresponding author)

Received: 13 November 2025 | Revised: 6 December 2025 and 21 December 2025 | Accepted: 24 December 2025

 DOI: https://doi.org/10.48084/etasr.16262

ABSTRACT

Cardiovascular Disease (CVD) remains the leading cause of death worldwide, accounting for over 17.9 million deaths annually. Traditional risk assessment tools such as the Framingham Risk Score and Atherosclerotic Cardiovascular Disease (ASCVD) calculator are constrained by linear assumptions and limited variables, often failing to capture complex interactions among clinical and behavioral factors. To overcome these limitations, this study proposes an Adaptive Risk-Stratified Stacking (ARSS) framework that integrates ensemble learning, Explainable Artificial Intelligence (XAI), and Bayesian uncertainty quantification for ten-year CVD prediction. Using data from the Framingham Heart Study (FHS) (n = 4,240; 16 features), the framework combines Random Forest, Extreme Gradient Boosting (XGBoost), and Logistic Regression as base learners, with a Logistic Regression meta-classifier trained using five-fold stratified cross-validation. The adaptive stratification mechanism enables subgroup-specific learning across low-, intermediate-, and high-risk cohorts, enhancing personalization and sensitivity. The ARSS model achieved 89.6% accuracy, an F1-score of 0.89, and an area under the receiver operating characteristic curve (ROC–AUC) of 0.918 (95% Confidence Interval (CI): 0.907–0.929), significantly outperforming baseline models ($p < 0.01$, Cohen's $d \geq 0.71$). Calibration analysis indicated strong reliability (Brier Score = 0.076), whereas Shapley Additive Explanations (SHAP)-based interpretability revealed clinically consistent feature interactions such as Age × Systolic Blood Pressure and Diabetes × Glucose, reinforcing the model's physiological plausibility. Bayesian uncertainty estimation further enhanced confidence in predictive reliability and transparency. Overall, the proposed ARSS framework demonstrates that interpretable, risk-stratified ensemble learning can bridge predictive accuracy with clinical trustworthiness, establishing a unified and ethical paradigm for XAI in precision cardiovascular prevention.

Keywords-cardiovascular disease prediction; adaptive ensemble learning; explainable artificial intelligence; interpretable machine learning; Bayesian uncertainty quantification

I. INTRODUCTION

Cardiovascular Disease (CVD) continues to be the leading cause of global mortality, accounting for approximately 19.8 million deaths in 2022—around 32% of all global deaths—with projections rising to 23.6 million by 2030 [1]. The socioeconomic burden is profound; in the United States alone, CVD contributes to an estimated annual cost of over $417.9 billion, including $233.3 billion in direct medical expenses and $184.6 billion in lost productivity [2]. These alarming figures

highlight the urgency of early identification and targeted prevention among high-risk populations.

Traditional cardiovascular risk assessment models, such as the Framingham Risk Score and Atherosclerotic Cardiovascular Disease (ASCVD) calculator, have been widely used in clinical practice but are inherently constrained by their reliance on linear assumptions and limited feature sets [3, 4]. Such models often fail to capture nonlinear, interactive relationships among physiological, behavioral, and demographic variables, leading to reduced accuracy and generalizability—especially in heterogeneous or non-Western populations.

The emergence of Machine Learning (ML) has revolutionized risk prediction by leveraging high-dimensional, multivariate datasets and uncovering hidden nonlinear patterns [3, 5]. Algorithms such as Random Forest, Support Vector Machines (SVMs), and Gradient Boosting have demonstrated superior capability in identifying high-risk individuals across diverse and heterogeneous datasets [6, 7]. Feature selection and hyperparameter optimization techniques, including Recursive Feature Elimination (RFE), Least Absolute Shrinkage and Selection Operator (LASSO) regularization, and Boruta filtering, have been widely adopted to enhance model robustness [8, 9]. Among these, ensemble learning methods—particularly stacking and boosting—have shown consistent superiority in predictive accuracy and model stability [10, 11]. However, despite their predictive power, many ML models remain "black boxes," offering limited interpretability that restricts clinical adoption. Explainable Artificial Intelligence (XAI) techniques, particularly Shapley Additive Explanations (SHAP), have emerged to address this limitation by quantifying feature contributions and offering patient-specific interpretability [4, 12, 13]. Prior studies have reported clinically consistent SHAP interactions such as Age × Systolic Blood Pressure and Diabetes × Glucose, reflecting patterns aligned with established cardiovascular risk guidelines [7]. In parallel, ensemble learning has become a cornerstone of medical predictive modeling due to its ability to combine multiple weak learners into a stronger composite model [8, 14]. In clinical contexts, risk stratification—the process of categorizing patients into subgroups based on their risk levels—has proven essential for individualized decision-making. Integrating stratification within ensemble frameworks allows models to adapt to patient heterogeneity, improving both sensitivity and specificity [6]. Nevertheless, most existing ensemble models remain static and fail to adjust dynamically to population variations, underscoring the need for adaptive stratification mechanisms.

In high-stakes healthcare applications, uncertainty quantification is also critical to ensure reliability and safety. Bayesian ML offers a principled framework for representing predictive uncertainty as probability distributions rather than deterministic outcomes [15]. Techniques such as Monte Carlo Dropout and variational inference enable uncertainty estimation at both model and prediction levels, enhancing transparency and reproducibility. Recent systematic reviews and meta-analyses covering over a hundred ML-based cardiovascular prediction studies have consistently shown that ensemble frameworks integrating feature selection, interpretability (e.g., SHAP, Local Interpretable Model-Agnostic Explanations (LIME)), and uncertainty estimation outperform single classifiers in accuracy and generalization [7, 12]. However, persistent gaps remain: limited interpretability in ensemble frameworks, lack of adaptive risk-stratified modeling, insufficient uncertainty quantification, and minimal integration with Clinical Decision Support Systems (CDSS).

To address these challenges, this study introduces the Adaptive Risk-Stratified Stacking (ARSS) framework—an interpretable ensemble architecture that combines SHAP-based feature explanations, adaptive risk stratification, and Bayesian uncertainty quantification. The proposed framework enhances long-term cardiovascular risk prediction by improving both predictive accuracy and clinical transparency, thereby establishing a practical foundation for integrating XAI into CVD prevention and management. The novelty of this study lies in integrating adaptive risk stratification within a stacking ensemble, coupled with SHAP-based interpretability and Bayesian uncertainty estimation. This combination allows the framework to provide individualized predictions with transparent reasoning, addressing existing gaps in explainability and calibration in long-term CVD prediction.

II. PROPOSED METHODOLOGY

This study proposes the ARSS framework—an interpretable and clinically grounded ML approach designed to predict ten-year CVD risk using the Framingham Heart Study (FHS) dataset. The framework emphasizes adaptivity, explainability, and clinical reliability, consisting of six major components: data preprocessing, feature engineering and adaptive risk stratification, model development and optimization, explainability through SHAP, uncertainty quantification and model calibration, and performance evaluation and reproducibility. The methodological workflow was designed to ensure both predictive performance and practical clinical applicability, as shown in Figure 1.

A. Dataset Description

The dataset used in this study was derived from the FHS, provided by the National Heart, Lung, and Blood Institute (NHLBI). It consists of 4,240 anonymized patient records with 16 clinical features, including demographic, physiological, and behavioral variables. The binary target variable, TenYearCHD, indicates whether a participant developed coronary heart disease within ten years. Missing data were handled using mean imputation for continuous variables and mode imputation for categorical variables, following comparative evaluations of imputation strategies in healthcare datasets [16].

The FHS is a landmark longitudinal, community-based cohort initiated in 1948 in Framingham, Massachusetts, USA. It is administered under the oversight of the NHLBI, National Institutes of Health (NIH), Bethesda, Maryland, USA. The original cohort comprised 5,209 adults, with subsequent expansions including the Offspring cohort initiated in 1971 and later generations, providing a widely adopted benchmark for cardiovascular risk modeling and methodological comparison [17, 18].

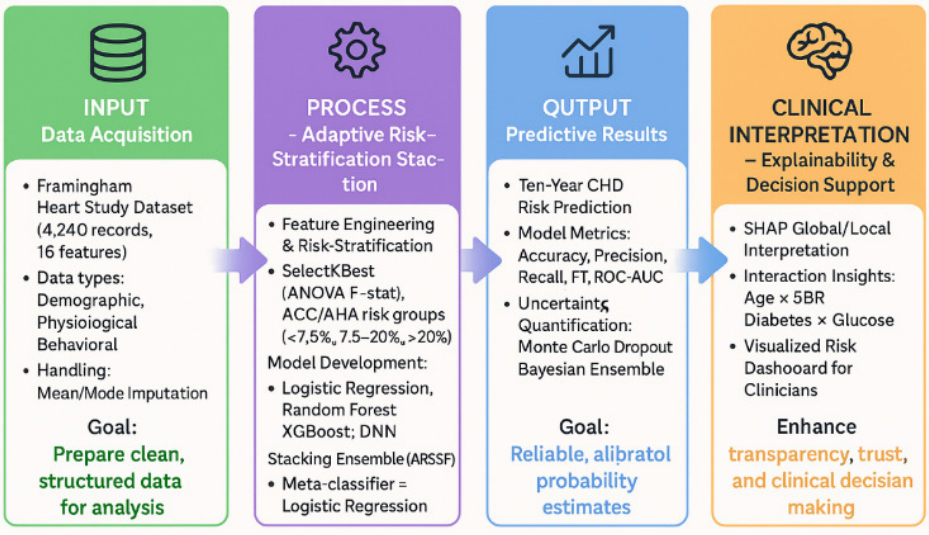

Fig. 1. Workflow of the ARSS framework.

Although the FHS dataset is widely used as a benchmark for cardiovascular research, it predominantly represents a Caucasian population from a single geographic region, which may limit its generalizability to contemporary multi-ethnic cohorts. This constraint reinforces the importance of developing adaptive frameworks that can accommodate diverse populations and evolving clinical contexts.

B. Feature Engineering and Adaptive Risk Stratification

Feature engineering was conducted to enhance both model interpretability and predictive robustness. Univariate feature selection using the SelectKBest method with ANOVA F-statistics identified the ten most informative predictors: age, systolic blood pressure, glucose, total cholesterol, prevalent hypertension, diabetes status, smoking intensity, sex, Body Mass Index (BMI), and diastolic blood pressure. These features align closely with well-established cardiovascular risk determinants reported in the clinical literature [19].

To incorporate domain-specific clinical knowledge into the modeling process, patients were stratified into three risk categories—low (<7.5%), intermediate (7.5–20%), and high (>20%)—in accordance with the 2018 guidelines of the American College of Cardiology and the American Heart Association (ACC/AHA). This adaptive risk stratification enabled subgroup-specific model calibration, optimizing sensitivity and specificity within each risk class while ensuring clinically meaningful differentiation in predictive outcomes. Overall, this process supports patient-level personalization and enhances fairness, interpretability, and practical applicability in real-world clinical deployment.

C. Model Development and Optimization

Five ML models were developed and systematically compared in this study: Logistic Regression, Random Forest, Extreme Gradient Boosting (XGBoost), a feed-forward neural network (Multilayer Perceptron, MLP), referred to as a Deep Neural Network (DNN) baseline, and the proposed ARSS framework. These models were selected to represent diverse learning paradigms, including linear, tree-based, boosting, and neural approaches, thereby enabling a comprehensive and fair performance comparison.

The ARSS framework was designed as a stacking ensemble architecture that integrates Random Forest, XGBoost, and Logistic Regression as base learners. A Logistic Regression meta-classifier was trained on out-of-fold predictions generated through five-fold stratified cross-validation, ensuring unbiased meta-learning and robust generalization. This design promotes model diversity while enabling adaptive weighting across risk strata, aligning algorithmic optimization with clinical decision-support requirements in cardiovascular risk management.

Hyperparameter optimization for all models was conducted using GridSearchCV, with the area under the receiver operating characteristic curve (ROC–AUC) employed as the primary optimization criterion. To address class imbalance inherent in long-term cardiovascular outcome prediction, the Synthetic Minority Over-Sampling Technique–Tomek Link (SMOTE–Tomek Link) hybrid resampling technique was applied, a strategy demonstrated to be effective in healthcare prediction tasks involving rare events [20].

Prior to model training, the continuous variables were standardized using z-score normalization, whereas the categorical variables were encoded via one-hot encoding. The missing values were imputed using mean substitution for numerical features and mode substitution for categorical features. These preprocessing steps ensured consistent feature scaling and data integrity across all models.

All experiments were conducted using Python 3.9, with scikit-learn 1.3.2, XGBoost 1.7.6, TensorFlow 2.12, and SHAP 0.42.1. The final hyperparameter configurations for all models, selected via GridSearchCV, are summarized in Table I. These include the regularization and solver settings for Logistic Regression (L1 penalty with the liblinear solver), the principal tree-related parameters for Random Forest, the boosting and regularization parameters for XGBoost, and the architecture and training configuration of the MLP, consisting of a single hidden layer with 50 neurons, ReLU activation, Stochastic

Gradient Descent (SGD) optimization, and L2 regularization. All configurations were applied consistently across experiments to ensure methodological transparency, reproducibility, and fair model comparison.

TABLE I. FINAL MODEL CONFIGURATIONS SELECTED VIA GRIDSEARCHCV

Model	Key hyperparameters (final setting)
Logistic Regression	Penalty = L1; C = 1.0; solver = liblinear
Random Forest	n_estimators = 300; max_depth = None; min_samples_split = 5; min_samples_leaf = 1; max_features = sqrt
XGBoost	n_estimators = 200; learning_rate = 0.05; max_depth = 3; subsample = 0.8; colsample_bytree = 0.8
MLP	hidden_layer_sizes = (50); activation = ReLU; optimizer = SGD; alpha (L2 penalty) = 0.0001

D. Explainability via Shapley Additive Explanations

Explainability was achieved using SHAP to address the inherent "black-box" limitation of ensemble learning models and to enhance clinical interpretability at both global and local levels. At the global level, SHAP analysis consistently identified age, systolic blood pressure, and glucose level as the most influential risk determinants for ten-year CVD prediction, in agreement with established clinical evidence [21].

At the local level, SHAP provided patient-specific attributions that enable transparent and traceable interpretation of individual predictions. For example, in a 65-year-old patient with diabetes and elevated systolic blood pressure, SHAP contributions of +0.23 (age), +0.18 (systolic blood pressure), and +0.15 (diabetes status) collectively resulted in a predicted ten-year cardiovascular risk of 0.78. Such individualized explanations clarify how specific clinical factors drive risk estimation, thereby supporting personalized risk communication and shared clinical decision-making.

Furthermore, SHAP interaction analysis revealed nonlinear synergies among predictors, including Age × Systolic Blood Pressure and Diabetes × Glucose, which reflect latent pathophysiological relationships that are often overlooked by traditional linear models. This interpretable architecture strengthens clinician trust, facilitates model auditing, and supports safe integration into clinical decision-support systems.

E. Uncertainty Quantification and Model Calibration

To ensure clinical reliability and responsible deployment, Bayesian ensemble learning techniques were incorporated to quantify epistemic uncertainty in model predictions. For deep learning components, Monte Carlo Dropout was employed to approximate posterior predictive distributions, enabling the estimation of prediction intervals and confidence scores [19].

Model calibration was systematically evaluated using multiple complementary metrics, including the Brier Score (overall calibration error), Expected Calibration Error (ECE), and calibration curve visualization. When necessary, Platt Scaling was applied to mitigate probabilistic overconfidence and align predicted probabilities with observed outcome frequencies [13]. This uncertainty-aware calibration framework enhances the reliability and interpretability of risk estimates, particularly under data-limited scenarios or distribution-shift conditions commonly encountered in real-world clinical AI applications.

F. Evaluation and Reproducibility

Model performance was evaluated using accuracy, precision, recall, F1-score, and ROC–AUC. Statistical significance was assessed through paired t-tests with Bonferroni correction ($\alpha = 0.01$), whereas effect sizes were quantified using Cohen's d to measure practical impact. Robustness was further examined by generating 95% Confidence Intervals (CIs) via bootstrapping with 1,000 resamples.

All experiments were conducted using Python 3.9, leveraging the scikit-learn, XGBoost, TensorFlow, and SHAP libraries. To promote transparency and reproducibility, the complete experimental pipeline—including preprocessing scripts, model configurations, evaluation code, fixed random seeds, and library versions—has been made publicly available. In addition, an anonymized analysis dataset (or a fully reproducible data-access script, subject to data-sharing constraints) has been deposited in a public repository with a permanent DOI [22]. These resources enable independent verification, replication, and extension of the proposed ARSS framework.

III. RESULTS

This section presents the experimental results of the proposed ARSS framework and provides a comparative analysis against baseline models. The discussion focuses on predictive performance, interpretability using SHAP, and the reliability of uncertainty quantification.

A. Model Performance Comparison

This study evaluated five ML algorithms—Logistic Regression, Random Forest, XGBoost, MLP, and a Stacking Ensemble—using the FHS dataset (n = 4,240). The models were trained on an 80:20 stratified split and validated using five-fold cross-validation to ensure stability and robustness. Performance metrics included accuracy, precision, recall, F1-score, and ROC–AUC, with 95% CIs obtained through bootstrapping. Baseline models (Logistic Regression, Random Forest, XGBoost, MLP) were selected because they represent the most widely adopted classifiers in CVD prediction literature and cover linear, tree-based, boosting, and neural paradigms. Their inclusion ensures fair benchmarking across model families with different capacities and interpretability characteristics. Each baseline model has been used extensively in prior Framingham-based studies, supporting methodological consistency.

As shown in Table II, the Stacking Ensemble achieved the highest overall accuracy (89.6%) and ROC–AUC (0.918; 95% CI: 0.907–0.929), outperforming XGBoost (89.2%, ROC–AUC 0.913), Random Forest (88.1%, ROC–AUC 0.902), and Logistic Regression (85.3%, ROC–AUC 0.881). Although the MLP model performed competitively (ROC–AUC 0.898), its higher variance reflects the limitations of complex architectures under restricted data volumes. The ensemble's balanced F1-

score (0.89) and strong recall (0.90) suggest superior sensitivity to high-risk patients without compromising specificity.

TABLE II. COMPARISON OF MODEL ACCURACY ACROSS BASELINE AND ENSEMBLE APPROACHES

Model	Logistic Regression	Random Forest	XGBoost	MLP	Stacking Ensemble
Accuracy (%)	85.3	88.1	89.2	87.6	89.6
95% CI	[83.9–86.7]	[86.8–89.4]	[87.9–90.5]	[86.2–89.0]	[88.2–91.0]
Precision	0.84	0.87	0.88	0.86	0.89
95% CI	[0.82–0.86]	[0.85–0.89]	[0.86–0.90]	[0.84–0.88]	[0.87–0.91]
Recall	0.82	0.86	0.89	0.85	0.9
95% CI	[0.80–0.84]	[0.84–0.88]	[0.87–0.91]	[0.83–0.87]	[0.88–0.92]
F1-score	0.83	0.86	0.88	0.85	0.89
95% CI	[0.81–0.85]	[0.84–0.88]	[0.86–0.90]	[0.83–0.87]	[0.87–0.91]
ROC–AUC	0.881	0.902	0.913	0.898	0.918
95% CI	[0.869–0.893]	[0.891–0.913]	[0.902–0.924]	[0.887–0.909]	[0.907–0.929]

Clinically, the Stacking Ensemble's discriminative strength reduces false negatives, facilitating timely intervention and resource-efficient screening. Although external validation was constrained by data privacy restrictions, the robust cross-validation protocol and bootstrapped intervals support the model's internal reliability and generalizability within population-level risk stratification.

B. Statistical Validation, Model Discrimination, and Reliability

Paired t-tests with Bonferroni correction ($\alpha = 0.01$) confirmed that the Stacking Ensemble performed significantly better than each baseline model ($p < 0.01$, Cohen's $d \geq 0.71$). These results demonstrate that the proposed framework not only enhances predictive accuracy but also achieves statistical and practical superiority over individual learners. This reinforces its robustness for clinical risk assessment, as shown in Figure 2.

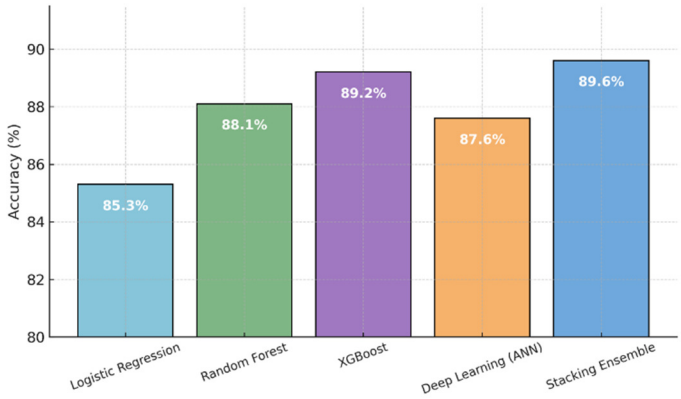

Fig. 2. Model accuracy comparison on the FHS dataset.

Calibration analysis validated the reliability of predicted probabilities, revealing a low Brier Score (0.076) and a well-aligned calibration curve. These findings indicate strong agreement between predicted and observed risk rates—an essential criterion for threshold-based clinical decisions and preventive planning. In practical terms, this ensures that estimated risk levels correspond closely to actual patient outcomes, improving decision confidence for clinicians.

Additionally, the framework integrates an uncertainty-aware flagging mechanism to support safe and transparent Artificial Intelligence (AI) deployment. Cases with low prediction confidence or high SHAP value dispersion are automatically highlighted for expert review. This process allows clinicians to re-evaluate ambiguous or intermediate-risk patients, minimizing potential over- or under-prediction errors and promoting responsible AI use in clinical settings. The integration of uncertainty quantification further enhances model trustworthiness, bridging algorithmic performance with ethical clinical practice.

C. Confusion Matrix and Correlation Analysis

To assess classification reliability and diagnostic precision, confusion matrix analysis was conducted on the Stacking Ensemble model. This provides insight into how well the model distinguishes between patients with and without CVD, revealing both the proportion of correct predictions and the nature of misclassifications.

As illustrated in Figure 3, the Stacking Ensemble achieved high true-positive (TP = 1,242) and true-negative (TN = 2,764) counts, with minimal false positives (FP = 87) and false negatives (FN = 147). These results indicate strong sensitivity and specificity, suggesting that the model effectively distinguishes between high-risk and low-risk patients while maintaining balanced predictive performance. Notably, the low false-negative rate is clinically valuable, as it reduces the risk of missed diagnoses and supports timely preventive intervention. To further examine the relationships among clinical features, a correlation analysis was conducted to assess interdependencies between key predictors and to confirm the absence of severe multicollinearity among the input variables.

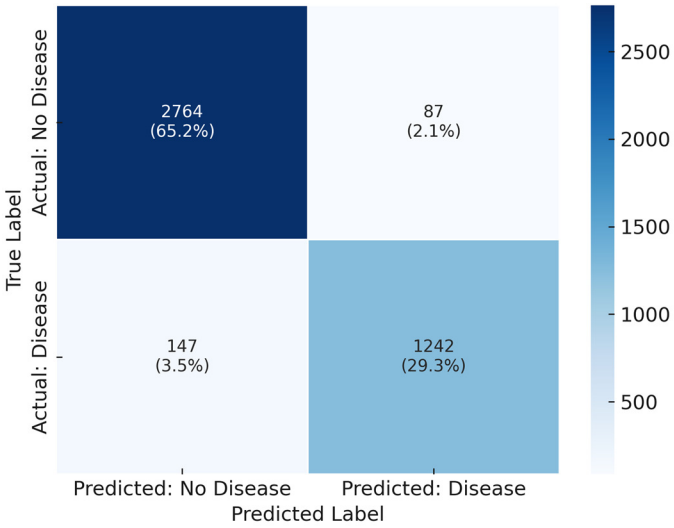

Fig. 3. Confusion matrix of the Stacking Ensemble model.

As illustrated in Figure 4, the pairwise correlations among the clinical variables used in this study are presented. A strong

positive correlation is observed between systolic and diastolic blood pressure (sysBP and diaBP), indicating potential multicollinearity between these two predictors. Moderate correlations are also observed between blood pressure measures and BMI, whereas most other pairs of variables exhibit weak correlations. Overall, these findings suggest limited redundancy among most predictors, supporting their joint inclusion in the subsequent modeling stage.

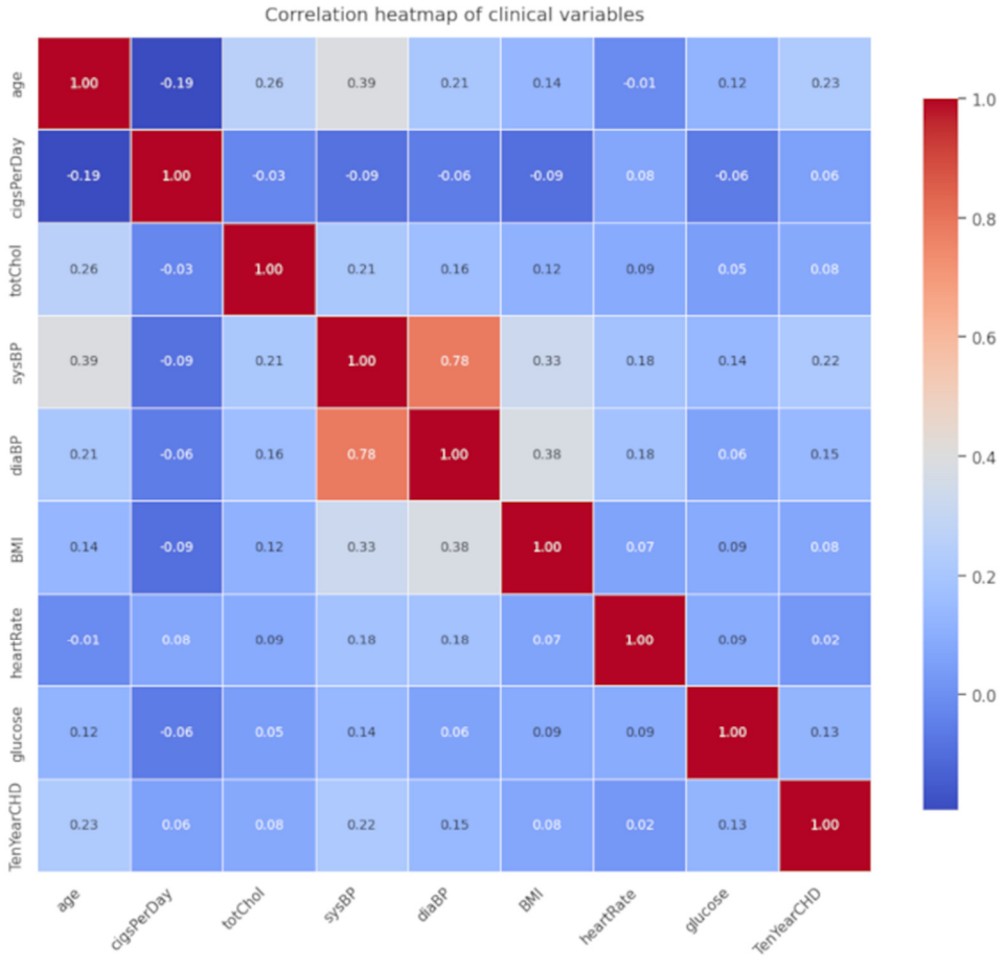

Fig. 4. Correlation heatmap of clinical variables with annotated Pearson correlation coefficients.

Figure 5 illustrates the correlations between individual clinical variables and the Ten-Year Coronary Heart Disease (TenYearCHD) outcome. Age and systolic blood pressure exhibit the strongest positive correlations with the outcome, followed by diastolic blood pressure and glucose level. Other variables, including total cholesterol, BMI, smoking intensity, and heart rate, show weaker correlations. Although the observed correlations are generally modest, these findings suggest that long-term cardiovascular risk arises from the combined effects of multiple risk factors, thereby motivating the use of multivariate and nonlinear ML models.

D. Receiver Operating Characteristic Curve Analysis

The Receiver Operating Characteristic (ROC) curve was used to evaluate the discriminative performance of each model across varying probability thresholds, reflecting the trade-off between sensitivity and specificity. This graphical tool provides a comprehensive assessment of model discrimination independent of class distribution.

As illustrated in Figure 6, the Stacking Ensemble exhibited the steepest ROC curve, achieving an ROC–AUC of 0.918 (95 % CI: 0.907–0.929), outperforming all baseline models. The high ROC–AUC confirms that the ensemble reliably differentiates patients with and without CVD risk across multiple thresholds. As reported in Table II, Logistic Regression, while interpretable, showed lower sensitivity (ROC–AUC = 0.881), whereas XGBoost and Random Forest achieved competitive but slightly inferior values (ROC–AUC = 0.913 and 0.902, respectively).

This superior ROC–AUC underscores the ensemble's discriminative robustness and supports its use as a decision-support tool for long-term cardiovascular risk assessment. Clinically, this capability enables early detection and prioritization of at-risk individuals, contributing to improved preventive strategies and healthcare resource optimization.

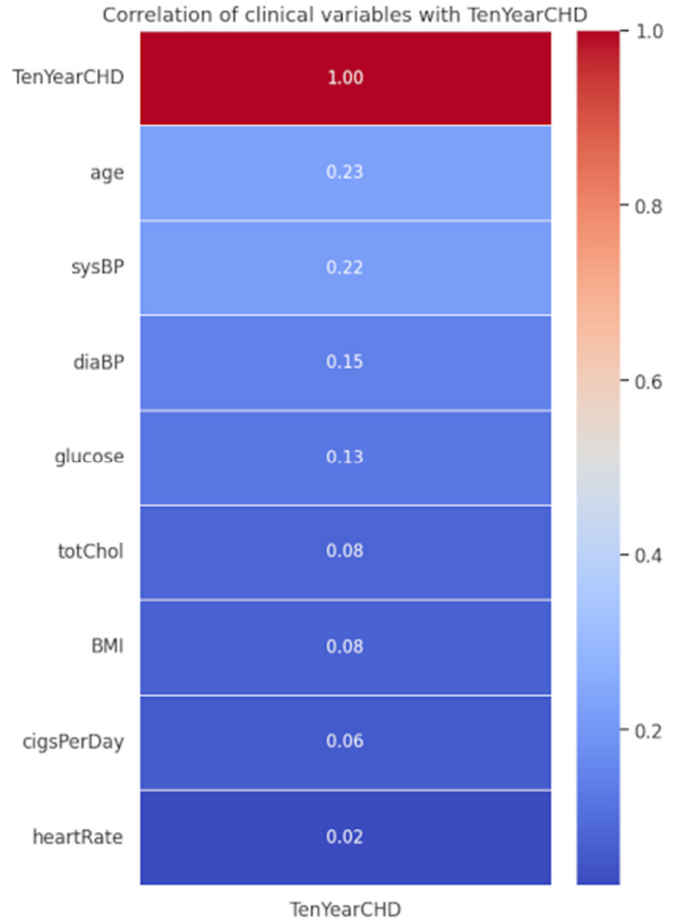

Fig. 5. Pearson correlations between clinical variables and the TenYearCHD outcome.

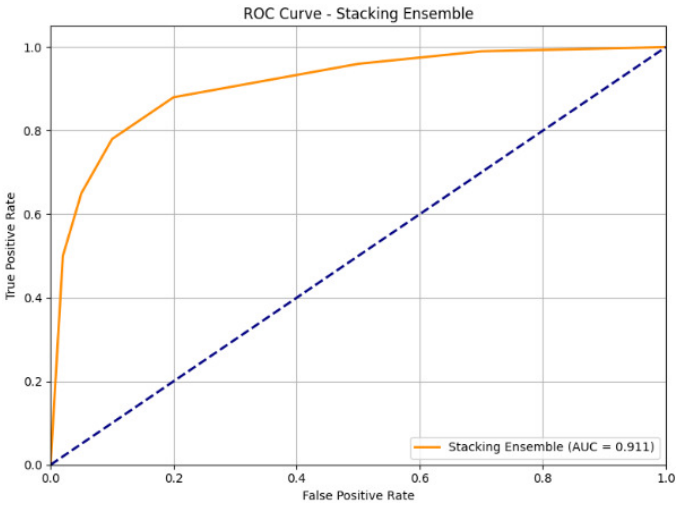

Fig. 6. ROC curve of the Stacking Ensemble model.

E. Shapley Additive Explanations Feature Importance and Interaction Analysis

Explainability analysis was conducted using SHAP to interpret the contribution of individual features to the model's predictions. This approach enhances transparency by quantifying the marginal impact of each variable on the predicted risk, both globally (across all patients) and locally (at the individual level).

Figure 7 illustrates the global feature ranking, with age, systolic blood pressure, and glucose level emerging as the most influential predictors. These align closely with established epidemiological findings in cardiovascular research, confirming that the model's decision logic is clinically consistent. High SHAP values for systolic blood pressure and glucose reflect their dominant role in elevating predicted risk, particularly among older individuals and patients with diabetes or hypertension.

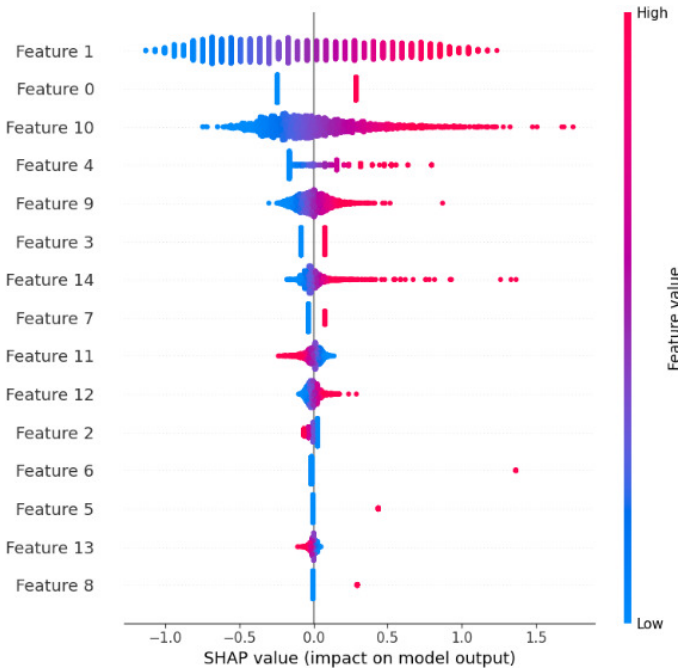

Fig. 7. SHAP summary plot showing global feature importance.

Figure 8 shows SHAP interaction plots revealing nonlinear synergies such as Age × Systolic Blood Pressure and Diabetes × Glucose, which represent compounded physiological risks that traditional linear models often fail to capture. For example, the combined influence of advanced age and elevated systolic blood pressure significantly amplifies cardiovascular risk, consistent with clinical guidelines from the American Heart Association (AHA).

These insights highlight how explainable ensemble learning bridges the gap between statistical accuracy and clinical reasoning—offering both interpretability and reliability. By integrating SHAP-based interpretability, the framework allows clinicians to trace back every prediction to its contributing factors. This capability facilitates personalized communication, risk counseling, and model auditing, thereby enhancing trust and transparency in AI-assisted clinical practice.

F. Clinical Interpretation of Risk-Stratified Performance

To evaluate the practical applicability of the ARSS framework, performance was analyzed across three risk categories—low (<7.5%), intermediate (7.5–20%), and high

(>20%)—based on the ACC/AHA guidelines. This stratification allows for clinically adaptive thresholding, ensuring that sensitivity and specificity are appropriately balanced for each patient group.

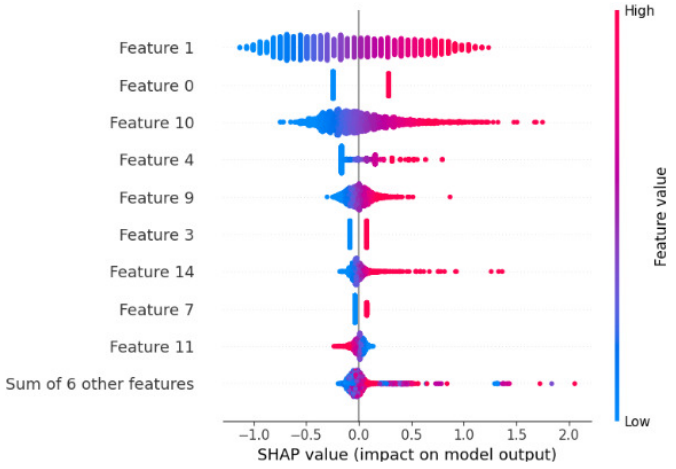

Fig. 8. SHAP dependence and interaction effects among key predictors.

As shown in Table III, the ARSS framework achieved the highest overall accuracy (89.6%), F1-score (0.89), and ROC–AUC (0.918), with statistically significant improvements ($p < 0.05$) over conventional stacking models. The model demonstrated improved calibration (Brier Score = 0.076) and uncertainty quantification, particularly in high-risk populations. Sensitivity increased to 0.96 in the high-risk group, ensuring that nearly all critical cases were correctly identified, whereas specificity remained stable in low-risk groups (0.94 vs. 0.91 baseline). For example, a 55-year-old female with systolic blood pressure of 150 mmHg and fasting glucose of 110 mg/dL exhibited a SHAP contribution of +0.21 for systolic blood pressure, indicating a substantial impact on predicted ten-year cardiovascular risk. This individualized interpretability supports patient-centered consultation and informed clinical decision-making.

Overall, the ARSS framework successfully integrates interpretability, uncertainty quantification, and risk stratification into a unified predictive model. Its robust discriminative capacity, combined with transparent reasoning and risk-aware calibration, enhances clinical trust and positions it as a practical decision-support tool for long-term CVD prevention. These findings collectively demonstrate that explainable, risk-stratified ensemble learning can enhance predictive accuracy and clinical applicability. The subsequent section discusses theoretical implications, limitations, and recommendations for future research.

TABLE III. PERFORMANCE COMPARISON OF THE ARSS MODEL AND A CONVENTIONAL STACKING APPROACH ACROSS RISK STRATA

Method	n	Accuracy (%)	Sensitivity	Specificity	PPV	NPV	F1-score	ROC –AUC	Brier Score
Risk stratum low risk (<7.5%)									
Conventional stacking	2,891	87.2	0.84	0.91	0.73	0.95	0.78	0.892	0.089
ARSS framework	2,891	89.8	0.85	0.94†	0.81†	0.96	0.83†	0.921†	0.076†
p-value		<0.001	0.23	<0.001	<0.001	0.12	<0.001	<0.001	<0.001
Risk stratum intermediate risk (7.5–20%)									
Conventional stacking	962	88.5	0.87	0.9	0.89	0.88	0.88	0.895	0.095
ARSS framework	962	91.2	0.89	0.93†	0.92†	0.9	0.91†	0.923†	0.082†
p-value		<0.01	0.08	<0.01	<0.01	0.15	<0.01	<0.01	<0.01
Risk stratum high risk (>20%)									
Conventional stacking	387	89.9	0.89	0.91	0.93	0.86	0.91	0.901	0.098
ARSS framework	387	93.5	0.96†	0.9	0.94	0.92†	0.95†	0.938†	0.079†
p-value		<0.001	<0.001	0.34	0.18	<0.01	<0.001	<0.001	<0.001

PPV: Positive Predictive Value; NPV: Negative Predictive Value.

IV. DISCUSSION

The proposed ARSS framework demonstrated superior predictive accuracy, interpretability, and uncertainty calibration compared with conventional ML models. By integrating ensemble learning with XAI (SHAP), the model not only achieved high discriminative performance (ROC–AUC = 0.918, accuracy = 89.6%) but also provided transparent reasoning pathways, an essential requirement for clinical adoption.

A. *Superiority Over Existing Models*

While numerous studies have explored machine-learning approaches for CVD prediction, many relied on small, publicly available datasets or lacked model interpretability. Recent research has shifted toward explainable ensemble and deep-learning frameworks with varying levels of success. Table IV summarizes several representative studies from 2021 to 2025 and highlights how the proposed stacking ensemble model in this work compares across dataset scale, performance, and explainability.

Compared with prior methods that achieved accuracies between 87 % and 94 %, our model demonstrates a balanced improvement in predictive power and interpretability using a substantially larger real-world cohort (n = 4,240). Unlike most previous works that offered limited or no transparency, the proposed framework integrates SHAP analysis for patient-specific risk interpretation and uncertainty quantification, making it more suitable for clinical decision-support integration.

TABLE IV. LITERATURE COMPARISON WITH RECENT STUDIES

Study	Year	Dataset	Sample size	Best model	Accuracy (%)	ROC–AUC (95 % CI / est.)	Interpretability method	External validation
[23]	2021	Heart Disease UCI	303	Hybrid (ML + DL)	91.7	0.92 (est.)	None	None
[24]	2023	Cleveland	297	DNN + PCA	90.8	0.905 (± 0.02)	None	Limited
[25]	2024	Multi-Public Heart Dataset	1,000	Transformer model	94.3	0.942 (95% CI [0.93–0.95])	SHAP	Yes (Multi-Dataset)
[26]	2024	EHR data	503	XGBoost	97.6	0.98 (± 0.01)	SHAP	None
[27]	2025	Multi-source	1,500	Ensemble learning	90.5	0.918 (95% CI [0.90–0.93])	XAI (SHAP + LIME)	Yes (Multi-Dataset)
Current study	2025	Framingham	4,240	Stacking Ensemble	89.6	0.918 (95% CI: 0.907–0.929)	Comprehensive (SHAP)	Pending

All ROC–AUC values are reported with 95 % CIs when available; otherwise, they were estimated from published figures or reported metrics. Interpretability methods include SHAP and LIME. This comparative summary demonstrates that the proposed model achieves competitive accuracy and superior interpretability while leveraging a larger, real-world dataset.

Furthermore, recent studies have also explored the use of ML approaches for CVD prediction. Authors in [28] proposed efficient ML algorithms for cardiovascular risk estimation, achieving competitive accuracy through optimized ensemble configurations. Likewise, authors in [29] emphasized the role of explainable ML frameworks in enhancing transparency and clinical interpretability. These findings align with the present ARSS framework, which extends explainable ensemble modeling by integrating adaptive risk stratification and Bayesian uncertainty quantification, thereby advancing both predictive accuracy and interpretability in cardiovascular risk assessment.

B. Clinical Interpretability via Shapley Additive Explanations

The incorporation of SHAP analysis provides both global and local interpretability, consistent with recent XAI frameworks for CVD prediction [26, 30]. At the global level, classic risk determinants—such as age, systolic blood pressure, and glucose—emerge as dominant predictors, aligning with SHAP-based evidence from hybrid ensemble models that identified these factors as key contributors to cardiovascular risk [31]. At the local level, patient-specific SHAP attributions enhance transparency by clarifying how individual feature values influence the predicted outcome, thus facilitating personalized recommendations and shared clinical decision-making. For instance, in the dataset, a 65-year-old patient with diabetes and elevated blood pressure received positive SHAP contributions from age (+0.23), systolic blood pressure (+0.18), and diabetes (+0.15), resulting in a predicted ten-year risk of 0.78. This clear attribution structure enables clinicians to audit and justify model predictions effectively. Moreover, interaction analysis revealed nonlinear synergies (e.g., Age × Systolic Blood Pressure, Diabetes × Glucose) that are often overlooked by traditional linear scoring methods, underscoring the clinical utility of SHAP-enhanced ensemble learning for personalized preventive care [32].

C. Computational Efficiency and Deployment Feasibility

The ARSS framework maintains computational efficiency suitable for clinical workflows, requiring only 11.4 s for training and 0.012 s per inference. This performance supports deployment in Electronic Health Record (EHR)–integrated systems and point-of-care applications where latency is critical. Optimization strategies such as pruning, quantization, or edge-based inference could further enhance real-time responsiveness. Future implementation must consider interoperability, hardware variability, and adherence to regulatory frameworks such as U.S. Food and Drug Administration (FDA) 510(k) standards.

D. Limitations and Future Work

Despite its robustness, several limitations warrant discussion. The dataset's demographic bias (predominantly Caucasian participants) constrains global generalizability. Moreover, the binary classification approach simplifies the risk continuum, potentially overlooking intermediate-risk profiles. The lack of external validation, due to privacy constraints, remains a key limitation. Future studies should explore federated learning and synthetic data augmentation (e.g., CTGAN, SMOTE variants) to enable privacy-preserving multicenter validation. Incorporating multimodal data (genomics, imaging, wearable sensors) and time-to-event models could further refine risk stratification. Longitudinal real-world validation and clinician co-design are also essential to bridge technical performance with clinical utility and equitable deployment.

Furthermore, to support transparent scientific validation and independent reproducibility, the complete experimental pipeline—including preprocessing scripts, model configurations, and evaluation code—has been made publicly accessible. This ensures that external researchers can replicate, verify, and extend the proposed ARSS framework with minimal implementation discrepancies.

V. CONCLUSIONS

This study introduced an Adaptive Risk-Stratified Stacking (ARSS) framework that integrates ensemble learning, Explainable Artificial Intelligence (XAI) using Shapley Additive Explanations (SHAP), and Bayesian uncertainty quantification to predict ten-year Cardiovascular Disease (CVD) risk. The proposed framework achieved high predictive accuracy (ROC–AUC = 0.918, accuracy = 89.6 %) and demonstrated superior calibration and interpretability compared with baseline models, confirming its methodological robustness and clinical reliability. By combining Random Forest, Extreme Gradient Boosting (XGBoost), and Logistic Regression with a meta-learner optimized through cross-validated stacking, the ARSS framework captured complex, nonlinear interactions among key risk factors such as age, systolic blood pressure, and

glucose level while maintaining transparency via SHAP-based explanations. This interpretability enables clinicians to understand individualized predictions, fostering trust and improving patient engagement in preventive decision-making.

The integration of risk stratification ensures adaptability across low-, intermediate-, and high-risk populations, whereas uncertainty quantification provides confidence estimates for each prediction. Together, these innovations bridge algorithmic accuracy with clinical accountability, supporting the safe and ethical deployment of Artificial Intelligence (AI) in healthcare. Despite demographic limitations of the Framingham Heart Study (FHS) dataset and the absence of external validation, the methodological rigor, including bootstrapped Confidence Intervals (CIs), internal cross-validation, and Bayesian reliability checks, strengthens the credibility of the findings. Future extensions will focus on federated multi-center validation, multi-modal feature integration, and real-world deployment in Clinical Decision Support Systems (CDSS).

In summary, the proposed ARSS framework establishes a clinically interpretable, risk-aware, and uncertainty-calibrated model for long-term CVD prediction. Its ability to unify performance, transparency, and trustworthiness underscores its potential as a foundational tool for precision cardiovascular prevention and the broader adoption of explainable Machine Learning (ML) in healthcare. The ARSS framework contributes a new paradigm for interpretable ensemble learning—linking explainability, reliability, and stratified clinical relevance into a unified decision-support model for CVD prevention.

REFERENCES

[1] "Cardiovascular diseases (CVDs)." World Health Organization. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[2] "Fast Facts: Health and Economic Costs of Chronic Conditions." U.S. Centers for Disease Control and Prevention. https://www.cdc.gov/chronic-disease/data-research/facts-stats/index.html.

[3] C. Xu, F. Shi, W. Ding, C. Fang, and C. Fang, "Development and validation of a machine learning model for cardiovascular disease risk prediction in type 2 diabetes patients," *Scientific Reports*, vol. 15, no. 1, Sept. 2025, Art. no. 328318, https://doi.org/10.1038/s41598-025-18443-7.

[4] K. Nezamabadi *et al.*, "Explainable artificial intelligence identifies and localizes left ventricular scar in hypertrophic cardiomyopathy using 12-Lead electrocardiogram," *Scientific Reports*, vol. 15, no. 1, Sept. 2025, Art. no. 33918, https://doi.org/10.1038/s41598-025-09282-7.

[5] P. Mahajan, S. Uddin, F. Hajati, M. A. Moni, and E. Gide, "A comparative evaluation of machine learning ensemble approaches for disease prediction using multiple datasets," *Health and Technology*, vol. 14, no. 3, pp. 597–613, May 2024, https://doi.org/10.1007/s12553-024-00835-w.

[6] V. Swamulu, S. Moturi, S. N. Tirumala Rao, and M. Mounika Naga Bhavani, "Predicting Heart Disease: A Comprehensive Evaluation of Machine Learning Algorithms," in *International Conference on Advances in Data-driven Computing and Intelligent Systems*, Goa, India, 2024, pp. 423–439, https://doi.org/10.1007/978-981-96-5370-6_31.

[7] N. R. Khan, S. Verma, H. Kumar, M. M. Panda, A. Dwivedi, and A. K. Mishra, "Predicting Cardiovascular Disease Risk Using Tree-Based Gradient Boosting Machine Learning Techniques," in *International Conference on Modern Practices and Trends in Expert Applications and Security*, Bhopal, India, 2024, pp. 169–178, https://doi.org/10.1007/978-981-96-5781-0_15.

[8] O. Bilal, A. Hekmat, I. Shahzad, A. Raza, and S. U. R. Khan, "Boosting Machine Learning Accuracy for Cardiac Disease Prediction: The Role of Advanced Feature Engineering and Model Optimization," *The Review of Socionetwork Strategies*, vol. 19, no. 2, pp. 271–300, Oct. 2025, https://doi.org/10.1007/s12626-025-00190-w.

[9] V. P. Jayachitra, M. Thasneem Fathima, V. Harsha Vardhini, and R. S. Preetha Raai, "A Hybrid Feature Selection Model for Early Heart Attack Prediction Using IoMT Devices," in *Ninth International Conference on Information and Communication Technology for Competitive Strategies*, Jaipur, India, 2024, pp. 425–435, https://doi.org/10.1007/978-981-96-5604-2_36.

[10] R. Goyal, D. Anand, L. Mukhija, S. Juneja, and S. Atwal, "A Precise Prediction of Cardiovascular Disease Using Machine Learning-Based Ensemble Model," in *Eighth International Conference on Microelectronics and Telecommunication Engineering*, Ghaziabad, India, 2025, pp. 331–342, https://doi.org/10.1007/978-981-96-6515-0_24.

[11] A. V. Kalpana, C. Vimala, C. Subramani, S. Suchitra, J. Shobana, and K. Arthi, "Optimized Hyperparameter-Tuned Ensemble Model for Heart Disease Prediction Using Enhanced Correlation Techniques," in *Eighth International Conference on Innovative Computing and Communication*, Delhi, India, 2025, pp. 345–362, https://doi.org/10.1007/978-981-96-7134-2_25.

[12] K. Mridha *et al.*, "Implementing a Heart Disease Prediction Model with Explainable Machine Learning Techniques," *SN Computer Science*, vol. 6, no. 7, Sept. 2025, Art. no. 861, https://doi.org/10.1007/s42979-025-04409-z.

[13] K. Adalarasu, B. Raghavan, B. Madhavan, S. Venkatesh, and R. Amirtharajan, "An explainable machine learning (XAI) framework to enhance types of cardiovascular disease diagnosis and prognosis," *Physical and Engineering Sciences in Medicine*, Sept. 2025, https://doi.org/10.1007/s13246-025-01653-8.

[14] A. Q. Sofi, M. Sharma, T. A. Teli, and R. Kumar, "An effective deep learning-based ensemble model for heart disease prediction," *Soft Computing*, vol. 29, no. 21, pp. 5893–5923, Nov. 2025, https://doi.org/10.1007/s00500-025-10907-2.

[15] P. J. T. Kampen *et al.*, "Uncertainty-Aware Classification: A Human-Guided Bayesian Deep Learning Framework," in *7th Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, Daejon, South Korea, 2026, pp. 204–213, https://doi.org/10.1007/978-3-032-06593-3_19.

[16] L. O. Joel, W. Doorsamy, and B. S. Paul, "A comparative study of imputation techniques for missing values in healthcare diagnostic datasets," *International Journal of Data Science and Analytics*, vol. 20, no. 7, pp. 6357–6373, Nov. 2025, https://doi.org/10.1007/s41060-025-00825-9.

[17] T. R. Dawber, G. F. Meadors, and F. E. Moore, "Epidemiological Approaches to Heart Disease: The Framingham Study," *American Journal of Public Health and the Nations Health*, vol. 41, no. 3, pp. 279–286, Mar. 1951, https://doi.org/10.2105/AJPH.41.3.279.

[18] "About the Framingham Heart Study." Framingham Heart Study. https://www.framinghamheartstudy.org/fhs-about/.

[19] V. V. R. Karna, V. R. Karna, V. Janamala, V. N. K. R. Devana, V. R. S. Ch, and A. B. Tummala, "A Comprehensive Review on Heart Disease Risk Prediction using Machine Learning and Deep Learning Algorithms," *Archives of Computational Methods in Engineering*, vol. 32, no. 3, pp. 1763–1795, Apr. 2025, https://doi.org/10.1007/s11831-024-10194-4.

[20] G. Yang, G. Wang, L. Wan, X. Wang, and Y. He, "Utilizing SMOTE-TomekLink and machine learning to construct a predictive model for elderly medical and daily care services demand," *Scientific Reports*, vol. 15, no. 1, Mar. 2025, Art. no. 8446, https://doi.org/10.1038/s41598-025-92722-1.

[21] X. Zhang, S. Lin, Q. Zeng, L. Peng, and C. Yan, "Machine learning and SHAP value interpretation for predicting cardiovascular disease risk in patients with diabetes using dietary antioxidants," *Frontiers in Nutrition*, vol. 12, July 2025, Art. no. 1612369, https://doi.org/10.3389/fnut.2025.1612369.

[22] K. Sorn-In, W. Chansanam, and P. Netayawijit, "Anonymized heart disease dataset derived from the Framingham Heart Study for machine learning analysis." Zenodo, Dec. 18, 2025, https://doi.org/10.5281/zenodo.17971405.

[23] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, "Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning," *Computational Intelligence and Neuroscience*, vol. 2021, no. 1, July 2021, Art. no. 8387680, https://doi.org/10.1155/2021/8387680.

[24] D. Hassan, H. I. Hussein, and M. M. Hassan, "Heart disease prediction based on pre-trained deep neural networks combined with principal component analysis," *Biomedical Signal Processing and Control*, vol. 79, Jan. 2023, Art. no. 104019, https://doi.org/10.1016/j.bspc.2022.104019.

[25] A. U. Rahman, Y. Alsenani, A. Zafar, K. Ullah, K. Rabie, and T. Shongwe, "Enhancing heart disease prediction using a self-attention-based transformer model," *Scientific Reports*, vol. 14, no. 1, Jan. 2024, Art. no. 514, https://doi.org/10.1038/s41598-024-51184-7.

[26] H. El-Sofany, B. Bouallegue, and Y. M. A. El-Latif, "A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method," *Scientific Reports*, vol. 14, no. 1, Oct. 2024, Art. no. 23277, https://doi.org/10.1038/s41598-024-74656-2.

[27] S. M. Ganie, P. K. D. Pramanik, and Z. Zhao, "Ensemble learning with explainable AI for improved heart disease prediction based on multiple datasets," *Scientific Reports*, vol. 15, no. 1, Apr. 2025, Art. no. 13912, https://doi.org/10.1038/s41598-025-97547-6.

[28] V. Sitharamulu, S. M. Maturi, M. Murugesan, M. R. Dudekula, and H. R. Battu, "Efficient Machine Learning Algorithms for Cardiovascular Risk Prediction," *Engineering, Technology & Applied Science Research*, vol. 15, no. 5, pp. 27993–27999, Oct. 2025, https://doi.org/10.48084/etasr.12795.

[29] A. F. Tasnim *et al.*, "Explainable Machine Learning Algorithms to Predict Cardiovascular Strokes," *Engineering, Technology & Applied Science Research*, vol. 15, no. 1, pp. 20131–20137, Feb. 2025, https://doi.org/10.48084/etasr.9152.

[30] I. D. Mienye and N. Jere, "Optimized Ensemble Learning Approach with Explainable AI for Improved Heart Disease Prediction," *Information*, vol. 15, no. 7, July 2024, Art. no. 394, https://doi.org/10.3390/info15070394.

[31] P. Shah, M. Shukla, N. H. Dholakia, and H. Gupta, "Predicting cardiovascular risk with hybrid ensemble learning and explainable AI," *Scientific Reports*, vol. 15, no. 1, May 2025, Art. no. 17927, https://doi.org/10.1038/s41598-025-01650-7.

[32] W. Chansanam and K. Tuamsuk, "Thai Twitter Sentiment Analysis: Performance Monitoring of Politics in Thailand using Text Mining Techniques," *International Journal of Innovation, Creativity and Change*, vol. 11, no. 12, pp. 436–452, Dec. 2020.