Enhancing Cardiovascular Risk Stratification with Interpretable Ensemble Learning: A Comparative Analysis and Explainable AI-Driven Insights

Panji Arisaputra[1*], Pandu Wicaksono[2], Karli Eka Setiawan[3], and Anindhita Dewabharata[4]
[1–3]Computer Science Department, School of Computer Science, Bina Nusantara University
Jakarta, Indonesia 11480
[4]Information Systems Department, School of Information Systems, Bina Nusantara University
Jakarta, Indonesia 11480
Email: [1]panji.arisaputra@binus.ac.id, [2]pandu.wicaksono005@binus.ac.id,
[3]karli.setiawan@binus.ac.id, [4]anindhita.dewabharata@binus.ac.id

***Abstract*—Cardiovascular disease (CVD) is a leading cause of global mortality, demanding accurate and interpretable risk stratification models for early intervention. However, the tradeoff between model performance and clinical interpretability remains a gap. Therefore, the research evaluates lightweight parametric ensemble models against complex black-box alternatives. A comparative analysis of Logistic Regression (LR) and Gaussian Naive Bayes (GNB) is conducted using a publicly available dataset of 1,000 patient records, enhanced with Bagging and AdaBoost ensemble methods. The data undergo standardized preprocessing to mitigate bias, and a fivefold stratified cross-validation protocol ensures model generalizability. The Bagging LR ensemble achieves the best predictive performance, with a mean accuracy of 0.966 (95% Confidence Interval (CI): [0.958, 0.974]) and a Receiver Operating Characteristic-Area Under the Curve (ROC-AUC) of 0.993 (95% CI: [0.990, 0.996]), highly competitive with state-of-the-art baselines including LightGBM and CatBoost. These results show that computationally efficient and interpretable ensembles provide a viable alternative to more complex models, particularly in data-constrained clinical settings. A comprehensive Explainable Artificial Intelligence (XAI) analysis using SHapley Additive exPlanations (SHAP) identifies clinically congruent drivers of prediction, the slope of the peak exercise ST segment, ST depression, and chest pain type. The findings highlight the potential of interpretable ensemble learning to develop dependable, deployable clinical decision support tools for integration into hospital or e-health systems for early CVD risk stratification. Further validation on larger, more diverse datasets is recommended to confirm broader clinical applicability.**

***Index Terms*—Machine Learning, Ensemble Learning, Cardiovascular Disease Prediction, AdaBoost, Bagging, Explainable AI, SHapley Additive exPlanations (SHAP)**

I. Introduction

CARDIOVASCULAR Disease (CVD) is a constellation of diseases affecting the heart and blood vessels, including coronary artery disease, hypertension, and stroke, which collectively constitute the leading cause of mortality worldwide. According to the World Health Organization (WHO), CVD claims an estimated 17.9 million lives annually, accounting for approximately 31% of all global deaths [1]. A substantial proportion of these deaths is premature and preventable through early detection and timely intervention. The reality places the improvement of early-stage diagnostics and prognostic risk assessment at the forefront of global public health priorities. In this context, the paradigm in cardiovascular care is shifting from reactive treatment of acute events to proactive, preventive strategies modified to the risk profile of an individual.

Traditional diagnostic tools are effective and often include costly, invasive, or time-consuming procedures that require specialized expertise, limiting accessibility in resource-constrained environments. Data-driven technologies, particularly Machine Learning (ML), have appeared as a transformative method, offering the potential to analyze massive and complex health datasets to uncover subtle patterns indicative of disease risk. By leveraging information readily available in Electronic Health Records (EHRs), ML models improve clinical decision-making, enabling more accurate

Received: April 14, 2025; received in revised form: Nov. 04, 2025; accepted: Nov. 04, 2025; available online: Aug. 05, 2026.
*Corresponding Author

and efficient risk stratification [2, 3]. Recent reviews show that several CVD prediction models suffer from poor data quality, limited sample sizes, and overfitting, leading to reduced clinical reliability and reproducibility [4].

The pursuit of higher predictive accuracy has driven the adoption of increasingly complex models, such as deep neural networks and large-scale transformers. However, this trend has created a critical challenge because clinical medicine requires model interpretability. The reasoning process needs to be transparent and consistent with established medical knowledge for a predictive model to gain the trust and acceptance of clinicians. Complex black-box models often fail to provide this transparency regardless of high predictive performance, forming a significant barrier to real-world implementation. Previous research has observed that models such as Random Forest (RF) and deep neural networks suffered from limited interpretability when achieving high predictive accuracy, making validation and deployment in high-stakes clinical environments more difficult [11]. Therefore, a distinct study gap exists because predictive models need to achieve an optimal balance among predictive performance, computational efficiency, and clinical interpretability [12, 13]. Simple parametric models, including Logistic Regression (LR) and Naive Bayes (NB), are highly interpretable and computationally efficient. These models often show lower predictive performance than more complex architectures. As a result, this research evaluates how the predictive performance of simple and transparent models can be improved to a level comparable to state-of-the-art baseline models through the systematic application of ensemble methods. The strategy provides a practical solution that combines strong predictive performance with clinical interpretability. The proposed method is particularly suitable for data-constrained scenarios, in which complex models are more susceptible to overfitting while simpler, well-regularized models often achieve more robust and reliable performance.

The research aims to bridge the gap by providing a rigorous and comprehensive evaluation of interpretable ensemble models for CVD risk stratification through four primary contributions. First, a systematic comparative analysis is conducted to evaluate two fundamental parametric models, LR and Gaussian NB (GNB), with improved variants based on Bagging and AdaBoost ensemble methods using a robust fivefold stratified cross-validation framework. Second, the best-performing interpretable ensemble model is benchmarked against state-of-the-art non-parametric models, including RF, Gradient Boosting (GBDT), Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and Categorical Boosting (CatBoost), to evaluate the relative predictive performance. Third, a multifaceted Explainable Artificial Intelligence (XAI) analysis is applied to the best-performing model using SHapley Additive exPlanations (SHAP), Partial Dependence Plots (PDP), and Individual Conditional Expectation (ICE) plots to explain the decision-making process, identify the most influential predictive features, and analyze failure modes. Fourth, the research assesses the potential clinical utility of the model through calibration analysis and a discussion of the diagnostic tradeoff between false positive and false negative prediction, providing a pathway towards reliable clinical integration.

A. Related Works

The application of ML to CVD prediction has a rich history, with numerous studies showing the utility of various algorithms. Foundational machine learning models, including Support Vector Machines (SVM), Decision Trees, and NB, have been widely applied [14, 15]. Ensemble methods, such as RF and GBDT, have consistently presented superior predictive performance by combining multiple base learners to improve predictive accuracy and model robustness. Table I shows recent advancements in the field to provide context for this research. The comparison shows the tradeoffs among various methods, signifying that several recent studies have prioritized high predictive accuracy or advanced architectures at the expense of model simplicity and clinical interpretability.

The field has experienced a significant shift in recent years towards more complex architectures, driven by the success of deep learning (DL) in other domains. Models such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks are explored to analyze medical imaging and time series data, such as Electrocardiograms (ECGs). For instance, CNNs excel in analyzing medical signals such as ECGs by automatically extracting hierarchical features that are often missed by manual feature engineering methods [16]. Transformer-based models, originally developed for natural language processing, are increasingly adapted for tabular health data and unstructured clinical notes in EHRs, showing strong potential for capturing complex patient-specific patterns. Transformer-based patient embeddings have presented strong predictive performance on EHR data, with some models achieving an average ROC-AUC of approximately 0.87 for future disease prediction across diverse patient populations [17]. Federated learning (FL) has also appeared as an advanced method to

Cite this article as: P. Arisaputra, P. Wicaksono, K. E. Setiawan, and A. Dewabharata, "Enhancing cardiovascular risk stratification with interpretable ensemble learning: A comparative analysis and Explainable AI-driven insights", CommIT Journal 20(2), 283–296, 2026.

TABLE I
COMPARATIVE ANALYSIS OF RECENT STUDIES IN CARDIOVASCULAR DISEASE (CVD) PREDICTION.

Reference	Models Used	Dataset	Key Performance Metrics	Stated Advantages	Identified Limitations/Gaps
[5]	Random Forest	University of California, Irvine (Machine Learning Repository) (UCI) Heart Disease (303 records)	Accuracy: 89.01%	High accuracy with specific cross-validation	No comparison against boosting methods; single dataset.
[6]	Hybrid Ensemble (AdaBoost)	Custom	Accuracy: 83.0%	Robust feature selection enhances performance	High complexity of the hybrid framework.
[7]	Extra Tree classifier	Combined Kaggle datasets	Accuracy: 98.15%	Emphasizing hyperparameter tuning	Performance highly dependent on tuning process.
[8]	Deep Learning (DL) vs. Logistic Regression (LR)	Korean National Health Database	Area Under the Curve (AUC): 0.979 (DL) vs. 0.780 (LR) for mortality	High performance on large-scale data	Using "black box," lacking interpretability.
[9]	Extreme Gradient Boosting (XGBoost)	Custom (Combined Mendeley & Kaggle)	Accuracy: 98.5% (Tuned XGBoost)	Meticulous fine-tuning with high accuracy	Primarily focusing on a single model's optimization.
[10]	Federated Learning Modified Artificial Bee Colony (optimization) with Support Vector Machine (MABC-SVM)	Combined CVD dataset	Accuracy: 93.8%	Preserving data privacy	Higher complexity due to federated architecture.

address critical issues of data privacy and fragmentation across institutions by enabling collaborative model training without centralizing sensitive patient data. This model preserves patient privacy while enabling the development of robust global models across decentralized datasets by training models locally on each client and sharing only model updates rather than raw patient data [10].

II. RESEARCH METHOD

A. Dataset and Cohort Characteristics

The research uses the publicly available CVD dataset, sourced from a multispecialty hospital in India and made available on Kaggle by Bhanu Prakash Doppala and Debnath Bhattacharyya of Lincoln University College under a CC0 Public Domain license [18]. The dataset comprises 1,000 anonymized patient records, each containing 14 attributes, including a binary target variable with the presence (1) or absence (0) of heart disease. During the process, the analysis uses 13 predictor variables derived from demographic, clinical, and laboratory measurements, as shown in Table II.

An initial Exploratory Data Analysis (EDA) has been conducted to understand the characteristics of the dataset. The target variable shows a relatively balanced class distribution, with 580 patients (58%) diagnosed with heart disease and 420 (42%) without heart disease, providing a favorable class balance for model training. Moreover, patient age ranges from 20 to 80 years, with a significant concentration in the 50-to-60-year age group. The cohort is predominantly male, comprising 765 males and 235 females. No missing values are identified in the dataset based on the initial info() check, eliminating the need for imputation. Additionally, the patientid column, which serves as a unique identifier without predictive value, is excluded from the analysis to prevent data leakage and potential model bias.

B. The End-to-End Experimental Framework

A comprehensive experimental framework is designed to ensure methodological rigor and reproducibility. It is shown in Fig. A1 in Appendix. The entire process from data preprocessing to model evaluation is conducted in a cross-validation framework to provide an unbiased estimate of model performance and prevent information leakage from the test folds into the training process.

A major component of the framework is the preprocessing pipeline, which is constructed using the Column-Transformer in scikit-learn [19]. The preprocessing pipeline applies StandardScaler to all numerical features to normalize feature distributions to an average of 0 and a standard deviation of 1. This step is crucial for models such as LR because model performance is sensitive to the scale of input features. Then, OneHotEncoder is applied to the categorical features (chest pain, resting electro, and slope) to convert the categorical features into a numerical format suitable for processing by the ML algorithms. In addition, the

Cite this article as: P. Arisaputra, P. Wicaksono, K. E. Setiawan, and A. Dewabharata, "Enhancing cardiovascular risk stratification with interpretable ensemble learning: A comparative analysis and Explainable AI-driven insights", CommIT Journal 20(2), 283–296, 2026.

TABLE II
DESCRIPTION OF DATASET FEATURES.

Feature Name	Description	Data Type	Units/Categories
patientid	Patient identification number	Numeric	Number
age	Age of the patient	Numeric	Years
gender	Gender of the patient	Binary	0: Female, 1: Male
chestpain	Type of chest pain experienced	Nominal	0: Typical Angina, 1: Atypical Angina, 2: Non-anginal Pain, 3: Asymptomatic
restingBP	Resting blood pressure	Numeric	94–200 (in mmHg)
serumcholesterol	Serum cholesterol level	Numeric	126–564 (in mg/dl)
fastingbloodsugar	Fasting blood sugar $>$ 120 mg/dl	Binary	0: False, 1: True
restingelectro	Resting electrocardiogram results	Nominal	0: Normal, 1: ST-T wave abnormality, 2: Probable or definite left ventricular hypertrophy
maxheartrate	Maximum heart rate achieved during exercise	Numeric	Beats per minute (71–202)
exerciseangina	Exercise-induced angina	Binary	0: No, 1: Yes
oldpeak	ST depression induced by exercise relative to rest	Numeric	0–6.2
slope	The slope of the peak exercise in ST segment	Nominal	1: upsloping, 2: flat, 3: downsloping
noofmajorvessels	Number of major vessels (0–3) colored by fluoroscopy	Numeric	0, 1, 2, 3
target	Diagnosis of heart disease	Binary	0: Absence, 1: Presence

TABLE III
HYPERPARAMETER CONFIGURATION FOR ALL EVALUATED MODELS.

Model	Key Hyperparameters
Logistic Regression (LR)	`penalty='l2', C=1.0, solver = 'lbfgs', max_iter = 100`
Gaussian Naive Bayes (GNB)	`var_smoothing = 1e-9`
Bagging with LR	`base_estimator=LogisticRegression(), n_estimators = 10, max_samples = 1.0, bootstrap = True`
Bagging with GNB	`base_estimator = GaussianNB(), n_estimators = 10, max_samples = 1.0, bootstrap = True`
AdaBoost with LR	`base_estimator = LogisticRegression(), n_estimators = 50, learning_rate = 1.0`
AdaBoost with GNB	`base_estimator = GaussianNB(), n_estimators = 50, learning_rate = 1.0`
Random Forest	`n_estimators = 100, criterion = 'gini', max_depth = None`
Gradient Boosting (GBDT)	`n_estimators = 100, learning_rate = 0.1, max_depth = 3`
Extreme Gradient Boosting (XGBoost)	Default parameters as per xgboost library
Light Gradient Boosting Machine (LightGBM)	Default parameters as per lightgbm library
Categorical Boosting (CatBoost)	Default parameters as per catboost library

entire preprocessing pipeline is fitted only to the training data in each fold of the cross-validation procedure to avoid data leakage. The step internally renames the features by adding prefixes such as `num__`. For clarity, all results and interpretations use the original feature names presented in Table II.

All experiments use a fixed random seed (`random_state = 123`) for the StratifiedKFold cross-validation so that the fold partitions, model initialization, and reported results are fully reproducible and can be replicated independently. All models are implemented in Python 3.10 with scikit-learn 1.5.1, while data handling and numerical computation use pandas 2.2.2 and numpy 1.26.4. Then, model interpretation is performed with shap 0.45.0, and all figures are generated using matplotlib 3.8.4 and seaborn 0.13.2. Fixing both the random seed and the specific library versions ensures that the preprocessing pipeline, the cross-validation splits, and the resulting performance metrics are deterministic and verifiable

C. Predictive Modeling Suite

The research evaluates a suite of models to provide a comprehensive comparison between interpretable parametric models and complex non-parametric baselines. All models are implemented using the default hyperparameters without specific tuning to provide a fair baseline comparison of the out-of-the-box performance. All the details of the model configurations are shown in Table III.

The parametric and ensemble models include LR, a linear model that estimates the probability of a binary outcome and serves as a highly interpretable baseline because of direct feature coefficients [20]. GNB, a probabilistic classifier based on the theorem of Bayes that assumes conditional independence among features, is also included in the analysis. Moreover, the Gaussian variant is selected because it is suitable for continuous features assumed to follow a normal distribution [21]. Bootstrap Aggregating (Bagging) is implemented using Bagging classifier with LR and GNB as the base estimators. This method reduces variance by training multiple base models on different

random subsets of the training data and averaging the prediction. AdaBoost is implemented using AdaBoost classifier with LR and GNB as the base estimators. This method builds models sequentially, with each subsequent model focusing on instances misclassified by previous models to reduce bias [19, 22].

The parametric ensemble models are benchmarked against a suite of state-of-the-art tree-based models to establish a performance baseline. The benchmark models include RF, an ensemble of decision trees that use Bagging, and GBDT, a collective method building decision trees sequentially to correct prediction errors. The comparison also includes XGBoost (a highly optimized implementation of GBDT) LightGBM (a high-performance GBDT framework recognized for speed and efficiency), and CatBoost (a GBDT algorithm designed to handle categorical features effectively).

D. Validation and Evaluation Protocol

A single train-test split, as used in the initial version of the research, is susceptible to sampling bias and does not provide a reliable estimate of model performance on unseen data. A fivefold stratified cross-validation (StratifiedKFold) strategy is used to overcome the limitation [23]. The method divides the dataset into five folds while maintaining the proportion of positive and negative classes in each fold. The model is trained on four folds and evaluated on the remaining fold. In addition, the procedure is repeated five times, with each fold serving as the testing set once. The final performance metrics are averaged across all five folds, providing a more robust and stable estimate of model performance. Hyperparameter tuning is not performed in a nested cross-validation loop, and the models are evaluated using the default parameters. This method provides a baseline comparison but leads to slightly optimistic performance estimates. The limitation can be addressed in future work through the implementation of a nested cross-validation protocol.

A comprehensive set of evaluation metrics is calculated for each model across the cross-validation folds to provide a holistic evaluation of model performance. The evaluation metrics include accuracy, which measures the proportion of correctly classified instances (Eq. (1)). Precision measures the proportion of positive prediction that are actually correct, reflecting the ability of the model to avoid classifying a negative sample as positive by minimizing false positives (Eq. (2)). Recall (sensitivity) is the ability of the model to find all the positive samples (Eq. (3)). All results are reported as the average F1-Score, the harmonic mean of precision and recall (Eq. (4)) [24]. The equations have TP as True Positive, TN as True Negative, FP as False Positive, and FN as False Negative. Here are the equations used:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{3}$$

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{4}$$

Next, ROC-AUC measures the ability of a model to distinguish between classes across all classification thresholds. Precision-Recall Area Under the Curve (PR-AUC) is particularly informative for imbalanced datasets. Then, balanced accuracy represents the average recall across all classes and reduces inflated performance estimates for imbalanced datasets, as shown in (5) [24]:

$$\text{Balanced Accuracy} = \frac{\text{Recall} + \text{Specificity}}{2}. \tag{5}$$

A total of 95% Confidence Interval (CI) are calculated for each evaluation metric to quantify the uncertainty and stability of the performance estimates [13]. The confidence intervals are calculated by computing the standard error of the average across the five folds and using the t-distribution to define the interval, providing the statistical rigor required for high-quality scientific reporting. The research recognizes that the variance estimates from the fivefold cross-validation procedure are unstable, and evaluation metrics, such as AUC, do not follow a normal distribution. Future studies can use more robust resampling methods, such as stratified bootstrapping, to generate more stable confidence intervals.

E. Framework for Model Interpretability and Clinical Utility

The research places a strong priority on model interpretability and clinical utility more than predictive accuracy. A suite of Explainable Artificial Intelligence (XAI) methods is used to deconstruct the decision-making process of the best-performing model. These methods include SHAP, a game-theoretic model used to explain the output of ML models. SHAP values quantify the contribution of each feature to an individual prediction. The analysis uses several SHAP visualizations, including summary plots for global feature importance, force and waterfall plots for local prediction explanations, dependence plots for feature interactions, and heatmaps for cohort analysis. PDP and ICE plots are also used. PDP model shows the marginal effect of a feature on the predicted outcome.

Cite this article as: P. Arisaputra, P. Wicaksono, K. E. Setiawan, and A. Dewabharata, "Enhancing cardiovascular risk stratification with interpretable ensemble learning: A comparative analysis and Explainable AI-driven insights", CommIT Journal 20(2), 283–296, 2026.

TABLE IV
CROSS-VALIDATION PERFORMANCE OF PARAMETRIC AND ENSEMBLE MODELS (MEAN AND 95% CONFIDENCE INTERVALS (CIS)).

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	PR-AUC	Balanced Acc.
LR	0.964 [0.954, 0.974]	0.968 [0.954, 0.981]	0.971 [0.965, 0.976]	0.969 [0.965, 0.974]	**0.993 [0.990, 0.996]**	**0.994 [0.991, 0.997]**	0.963 [0.950, 0.976]
GNB	0.948 [0.940, 0.956]	0.946 [0.937, 0.955]	0.966 [0.953, 0.978]	0.956 [0.948, 0.963]	0.988 [0.981, 0.995]	0.991 [0.983, 0.998]	0.945 [0.936, 0.954]
Bagging with LR	**0.966 [0.958, 0.974]**	**0.969 [0.953, 0.986]**	0.972 [0.968, 0.977]	**0.971 [0.964, 0.977]**	**0.993 [0.990, 0.996]**	**0.994 [0.991, 0.998]**	**0.965 [0.954, 0.976]**
Bagging with GNB	0.946 [0.941, 0.951]	0.944 [0.934, 0.955]	0.964 [0.950, 0.978]	0.954 [0.950, 0.958]	0.988 [0.982, 0.995]	0.991 [0.984, 0.998]	0.943 [0.937, 0.948]
AdaBoost with LR	0.961 [0.947, 0.975]	0.963 [0.940, 0.985]	0.971 [0.960, 0.982]	0.967 [0.956, 0.978]	0.988 [0.982, 0.995]	0.990 [0.985, 0.994]	0.959 [0.943, 0.975]
AdaBoost with GNB	0.959 [0.943, 0.975]	0.956 [0.945, 0.967]	**0.974 [0.944, 1.000]**	0.965 [0.949, 0.981]	0.991 [0.985, 0.997]	**0.994 [0.989, 0.998]**	0.956 [0.939, 0.973]

Note: Logistic Regression (LR), Gaussian Naive Bayes (GNB), Receiver Operating Characteristic-Area Under the Curve (ROC-AUC), and Precision-Recall Area Under the Curve (PR-AUC).

Lastly, calibration analysis is conducted to assess the reliability of the predicted probabilities. Standardization curves are plotted because a well-calibrated model produces predicted probabilities matching the true frequency of outcomes, which is crucial for clinical decision-making.

III. RESULTS AND DISCUSSION

This section presents the quantitative results of the comparative analysis, focusing on the performance of the parametric ensemble models and their benchmarking against state-of-the-art tree-based algorithms. The evaluation is organized in two stages: first, the six parametric and ensemble models are compared across all metrics; second, the best-performing interpretable model is benchmarked against the tree-based baselines. All results are reported as the mean and 95% CI across fivefold stratified cross-validation, so that both central performance and stability are captured. The reliability of the predicted probabilities is then assessed through a calibration analysis.

A. Comparative Analysis of Parametric Ensemble Models

The performance of the primary models (LR, GNB, and Bagging and AdaBoost variants) is shown in Table IV. The baseline LR model shows strong initial performance, while the GNB model is significantly weaker across most metrics. The high-performance metrics observed (e.g., ROC-AUC > 0.99) are partially attributable to the clean, well-structured nature of the dataset, as the risk of some overfitting without nested CV is considered.

The application of ensemble methods yields mixed results. Concerning the GNB model, Bagging leads to a slight decrease in mean accuracy from 0.948 to 0.946. As a result, ensemble methods improves the performance of the LR model. Bagging with LR achieves the highest mean accuracy (0.966), precision (0.969), F1-score (0.971), and balanced accuracy (0.965), enabling the model to have the best performance in the category. The narrow 95% CIs across all evaluation metrics show stable and reliable performance across the five cross-validation folds. The results signify that Bagging effectively reduces model variance and improves the robustness of a simple and interpretable base learner.

B. Benchmarking Against Advanced Baselines

The Bagging with LR model is benchmarked against five state-of-the-art tree-based ensemble models to provide a comprehensive performance comparison. The comparative results obtained during the results are shown in Table V. The benchmarking shows a major research finding, indicating that the lightweight and interpretable Bagging with LR model achieves predictive performance comparable to more complex non-parametric models, with no statistically significant differences. For instance, the mean accuracy of Bagging with LR (0.966) is very close to RF (0.977), GBDT (0.977), and CatBoost (0.978). The overlapping 95% CIs across most metrics signify that observed differences in mean performance are not statistically significant. This result challenges the conventional assumption that more complex models are often necessary for achieving top-tier performance. The analysis establishes that a well-tuned ensemble of a simple parametric model can provide a powerful and transparent alternative, offering a compelling trade-off between predictive accuracy and clinical interpretability.

C. Assessment of Model Calibration

The clinical utility of a model depends on classification accuracy and the reliability of the predicted probabilities. Calibration analysis is performed to assess the extent to which the predicted probabilities from each

Cite this article as: P. Arisaputra, P. Wicaksono, K. E. Setiawan, and A. Dewabharata, "Enhancing cardiovascular risk stratification with interpretable ensemble learning: A comparative analysis and Explainable AI-driven insights", CommIT Journal 20(2), 283–296, 2026.

TABLE V
BENCHMARKING PERFORMANCE AGAINST STATE-OF-THE-ART MODELS (MEAN AND 95% CONFIDENCE INTERVALS (CIS))

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	PR-AUC	Balanced Acc.
Bagging with LR	0.966 [0.958, 0.974]	0.969 [0.953, 0.986]	0.972 [0.968, 0.977]	0.971 [0.964, 0.977]	0.993 [0.990, 0.996]	0.994 [0.991, 0.998]	0.965 [0.954, 0.976]
Random Forest	0.977 [0.957, 0.997]	0.980 [0.957, 1.000]	0.981 [0.963, 0.999]	0.980 [0.963, 0.998]	0.996 [0.993, 1.000]	0.997 [0.994, 1.000]	0.976 [0.955, 0.998]
GBDT	0.977 [0.960, 0.994]	0.976 [0.956, 0.996]	0.985 [0.973, 0.996]	0.980 [0.965, 0.995]	0.997 [0.994, 1.000]	0.998 [0.995, 1.000]	0.976 [0.957, 0.994]
XGBoost	0.971 [0.949, 0.993]	0.974 [0.950, 0.999]	0.976 [0.957, 0.995]	0.975 [0.956, 0.994]	0.996 [0.992, 1.000]	0.997 [0.994, 1.000]	0.970 [0.947, 0.994]
LightGBM	0.980 [0.962, 0.998]	0.978 [0.953, 1.000]	0.988 [0.978, 0.998]	0.983 [0.968, 0.998]	0.998 [0.995, 1.000]	0.998 [0.996, 1.000]	0.979 [0.958, 0.999]
CatBoost	0.978 [0.960, 0.996]	0.980 [0.958, 1.000]	0.983 [0.966, 1.000]	0.981 [0.966, 0.996]	0.998 [0.996, 1.000]	0.999 [0.997, 1.000]	0.977 [0.958, 0.996]

Note: Gradient Boosting (GBDT), Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), Categorical Boosting (CatBoost), Logistic Regression (LR), Gaussian Naive Bayes (GNB), Receiver Operating Characteristic-Area Under the Curve (ROC-AUC), and Precision-Recall Area Under the Curve (PR-AUC).

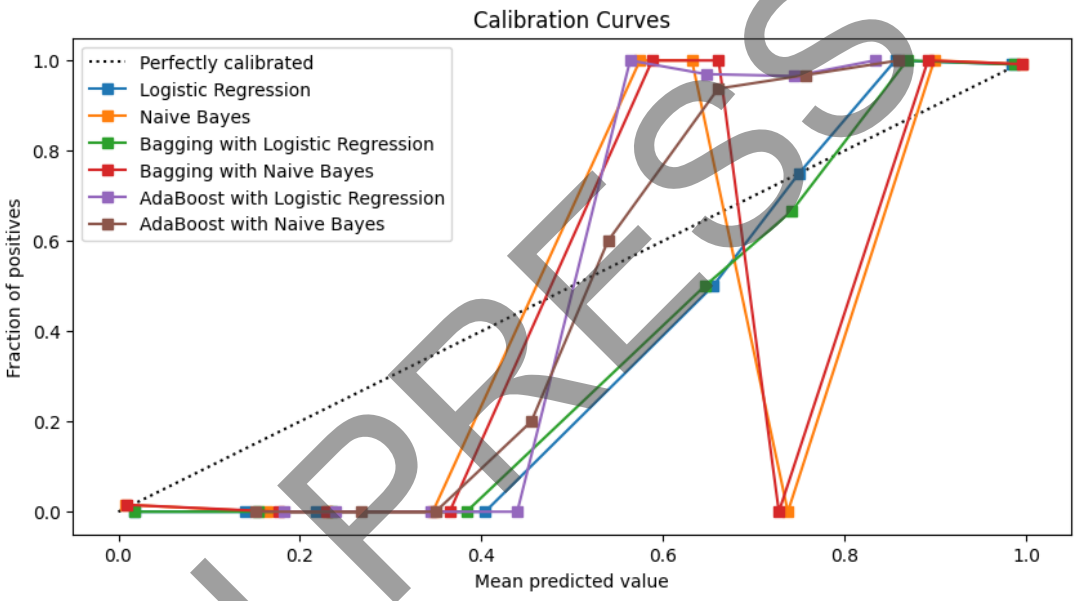

Fig. 1. Calibration curves for the evaluated models.

model are consistent with the observed frequencies of heart disease, as shown in Fig. 1. The calibration plot shows that the LR-based models, namely LR and Bagging with LR, indicate the best calibration performance, while AdaBoost-LR is an exception. The standardization curves are closest to the dashed diagonal line representing perfect calibration. The results show that a predicted probability of 70% closely corresponds to an observed heart disease prevalence of approximately 70%. Consequently, GNB-based models show poorer calibration, with standardization curves deviating more substantially from the ideal diagonal line. The superior calibration performance of the LR-based models increases confidence in clinical applications because the predicted probabilities more accurately reflect the true risk of heart disease.

D. Results of Explainable Artificial Intelligence (XAI)

The decision-making process of the best-performing model, Bagging with LR, is analyzed using a suite of XAI methods to improve the understanding of model behavior and strengthen clinical confidence. The analysis examines the model at three levels: the global level identifies the features that most influence predictions across the entire cohort; the local level explains individual predictions, including misclassified cases; and the cohort level groups patients with similar risk-response patterns. Across these levels, the analysis consistently identifies exercise-related electrocardiogram features, namely the peak-exercise ST-segment slope and ST depression (oldpeak), together with chest pain type, as the dominant drivers of the predicted CVD risk, confirming that the model relies on clinically

Cite this article as: P. Arisaputra, P. Wicaksono, K. E. Setiawan, and A. Dewabharata, "Enhancing cardiovascular risk stratification with interpretable ensemble learning: A comparative analysis and Explainable AI-driven insights", CommIT Journal 20(2), 283–296, 2026.

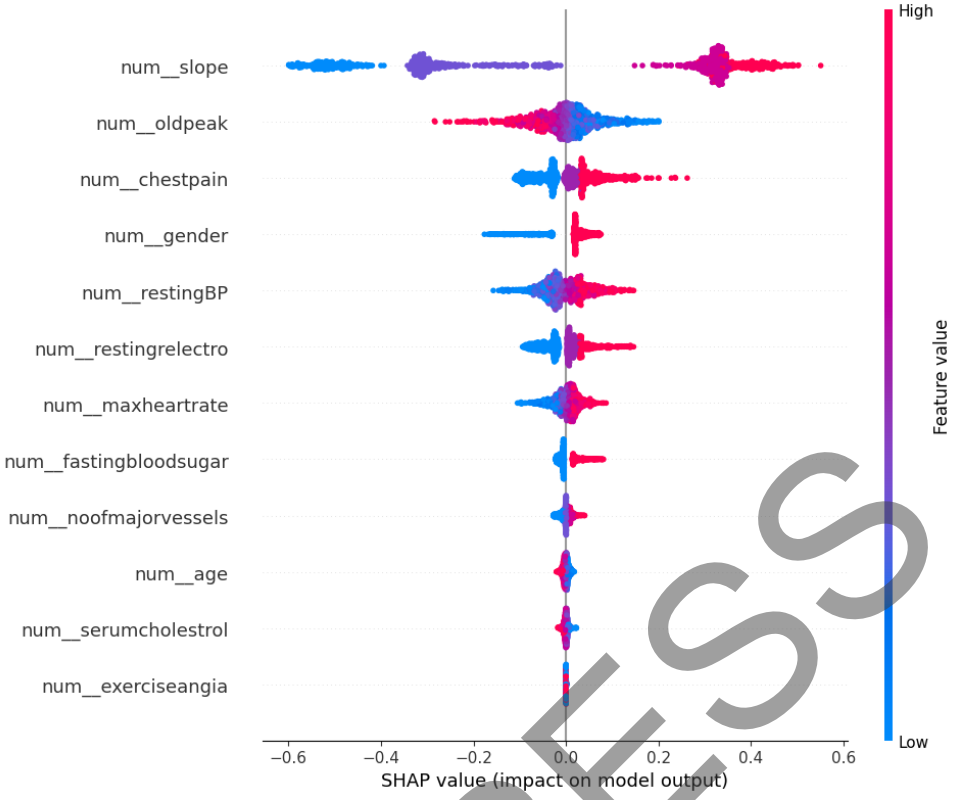

Fig. 2. SHapley Additive exPlanations (SHAP) summary plot.

meaningful factors

E. Global Drivers of Cardiovascular Risk

The SHAP review plot in Fig. 2 provides a global overview of feature importance by ranking the features according to the magnitude of the contribution to model prediction across the entire dataset. The results confirm that the model decision-making process is driven by clinically relevant features. The most influential feature is `num_slope`, representing the slope of the peak exercise ST segment, followed by `num_oldpeak`, which measures exercise-induced ST-segment depression. Both features are major indicators derived from exercise stress tests used to diagnose myocardial ischaemia. Other important features includes `num_chestpain`, `num_gender`, and `num_restingBP`. The SHAP review plot also shows the direction of the feature effects. Higher values of `num_slope` and `num_oldpeak` (red dots) are associated with positive SHAP values, indicating that the feature values increase the predicted probability of heart disease. Consequently, lower values of `num_chestpain` (blue dots, corresponding to typical angina) are connected with negative SHAP values. The pattern differs from conventional clinical expectations for chest pain and accurately reflects the relationships learned from the dataset, signifying the ability of the model to capture complex data patterns.

F. Uncovering Feature Dynamics and Interactions

PDP and ICE plots are generated to understand the response of the model to changes in major features. During the process, the PDP for slope in Fig. 3 shows that an increase in the feature value leads to a substantial increase in the average predicted probability of heart disease. It confirms the role of slope as a strong risk factor.

The combined PDP and ICE plot for oldpeak in Fig. 4 shows a more detailed understanding of the relationship between oldpeak and the predicted probability of heart disease. The average effect, represented by the red dashed PDP line, shows a positive association between the feature and the predicted probability of heart disease. Meanwhile, the individual ICE lines (blue) shows substantial heterogeneity, indicating that the effect of oldpeak is more pronounced for some patients than for others. The varying slopes of the lines signify the presence of interaction effects, where the influence of the feature is modified by other patient characteristics. Moreover, the identification of patient

Cite this article as: P. Arisaputra, P. Wicaksono, K. E. Setiawan, and A. Dewabharata, "Enhancing cardiovascular risk stratification with interpretable ensemble learning: A comparative analysis and Explainable AI-driven insights", CommIT Journal 20(2), 283–296, 2026.

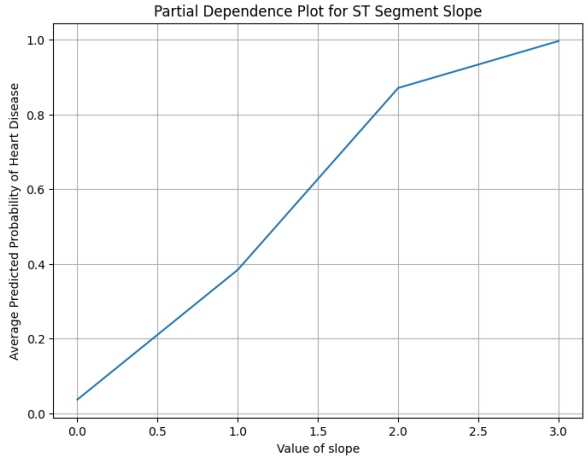

Fig. 3. Partial dependence plot for ST segment (slope).

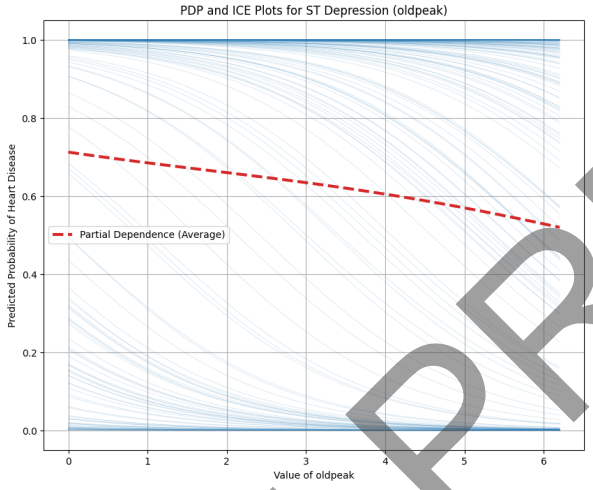

Fig. 4. Partial Dependence Plots (PDP) and Individual Conditional Expectation (ICE) plots for ST depression (oldpeak).

subgroups with different risk responses provides an important understanding that cannot be captured only through feature importance ranking.

G. Forensic Analysis of Individual Predictions and Model Errors

Understanding the circumstances under which the model produces incorrect prediction is essential for establishing clinical confidence. A forensic analysis of the misclassified cases is performed to identify the failure modes of the best-performing modes. The tradeoff between different types of classification errors is also an important consideration.

In a medical screening context, a FN, defined as failing to detect an existing case of heart disease, is generally considered more serious than a FP, which

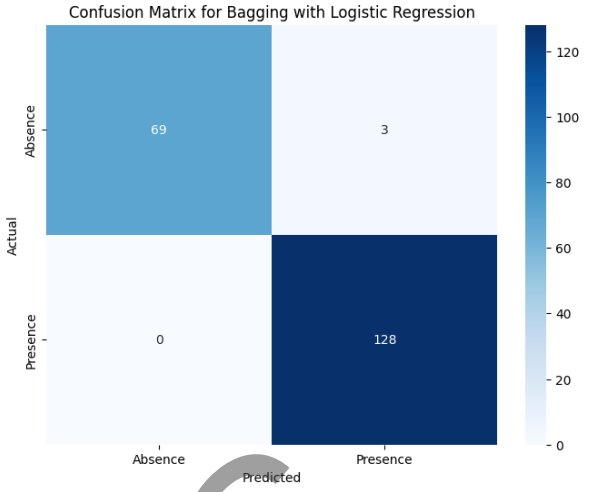

Fig. 5. Confusion matrix of the Bagging with Logistic Regression (LR) model.

incorrectly classifies a healthy individual as having heart disease. A missed diagnosis can lead to delayed treatment and poorer clinical outcomes. Following the discussion concerning the analysis, maximizing recall (sensitivity) is a primary objective. The confusion matrix for the Bagging with LR model on a representative test split, as shown in Fig. 5, indicates excellent performance, with 128 TP and 69 TN, alongside only 3 FP and no FN, respectively. The results corresponds to a recall of 100% for the positive class on the representative test split, indicating strong positioning with clinical screening priorities.

SHAP force plots provides a step-by-step explanation of individual prediction. Figure 6 shows the explanation for a misclassified instance, representing a FP in which the true class is 0 but the model predicts 1. The SHAP force plot shows that the incorrect prediction is driven by strong positive contributions from a small number of major features, outweighing the negative contributions from other features. For example, in patient index 195, high-risk contributions from chest pain and slope shift the prediction towards the positive class despite opposing evidence from the remaining variables. The results signify that the model is more sensitive to extreme values of specific features, providing an actionable understanding for future model refinement.

H. Discovering Patient Archetypes Through Cohort Analysis

Clustering patients according to the ICE plot curves for the oldpeak feature enables the identification of patient archetypes. It is defined as groups of patients

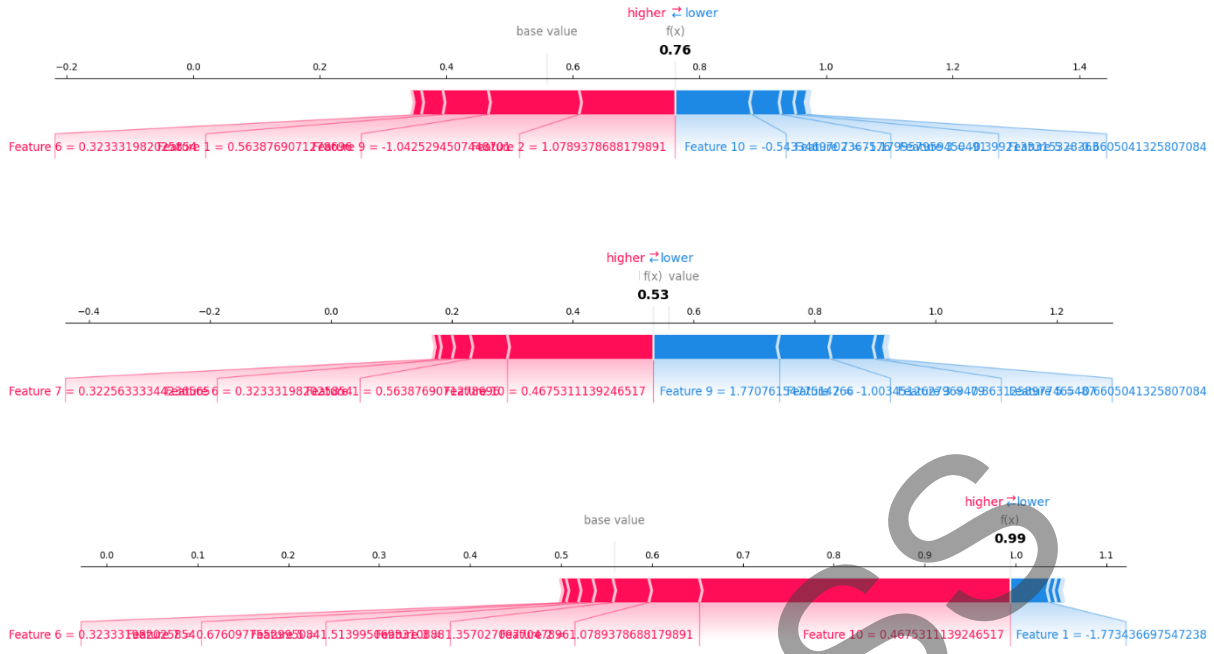

Fig. 6. SHapley Additive exPlanations (SHAP) force plot for a misclassified instance (False Positive (FP)) for patient index 195, 642, and 185.

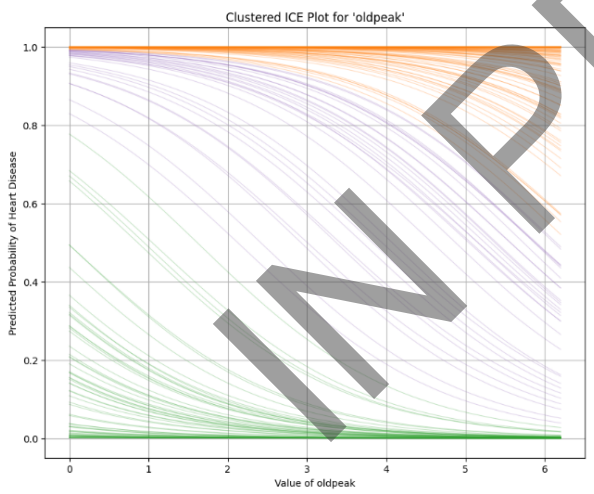

Fig. 7. Clustered Individual Conditional Expectation (ICE) plot for oldpeak.

for whom the model shows similar risk response patterns. The clustered ICE plot in Fig. 7 shows three distinct patient subgroups. The first cluster, represented in orange, shows a sharp increase in the predicted probability of heart disease as the oldpeak value increases. The second cluster in green shows a more moderate response, while the third in purple represents only a minimal response. The analysis surpasses the influence of individual features by showing the underlying prediction patterns of the model. The results show that the influence of a single clinical marker varies substantially across different patient profiles, providing a broader understanding of model behavior.

I. Discussion

The research is conducted to address the need for accurate, robust, and interpretable models for CVD risk stratification. A rigorous and comprehensive evaluation leads to several important results. First, the application of ensemble methods, particularly Bagging, significantly improves the predictive performance of simple parametric models, particularly LR. Second, the best-performing interpretable model, Bagging with LR, achieves predictive accuracy and discriminative performance comparable to those of state-of-the-art black-box models, including LightGBM and CatBoost, with no statistically significant differences. The results represent the principal contribution of the research by showing that high predictive performance and model interpretability can be achieved simultaneously. Third, the comprehensive XAI analysis confirms that model decision-making is driven by clinically relevant features and identifies the underlying prediction patterns,

interaction effects, and failure modes, providing a high level of transparency.

The results have important implications for the clinical application of ML. The development of a model that achieves both high predictive performance and high interpretability represents an important step towards establishing the confidence required for adoption by healthcare professionals. Although the Bagging with LR model shows considerable potential, the translation of the model into a clinical decision support tool should be conducted with caution.

The computational efficiency of the Bagging with LR model and the use of a limited set of routinely collected clinical variables enables the model to be a suitable candidate for deployment in resource-constrained settings or integration into real-time screening workflows. The availability of patient-specific explanations alongside the predicted risk score can support clinical decision-making, facilitate personalized patient consultations, and enable timely, targeted interventions. However, deployment in clinical practice should be preceded by extensive external validation.

IV. Conclusion

In conclusion, the research validates the significant potential of interpretable ensemble learning for CVD risk stratification. The results show that a carefully constructed Bagging ensemble based on a transparent LR model achieves predictive performance. It is competitive with complex state-of-the-art models. Therefore, the analysis supports a modelling method that prioritizes both predictive accuracy and interpretability. Moreover, the comprehensive XAI analysis increases confidence in model prediction by providing a detailed understanding of the underlying decision-making process and identifying the major factors influencing model behavior. Further external validation is required before clinical implementation, and the research establishes a robust methodological framework for developing ML models that are accurate, interpretable, reliable, as well as clinically applicable.

However, the research has several limitations that should be recognized to properly contextualize the results. The primary limitation is the use of a single and relatively small dataset comprising 1,000 patients from a single hospital in one country. This limitation reduces the generalizability of the analysis to populations with different demographic, genetic, or lifestyle characteristics. The dataset is also cross-sectional and lacks longitudinal information required to capture the dynamic evolution of cardiovascular risk factors over time. Considering a methodological perspective, the absence of a nested cross-validation protocol for hyperparameter tuning shows that the reported performance metrics are slightly optimistic. Furthermore, the calculation of confidence intervals from only five cross-validation folds lacks statistical stability.

The limitations identify several important directions in the research for future analysis. The primary and most important step is the external validation of the Bagging with LR model using larger, multicenter, and ethnically diverse clinical datasets to rigorously assess model robustness as well as generalizability. Future studies should implement a nested cross-validation framework for hyperparameter tuning to obtain a less biased estimate of model performance. The integration of additional data modalities, including longitudinal EHR data, genomic markers, and wearable devices data, may further improve predictive performance. Finally, prospective studies are needed to evaluate the practical impact of deploying the interpretable model on clinical workflows, clinical decision making, and patient outcomes.

Acknowledgement

The authors are grateful to Bina Nusantara University for providing the necessary academic resources and ongoing support that were essential for the success of the analysis. This study was supported by the INNOGEN Lab of Bina Nusantara University under the Initiative Project No. PP0060018, Year 2025.

Author Contribution

Conducted the experiments and the data analysis, P. A.; Assisted in drafting the manuscript and collecting the references, P. A.; Produced the initial draft, refined the overall structure, and conducted the revisions, P. W.; Initiated the project and developed the conceptual framework, K. E. S.; Contributed to the experiments and writing of the article, K. E. S.; Assisted in refining the experiments, A. D.; Finalized the manuscript, A. D.; and Supervised the research process and prepared the open-data documentation, A. D.

Data Availability

The data that support the results are openly available in Kaggle at https://www.kaggle.com/datasets/jocelyndumlao/cardiovascular-disease-dataset, reference number [18].

References

[1] T. Adam *et al.*, "The state of cardiac rehabilitation in Saudi Arabia: Barriers, facilitators, and policy implications," *Cureus*, vol. 15, no. 11, pp. 1–11, 2023.

[2] K. E. Setiawan, A. Kurniawan, A. Chowanda, and D. Suhartono, "Clustering models for hospitals

in Jakarta using Fuzzy C-Means and K-Means," *Procedia Computer Science*, vol. 216, pp. 356–363, 2023.

[3] P. Wicaksono, P. Samuel, I. N. Alam, and S. M. Isa, "Dealing with imbalanced sleep apnea data using DCGAN," *Traitement du Signal*, vol. 39, no. 5, pp. 1527–1536, 2022.

[4] Y. Q. Cai *et al.*, "Pitfalls in developing machine learning models for predicting cardiovascular diseases: Challenge and solutions," *Journal of Medical Internet Research*, vol. 26, pp. 1–21, 2024.

[5] S. Ouf and A. I. B. ElSeddawy, "A proposed paradigm for intelligent heart disease prediction system using data mining techniques," *Journal of Southwest Jiaotong University*, vol. 56, no. 4, pp. 220–240, 2021.

[6] S. P. Praveen *et al.*, "Enhanced feature selection and ensemble learning for cardiovascular disease prediction: Hybrid GOL2-2 T and adaptive boosted decision fusion with babysitting refinement," *Frontiers in Medicine*, vol. 11, pp. 1–18, 2024.

[7] D. Asif, M. Bibi, M. S. Arif, and A. Mukheimer, "Enhancing heart disease prediction through ensemble learning techniques with hyperparameter optimization," *Algorithms*, vol. 16, no. 6, pp. 1–17, 2023.

[8] S. J. Lee *et al.*, "Deep learning improves prediction of cardiovascular disease-related mortality and admission in patients with hypertension: Analysis of the Korean National Health Information Database," *Journal of Clinical Medicine*, vol. 11, no. 22, pp. 1–12, 2022.

[9] A. Ogunpola, F. Saeed, S. Basurra, A. M. Albarrak, and S. N. Qasem, "Machine learning-based predictive models for detection of cardiovascular diseases," *Diagnostics*, vol. 14, no. 2, pp. 1–19, 2024.

[10] M. M. Yaqoob, M. Nazir, M. A. Khan, S. Qureshi, and A. Al-Rasheed, "Hybrid classifier-based federated learning in health service providers for cardiovascular disease prediction," *Applied Sciences*, vol. 13, no. 3, pp. 1–17, 2023.

[11] X. Gao, S. Alam, P. Shi, F. Dexter, and N. Kong, "Interpretable machine learning models for hospital readmission prediction: A two-step extracted regression tree approach," *BMC medical informatics and decision making*, vol. 23, pp. 1–11, 2023.

[12] D. Y. Omkari and S. B. Shinde, "Opportunities and challenges of machine learning and deep learning techniques in cardiovascular disease prediction: A systematic review," *Journal of Biological Systems*, vol. 31, no. 02, pp. 309–344, 2023.

[13] M. A. Naser, A. A. Majeed, M. Alsabah, T. R. Al-Shaikhli, and K. M. Kaky, "A review of machine learning's role in cardiovascular disease prediction: recent advances and future challenges," *Algorithms*, vol. 17, no. 2, pp. 1–33, 2024.

[14] P. Aryawibowo, A. F. Hidayanto, Y. M. Toemali, K. E. Setiawan, and A. A. S. Gunawan, "Intelligent monitoring and diagnosing capability in healthcare: Systematic literature review," in *2023 International Conference on Information Management and Technology (ICIMTech)*. Malang, Indonesia: IEEE, Aug. 24–25, 2023, pp. 627–632.

[15] R. Katarya and S. K. Meena, "Machine learning techniques for heart disease prediction: A comparative study and analysis," *Health and Technology*, vol. 11, no. 1, pp. 87–97, 2021.

[16] A. Eleyan and E. Alboghbaish, "Electrocardiogram signals classification using deep-learning-based incorporated convolutional neural network and long short-term memory framework," *Computers*, vol. 13, no. 2, pp. 1–12, 2024.

[17] S. Xian *et al.*, "Transformer patient embedding using electronic health records enables patient stratification and progression analysis," *npj Digital Medicine*, vol. 8, no. 1, pp. 1–16, 2025.

[18] J. Dumlao, "Cardiovascular disease dataset," 2023. [Online]. Available: https://www.kaggle.com/datasets/jocelyndumlao/cardiovascular-disease-dataset

[19] G. Ngo, R. Beard, and R. Chandra, "Evolutionary bagging for ensemble learning," *Neurocomputing*, vol. 510, pp. 1–14, 2022.

[20] H. Jain, A. Khunteta, and S. Srivastava, "Churn prediction in telecommunication using logistic regression and logit boost," *Procedia Computer Science*, vol. 167, pp. 101–112, 2020.

[21] V. Vangara, S. P. Vangara, and K. Thirupathur, "Opinion mining classification using naive bayes algorithm," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 9, no. 5, pp. 495–498, 2020.

[22] S. González, S. García, J. Del Ser, L. Rokach, and F. Herrera, "A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities," *Information Fusion*, vol. 64, pp. 205–237, 2020.

[23] P. P. Pande, "Special issue on benchmarking machine learning systems and applications," *IEEE Design & Test*, vol. 39, no. 3, 2022.

[24] S. A. Hicks and Others, "On evaluation metrics for medical applications of artificial intelligence,"

Scientific Reports, vol. 12, pp. 1–9, 2022.

APPENDIX

The Appendix can be seen in the next page.

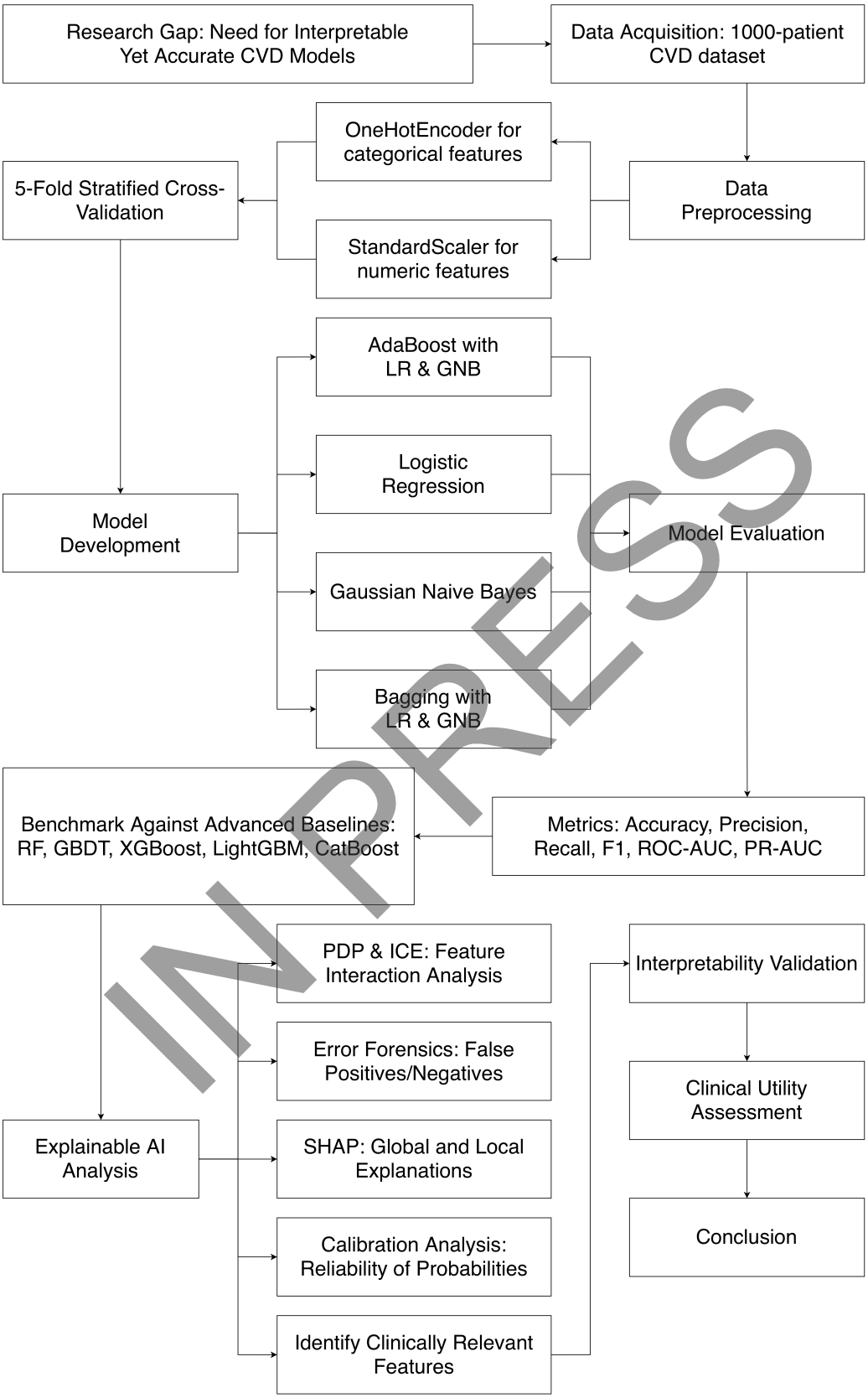

Fig. A1. Experimental framework. Note: Cardiovascular Disease (CVD), Logistic Regression (LR), Gaussian Naive Bayes (GNB), Random Forest (RF), Gradient Boosting (GBDT), Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), Categorical Boosting (CatBoost), Receiver Operating Characteristic-Area Under the Curve (ROC-AUC), Precision-Recall Area Under the Curve (PR-AUC), SHapley Additive exPlanations (SHAP), Partial Dependence Plots (PDP), and Individual Conditional Expectation (ICE).