Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15544
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Heart Disease Classification Using Optimised XGBoost and Random Forest with SHAP Explanations

Pancar Hizkia Hutagalung[1)], Andrianingsih[2)*]
[1)2)]Sistem Informasi, Fakultas Teknologi Komunikasi dan Informatika, Universitas Nasional
[1)]pancarhizkia2022@student.unas.ac.id, [2)]andrianingsih@civitas.unas.ac.id

Submitted :Nov 5, 2025 | **Accepted** : Nov 18, 2025 | **Published** : Jan 03, 2026

Abstract: Heart disease remains one of the leading causes of global morbidity, creating a need for accurate and interpretable computational tools to support early diagnosis. However, many existing studies on the Cleveland Heart Disease dataset rely on limited validation protocols, apply only a single hyperparameter optimisation strategy, or provide narrow explainability analyses, which can lead to optimistic performance estimates and inconsistent clinical insight. This study addresses these gaps by proposing a classification-based prediction framework that evaluates Random Forest and XGBoost for binary heart-disease classification under three hyperparameter optimisation strategies random search, Bayesian optimisation, and particle swarm optimisation (PSO) within a nested, anti-leakage cross-validation design, while SHAP is employed to analyse model interpretability across the best-performing configurations. The experimental results show that the ensemble classifiers achieve strong and consistent performance, with ROC–AUC values ranging from 0.8908 to 0.9089 across all scenarios; Random Forest optimised with PSO obtained the highest ROC–AUC (0.9089 ± 0.0146) and F1-score (0.8188 ± 0.0206), whereas XGBoost with Bayesian optimisation reached comparable performance without statistically significant differences. SHAP analyses identified oldpeak, ca, thal, cp, thalach, and exang as the most influential features, in line with established clinical indicators of myocardial ischemia and perfusion abnormalities. These findings indicate that combining tree-based ensemble classifiers with systematic hyperparameter optimisation and SHAP-based interpretability can enhance the reliability and transparency of heart-disease classification on the Cleveland dataset, while highlighting the need for further validation on contemporary, multi-centre clinical data.

Keywords: Heart Disease Prediction; Machine Learning; Ensemble Learning; XGBoost; Explainable Artificial Intelligence (XAI)

INTRODUCTION

Cardiovascular diseases (CVDs) remain a leading cause of death worldwide and contribute substantially to premature mortality. Early identification of individuals at high risk of coronary artery disease is therefore an important component of prevention and clinical decision making. The UCI Cleveland Heart Disease dataset, derived from the work of (Detrano et al., n.d.), is one of the most widely used benchmarks for computational heart-disease research. Because this dataset is relatively small, historical, and not fully representative of contemporary populations, careful model development and rigorous evaluation procedures are required.

Machine learning (ML) has been widely employed to support heart-disease classification. Studies using public heart-disease datasets report that ML models can achieve high accuracy when appropriate feature selection and validation strategies are used (Biswas et al., 2023). Evaluations on the Cleveland database and related datasets show that ensemble-based approaches often provide more robust discrimination than simpler models, and comparative investigations indicate that ensemble methods tend to outperform baseline classifiers for cardiovascular risk assessment (Al-Alshaikh et al., 2024; Bouqentar et al., 2024).

Among ensemble methods, Random Forest and Extreme Gradient Boosting (XGBoost) are frequently adopted because they handle heterogeneous tabular data effectively and offer strong predictive performance. Random Forest constructs an ensemble of decision trees trained on bootstrap samples with random feature subsampling,

*name of corresponding author

whereas XGBoost implements gradient-boosted trees with regularisation and advanced optimisation strategies (Breiman, 2001; Chen & Guestrin, 2016). In the heart-disease domain, a combination of Random Forest and XGBoost with particle swarm optimisation (PSO) on the Cleveland dataset has been shown to improve predictive performance compared with untuned baselines (Ansyari et al., 2023), which supports the use of tree-based ensembles as promising candidates for heart-disease classification.

The importance of explainable artificial intelligence (XAI) in medical prediction has grown alongside accuracy improvements. SHapley Additive exPlanations (SHAP) offer a game-theoretic framework for feature contribution attribution in complex models, notably in heart-disease classifiers. Research shows that integrating SHAP into these models enhances transparency and aids in the clinical interpretation of results (El-Sofany et al., 2024; Mienye & Jere, 2024; Rezk et al., 2024). A systematic review on XAI in disease prediction notes that, although SHAP and LIME are increasingly used, many studies still rely on small datasets, limited validation schemes, and predominantly qualitative interpretability analyses (Alkhanbouli et al., 2025).

Several methodological gaps remain, particularly in studies based on the Cleveland dataset. Many works evaluate only one ensemble model or one hyperparameter optimisation scheme, so the relative effect of strategies such as random search, Bayesian optimisation, and PSO on performance and robustness is still unclear (Al-Alshaikh et al., 2024; Ansyari et al., 2023). Even when ensembles are combined with XAI, evaluation is often based on non-nested cross-validation or a single train–test split, and SHAP-based explanations are usually analysed for only one model. As a result, the stability and clinical meaning of key features across different optimised models, and the practical benefit of combining advanced optimisation with SHAP in this setting, are not yet well established (Alkhanbouli et al., 2025; El-Sofany et al., 2024; Mienye & Jere, 2024; Rezk et al., 2024).

This research develops a classification-based framework for heart-disease prediction using the UCI Cleveland dataset, leveraging Random Forest and XGBoost alongside three hyperparameter optimisation strategies: random search, Bayesian optimisation, and PSO. It integrates preprocessing, class balancing with SMOTE, and classifier training into a single pipeline evaluated through nested stratified cross-validation. The study summarizes classification metrics (accuracy, recall, F1-score, and ROC-AUC) as mean ± standard deviation, with paired statistical tests to assess ROC-AUC differences. Additionally, SHAP is utilized to analyze feature attributions for the best configurations, focusing on clinically relevant predictors to jointly examine optimisation strategy, predictive performance, and interpretability in a reproducible setting.

LITERATURE REVIEW

This section reviews previous studies on heart-disease prediction and explainable artificial intelligence (XAI) in a structured way. The literature is organised into three main themes: (1) machine-learning models for heart-disease prediction, (2) optimisation strategies for ensemble classifiers, and (3) integration of XAI, especially SHAP, in clinical prediction models. Representative studies and their main contributions and research gaps are summarised in Table 1.

Table 1. Summary of previous studies on heart disease prediction and XAI

No.	Reference (Author, year, full title, DOI)	Main contribution	Main research gap that can be derived
1	Detrano, R., et al. (1989). *International application of a new probability algorithm for the diagnosis of coronary artery disease*. https://doi.org/10.1016/0002-9149(89)90524-9	Clinical probability algorithm for coronary artery disease; provides the original features and labels that form the basis of the Cleveland Heart Disease dataset.	Small, historical cohort that does not represent modern or local populations; requires external validation and larger, more diverse datasets.
2	Biswas, S., et al. (2023). *Machine learning-based model to predict heart disease in early stage*. https://doi.org/10.1155/2023/6864343	Develops ML models for early-stage heart disease prediction using public datasets and several feature-selection strategies.	Relies on public benchmark data and does not integrate XAI deeply; clinical interpretation of model behaviour remains limited.
3	Ansyari, H., et al. (2023). *Implementation of Random Forest and Extreme Gradient Boosting in heart disease classification using PSO feature*	Combines Random Forest and XGBoost with PSO-based feature selection for	Does not employ SHAP or other XAI methods to explain important features;

*name of corresponding author

	selection. https://doi.org/10.35882/jeemi.v5i4.322	heart disease classification on the Cleveland dataset.	focuses mainly on accuracy, without discussing fairness, bias, or data integrity.
4	Bouqentar, L., et al. (2024). *Early heart disease prediction using feature engineering and machine learning algorithms*. https://doi.org/10.1016/j.heliyon.2024.e38731	Combines feature engineering and multiple ML algorithms to improve early heart disease prediction on Cleveland and Statlog datasets.	Focuses on accuracy and feature engineering; XAI is limited and not strongly linked to deep clinical interpretation or robustness analysis.
5	Al-Alshaikh, H. A., et al. (2024). *Comprehensive evaluation and performance analysis of machine learning in heart disease prediction*. https://doi.org/10.1038/s41598-024-58489-7	Provides a broad evaluation of multiple ML models for heart disease prediction, reporting strong results for ensemble-based approaches.	XAI is not the main focus; the work centres on performance metrics and relies mainly on Cleveland, which limits demographic diversity and generalisability.
6	Mienye, I. D., & Jere, N. R. (2024). *Optimized ensemble learning approach with explainable AI for improved heart disease prediction*. https://doi.org/10.3390/info15070394	Proposes an optimised ensemble with XAI for heart disease prediction using Cleveland and Framingham datasets, employing SHAP to analyse feature importance.	Uses benchmark datasets; generalisation to local clinical populations and systematic statistical comparison of alternative configurations are not addressed.
7	Rezk, E., et al. (2024). *XAI-augmented voting ensemble models for heart disease prediction*. https://doi.org/10.3390/bioengineering11101016	Develops voting ensemble models augmented with SHAP and LIME to explain heart-disease predictions on a clinical dataset.	Uses a single clinical dataset; provides limited investigation of multi-centre generalisation and stability of XAI explanations across different patient subgroups.
8	El-Sofany, H. F., et al. (2024). *A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method*. https://doi.org/10.1038/s41598-024-74656-2	Combines multiple ML algorithms with XAI methods to generate more transparent heart-disease risk scores on public datasets.	Relies on public data and does not consider robust anti-leakage validation, real-time clinical deployment, or explicit data-integrity mechanisms.
9	Teja, K. R., & Rayalu, A. (2025). *Optimizing heart disease diagnosis with advanced machine learning models*. https://doi.org/10.1186/s12872-025-04627-6	Uses several heart-disease datasets (UCI and Kaggle) to train and compare advanced ML models for diagnosis.	Although multiple datasets are used, XAI and data-integrity aspects are not explored; the combined effect of optimisation and interpretability remains open.
10	Alsabhan, W., & Alfadhly, A. (2025). *Effectiveness of machine learning models in diagnosis of heart disease: A comparative study*. https://doi.org/10.1038/s41598-025-09423-y	Provides a comparative study of several ML models for heart-disease diagnosis, with emphasis on predictive performance.	Emphasises performance but does not investigate XAI in depth or robustness under distribution

*name of corresponding author

			shifts; interpretability and stability of model decisions remain underexplored.

Computational research on heart disease prediction is strongly influenced by the clinical work of (Detrano et al., n.d.), who developed a probability-based algorithm for the diagnosis of coronary artery disease. The variables and outcome from that study form the basis of the UCI Cleveland Heart Disease dataset, which has become a standard benchmark for evaluating machine-learning models in this domain.

From the perspective of predictive modelling, several works focus primarily on improving performance using conventional machine-learning models. (Biswas et al., 2023) developed ML-based models for early-stage heart-disease prediction using public datasets and multiple feature-selection strategies, reporting increased accuracy compared with traditional baselines. (Bouqentar et al., 2024) combined feature engineering with several algorithms on the Cleveland and Statlog datasets and showed that refined input representations can lead to higher accuracy and ROC–AUC. (Al-Alshaikh et al., 2024) conducted a comprehensive evaluation of multiple ML models for heart-disease prediction and found that ensemble-based models tend to outperform simpler classifiers. More recent comparative works by (Teja & Rayalu, 2025) and by (Alsabhan & Alfadhly, 2025) considered multiple datasets, including UCI and Kaggle collections, and systematically compared advanced models for heart-disease diagnosis. Across these studies, ensemble methods consistently emerge as strong baselines, but the focus remains on performance metrics rather than on detailed interpretability or rigorous anti-leakage validation.

Beyond accuracy alone, some studies explicitly explore optimisation strategies for ensemble models. (Ansyari et al., 2023) utilized Random Forest and XGBoost in heart-disease classification with the Cleveland dataset, employing particle swarm optimization (PSO) for feature selection. Results show significant improvements through metaheuristic optimization compared to untuned models. However, model interpretability was not addressed, lacking techniques like SHAP for explanation. Furthermore, the research highlights a gap in existing literature, where most optimization studies use a single scheme and fail to compare various strategies like random search, Bayesian optimization, and PSO regarding performance and robustness.

Parallel to these developments, there is growing interest in integrating XAI into heart-disease prediction models. (Mienye & Jere, 2024) proposed an optimised ensemble learning approach with XAI for heart-disease prediction using Cleveland and Framingham datasets and employed SHAP to analyse feature importance. (Rezk et al., 2024) developed XAI-augmented voting ensemble models for heart disease, combining SHAP and LIME to interpret predictions on a clinical dataset. (El-Sofany et al., 2024) combined several machine-learning algorithms with XAI methods to produce more transparent risk scores on public heart-disease datasets. These works demonstrate that XAI can be effectively integrated with ensemble models and that SHAP-based analyses can provide clinically meaningful insights into the role of key features.

At a broader level, (Alkhanbouli et al., 2025) performed a systematic literature review on the role of explainable artificial intelligence (XAI) in disease prediction, revealing that methods like SHAP and LIME are gaining traction, with transparency deemed essential for clinical decision support. However, the review pointed out a lack of rigorous evaluation protocols that differentiate hyperparameter tuning from performance estimation, as well as insufficient research on the stability of explanations across various models and optimization settings.

Table 1 provides a consolidated view of these studies, including their main contributions and the research gaps that can be derived. Ensemble methods, especially tree-based models such as Random Forest and XGBoost, provide strong predictive performance for heart-disease classification on Cleveland and related cohorts (Al-Alshaikh et al., 2024; Alsabhan & Alfadhly, 2025; Biswas et al., 2023; Bouqentar et al., 2024; Teja & Rayalu, 2025). Metaheuristic optimisation, such as PSO, has been shown to further improve performance in some settings (Ansyari et al., 2023). At the same time, XAI techniques such as SHAP and LIME have been successfully used to interpret ensemble predictions and to make model behaviour more transparent to clinicians (Mienye & Jere, 2024; Rezk et al., 2024; El-Sofany et al., 2024; Alkhanbouli et al., 2025).

Despite its strengths, the research identifies methodological gaps in heart-disease studies using the Cleveland dataset, which typically focus on a single ensemble model or hyperparameter optimization approach. This limitation prevents a comprehensive assessment of the effects of various strategies like random search, Bayesian optimisation, and PSO on predictive performance and robustness (Al-Alshaikh et al., 2024; Ansyari et al., 2023; Biswas et al., 2023). A reliance on single-strategy tuning is noted in optimization studies across various domains, including PSO-based deep learning systems for assistive technologies and gradient-boosting case studies in geospatial risk modeling (Hindarto, 2024; Islam et al., 2023). Second, evaluation protocols are often based on non-nested k-fold cross-validation or single train–test splits, with limited attention to information leakage between tuning and testing phases, which raises concerns about over-optimistic performance estimates, particularly when working with small datasets such as Cleveland (Bouqentar et al., 2024; Gnanavelu et al., 2025; Teja & Rayalu,

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15544
e-ISSN : 2541-2019
p-ISSN : 2541-044X

2025). Third, although XAI methods such as SHAP and LIME are increasingly integrated into clinical prediction models, SHAP-based explanations are typically analysed for only one selected model, and there is little investigation of whether feature rankings and clinical interpretations remain stable across different optimised configurations (Alkhanbouli et al., 2025; El-Sofany et al., 2024; Lundberg et al., n.d.; Mienye & Jere, 2024; Rezk et al., 2024). These gaps form the methodological basis for the framework proposed in this study, which compares ensemble classifiers and optimisation strategies under a nested, anti-leakage validation scheme while simultaneously examining SHAP-based explanations for the best-performing configurations.

METHOD

This study develops a heart disease classification model based on ensemble *machine learning* and *explainable AI*. The overall research process starts from data acquisition and preprocessing, continues with class balancing and model training using three hyperparameter optimisation strategies, and ends with performance evaluation, statistical testing, and SHAP-based interpretation. The methodological pipeline can be summarised visually as follows.

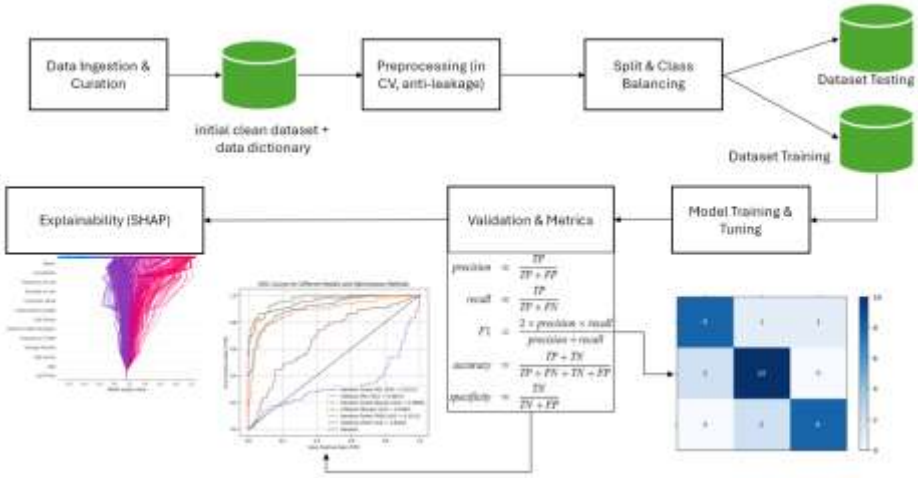

Fig 1 Proposed method Predicting Heart Disease with Machine Learning

The experiment uses the Cleveland Heart Disease dataset from the UCI Machine Learning Repository, which is widely adopted as a benchmark in the literature. The dataset contains 303 patient records with routinely collected clinical attributes and a diagnosis variable. In this study, the original multi-class diagnosis num is reformulated as a binary label, where 0 denotes “no heart disease” and 1 denotes “heart disease”. This formulation aligns the task with a binary classification problem, and the term “classification” is used consistently throughout the article. The predictor variables include demographic information (age and sex), chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiogram, maximum heart rate achieved, exercise-induced angina, ST-segment depression (*oldpeak*), slope of the peak exercise ST segment, number of major vessels coloured by fluoroscopy (*ca*), and thallium stress test result (*thal*). These variables represent clinically meaningful risk factors that are commonly used in cardiovascular assessment and provide a realistic basis for developing a classification-based prediction model.

Data preprocessing is designed to be fully embedded inside the cross-validation pipeline to prevent information leakage. Numeric features (age, resting blood pressure, serum cholesterol, maximum heart rate, and *oldpeak*) are processed with median imputation using SimpleImputer(strategy="median"). Categorical features (sex, chest pain type, fasting blood sugar, resting ECG, exercise-induced angina, ST-slope, number of vessels, and thallium test result) are imputed with the most frequent category and then transformed via one-hot encoding using OneHotEncoder(handle_unknown="ignore"). The numeric and categorical transformers are combined in a ColumnTransformer and attached to the classifier through an ImbPipeline from the imbalanced-learn library. After one-hot encoding, a SelectKBest step is applied to control the dimensionality of the transformed feature space by retaining the most informative variables according to a univariate statistical test. No feature scaling or normalisation is applied, because both *Random Forest* and *XGBoost* are tree-based ensemble classifiers that do not require standardised input features. All preprocessing components are fitted only on the training portion within each fold and then applied to the corresponding validation or test data, so that the anti-leakage principle is strictly maintained.

The original dataset exhibits a moderate imbalance between the positive and negative classes. To mitigate bias towards the majority class, the Synthetic Minority Oversampling Technique (*SMOTE*) is integrated into the training pipeline. SMOTE is placed after preprocessing and feature selection. For each training split within cross-

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

validation, SMOTE is fitted on the training subset only and used to generate synthetic minority class samples; the corresponding validation or test subset remains untouched. Because SMOTE is implemented as a step inside the ImbPipeline, it is automatically refitted in every fold, and synthetic observations are never created using information from outer test folds. This design enforces a strict training-only oversampling strategy and supports the anti-leakage cross-validation scheme requested by the reviewers.

Two tree-based ensemble classifiers are investigated: *Random Forest* and *Extreme Gradient Boosting (XGBoost)*. Both models are used strictly as binary classifiers. *Random Forest* constructs an ensemble of decision trees, each trained on a bootstrap sample of the data with random feature subsampling at each split. The main hyperparameters considered for tuning include the number of trees (n_estimators), the maximum depth (max_depth), the minimum number of samples required to split an internal node (min_samples_split), the minimum number of samples at leaf nodes (min_samples_leaf), and the fraction of features considered at each split (max_features). *XGBoost* implements gradient-boosted decision trees with explicit regularisation and sampling strategies. Its key hyperparameters include the number of boosting iterations (n_estimators), learning rate, maximum tree depth, subsampling ratio of training instances (subsample), column subsampling ratio (colsample_bytree), and L1/L2 regularisation strengths (reg_alpha and reg_lambda). For all scenarios, the classifiers share the same preprocessing, feature-selection, and SMOTE components within the unified pipeline, so that differences in performance can be attributed to the model type and hyperparameter optimisation strategy rather than to differences in data handling.

To obtain competitive models and to compare optimisation strategies in a fair way, three hyperparameter optimisation methods are applied to both classifiers, giving six classification scenarios: *Random Forest* with Randomized Search (RF–RS), *Random Forest* with Bayesian optimisation (RF–Bayes), *Random Forest* with Particle Swarm Optimisation (RF–PSO), *XGBoost* with Randomized Search (XGB–RS), *XGBoost* with Bayesian optimisation (XGB–Bayes), and *XGBoost* with Particle Swarm Optimisation (XGB–PSO). Model selection and performance estimation use the same nested stratified k-fold cross-validation scheme for all scenarios. In the outer loop, the dataset is divided into k folds with approximately equal class proportions. For each outer fold, the pipeline is trained on k−1 folds and evaluated on the remaining fold, which serves as an external test set. Inside each outer training set, an inner stratified k-fold cross-validation loop is used to search for the best hyperparameter configuration according to mean ROC–AUC.

In the Randomized Search setting, candidate hyperparameter vectors are sampled at random from predefined ranges and evaluated via inner cross-validation; the configuration with the highest mean inner ROC–AUC is selected for that outer fold. In the Bayesian optimisation setting, a probabilistic surrogate model of the hyperparameter–performance landscape guides the selection of new candidate configurations by maximising an acquisition function that balances exploration and exploitation. The same inner cross-validation design and ROC–AUC objective are used, but the search trajectory is adaptive rather than purely random. In the PSO setting, each particle in the swarm represents a candidate hyperparameter vector and updates its position by combining its personal best and the global best positions observed so far. Continuous hyperparameters such as learning rate and subsampling ratios are optimised directly, while integer hyperparameters such as the number of trees and maximum depth are rounded to valid values. For each outer training set, PSO runs for a fixed number of iterations and evaluates each particle via inner cross-validation, again using mean ROC–AUC as the objective. In all three optimisation strategies, once the best configuration for a given outer fold is identified, the full pipeline is refitted on the corresponding outer training data and evaluated on the held-out outer test fold. This procedure results in k unbiased test predictions per scenario and ensures that hyperparameter tuning remains strictly separated from final performance estimation.

Classification performance is quantified on the outer test folds using five metrics: accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (ROC–AUC). For each scenario, the mean and standard deviation of these metrics across the outer folds are reported in the Results section. To compare scenarios more rigorously, paired two-tailed t-tests are carried out on the outer-fold ROC–AUC values for every pair of models, with a significance level of 0.05. This statistical analysis highlights whether any observed differences in AUC between scenarios are statistically significant or whether they can be attributed to sampling variability.

To analyse interpretability, the optimised models are further examined using SHapley Additive exPlanations (*SHAP*). After nested cross-validation, the best configurations for each scenario are refitted on the full dataset using the same preprocessing pipeline, and a SHAP KernelExplainer is instantiated on a representative background sample. Global explanations are obtained from SHAP summary plots, which rank features by their mean absolute SHAP value and display the distribution of feature contributions across patients. Local behaviour and potential interactions are explored through SHAP dependence plots for the most influential features, such as chest pain type, number of major vessels, thallium stress-test categories, *oldpeak*, maximum heart rate, and exercise-induced angina. These plots provide a clinically oriented view of how changes in each variable affect the predicted probability of heart disease and form the basis for the discussion of model interpretability.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15544
e-ISSN : 2541-2019
p-ISSN : 2541-044X

All experiments were implemented in Python in a single Jupyter notebook executed in a Google Colab environment. The code was run on Linux (x86_64) with Python 3.12, an Intel Xeon CPU at 2.00 GHz (2 logical cores), approximately 13 GB of RAM, and access to an NVIDIA Tesla T4 GPU with 16 GB of VRAM. Data manipulation was performed with pandas (2.2.2) and numpy (2.0.2), model building and cross-validation with scikit-learn (1.6.1), Random Forest with RandomForestClassifier, XGBoost with the xgboost library (3.1.1), class balancing and imbalanced pipelines with imbalanced-learn (0.14.0), Bayesian optimisation with scikit-optimize (0.10.2), Particle Swarm Optimisation with pyswarm (0.6), and explainability analysis with shap (0.50.0). All Python dependencies were installed via pip at the beginning of the notebook, and a fixed random seed of 42 was used consistently for all stochastic components, including data splitting, classifier initialisation, SMOTE, and optimisation routines. The complete notebook, including environment configuration and the code that generates all reported tables and figures, can be re-executed on the Cleveland dataset to reproduce the results.

RESULT

The proposed heart disease classification pipeline was evaluated using nested stratified cross-validation with an outer 5-fold loop for unbiased performance estimation and an inner 5-fold loop for hyperparameter optimisation. Six classification scenarios were compared: Random Forest and *XGBoost* combined with three optimisation strategies, namely Randomized Search (RS), Bayesian optimisation, and Particle Swarm Optimisation (PSO). For each scenario, the mean and standard deviation of Accuracy, Precision, Recall, F1-score, and ROC-AUC across the outer folds were reported as the primary evaluation metrics.

Table 1. Nested cross-validation performance of heart disease classification models

Scenario	Accuracy (mean ± SD)	Precision (mean ± SD)	Recall (mean ± SD)	F1-score (mean ± SD)	ROC-AUC (mean ± SD)
Random Forest (PSO)	0.8383 ± 0.0164	0.8428 ± 0.0333	0.7981 ± 0.0391	0.8188 ± 0.0206	0.9089 ± 0.0146
XGBoost (Bayes)	0.8381 ± 0.0272	0.8527 ± 0.0369	0.7836 ± 0.0486	0.8157 ± 0.0342	0.9084 ± 0.0162
Random Forest (RS)	0.8349 ± 0.0186	0.8508 ± 0.0259	0.7767 ± 0.0289	0.8117 ± 0.0225	0.9065 ± 0.0216
Random Forest (Bayes)	0.8383 ± 0.0164	0.8438 ± 0.0395	0.7981 ± 0.0451	0.8187 ± 0.0210	0.9064 ± 0.0188
XGBoost (RS)	0.8316 ± 0.0221	0.8400 ± 0.0362	0.7839 ± 0.0342	0.8102 ± 0.0250	0.8998 ± 0.0207
XGBoost (PSO)	0.8283 ± 0.0229	0.8399 ± 0.0480	0.7767 ± 0.0289	0.8060 ± 0.0245	0.8908 ± 0.0257

Table 1 summarises the nested cross-validation results. All six scenarios achieved relatively high and stable classification performance, with mean ROC-AUC values ranging from 0.8908 to 0.9089. The best overall configuration in terms of ROC-AUC was Random Forest with PSO optimisation, which obtained an Accuracy of 0.8383 ± 0.0164, a Precision of 0.8428 ± 0.0333, a Recall of 0.7981 ± 0.0391, an F1-score of 0.8188 ± 0.0206, and a ROC-AUC of 0.9089 ± 0.0146. *XGBoost* with Bayesian optimisation reached very similar performance (Accuracy 0.8381 ± 0.0272, ROC-AUC 0.9084 ± 0.0162), followed by Random Forest with RS and Bayesian optimisation, which also produced ROC-AUC values above 0.9060. The *XGBoost* RS and PSO scenarios showed slightly lower yet still competitive AUC values (0.8998 and 0.8908 respectively). These results indicate that, under a properly regularised and nested evaluation, both Random Forest and *XGBoost* can provide robust heart disease classification, while different optimisation strategies mainly yield incremental rather than drastic improvements.

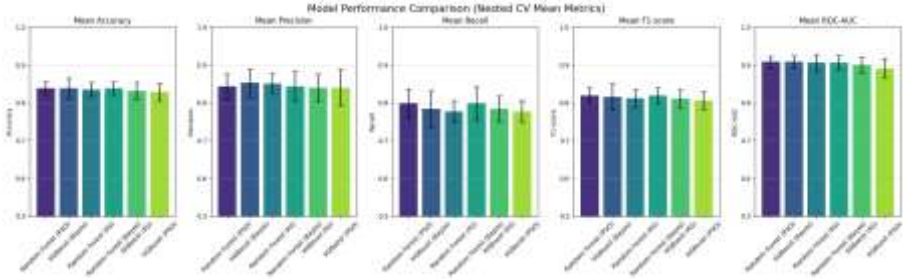

Fig 2 Model performance comparison (nested cross-validation mean metrics) for Random Forest and XGBoost under Randomized Search, Bayesian optimisation, and PSO.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15544
e-ISSN : 2541-2019
p-ISSN : 2541-044X

To assess whether the observed differences in ROC-AUC are statistically meaningful, paired t-tests were conducted on the outer-fold AUC values for all pairs of scenarios. The p-values indicate that most pairwise differences are not statistically significant at the 0.05 level, especially among the top-performing configurations (Random Forest PSO, Random Forest Bayes, Random Forest RS, and *XGBoost* Bayes). A statistically significant difference ($p \approx 0.047$) was observed only when comparing Random Forest with Bayesian optimisation against *XGBoost* with PSO, where the latter scenario produced noticeably lower ROC-AUC. Overall, this analysis suggests that, although Random Forest PSO achieves the highest mean AUC, the performance gap relative to the other well-tuned models is modest and should be interpreted as an incremental improvement rather than a fundamentally different regime of accuracy.

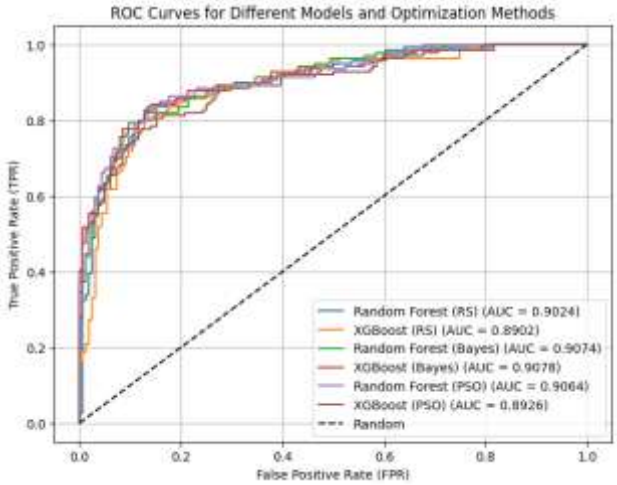

Fig 3. ROC curves for Random Forest and XGBoost with Randomized Search, Bayesian optimisation, and PSO under the best nested-CV configurations.

A detailed confusion matrix was generated for each scenario using the predictions aggregated over the outer folds. For the best configuration, Random Forest PSO, the confusion matrix shows 143 true negatives, 21 false positives, 28 false negatives, and 111 true positives.

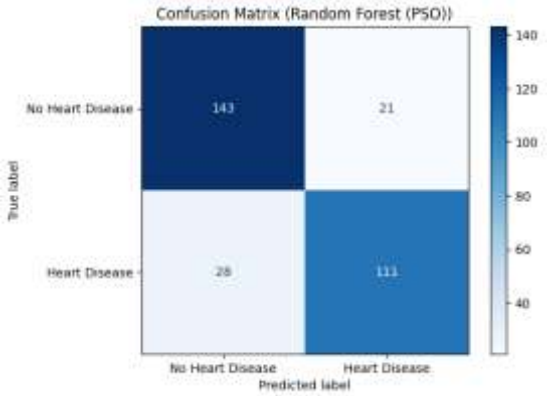

Fig 4. Confusion matrix of the best heart disease classification model (Random Forest PSO).

These counts correspond to an empirical Accuracy of 0.838, a Precision of approximately 0.841, a Recall of 0.789, and an F1-score of 0.819 on the combined test samples, which is consistent with the nested cross-validation summary. The relatively low number of false negatives compared to true positives indicates that the model can correctly identify most patients with heart disease, while maintaining an acceptable level of false alarms. From a clinical perspective, this balance between sensitivity (Recall) and specificity (driven by the number of false positives) is important because missing positive cases is usually more critical than over-diagnosing a subset of low-risk patients.

To further characterise the behaviour of the best model, a separate bar chart was produced that visualises its mean metrics with standard deviation bars.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15544
e-ISSN : 2541-2019
p-ISSN : 2541-044X

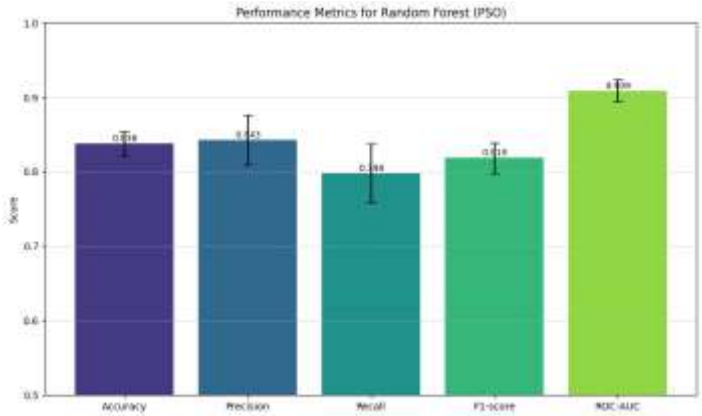

Fig 5. Performance metrics (nested cross-validation mean ± standard deviation) for the best heart disease classification model (Random Forest PSO).

Beyond pure predictive performance, this study examined the interpretability of the optimised classifiers using SHapley Additive exPlanations (*SHAP*). Global feature importance was first analysed through SHAP summary plots for the most competitive scenarios (Random Forest RS, Random Forest Bayes, Random Forest PSO, *XGBoost* Bayes, and *XGBoost* PSO). Across all models, a consistent pattern of influential predictors emerged. Binary encodings of the number of major vessels visualised by fluoroscopy (ca_0.0 and related dummy variables), chest-pain types (particularly cp_4.0, representing asymptomatic chest pain), thallium stress-test categories (thal_3.0 and thal_7.0), ST-segment depression (*oldpeak*), and maximum heart rate achieved (*thalach*) repeatedly appeared among the top features. This stability of the global SHAP rankings across different algorithms and optimisation strategies indicates that the models rely on a clinically coherent subset of cardiovascular risk factors rather than on spurious artefacts of the dataset.

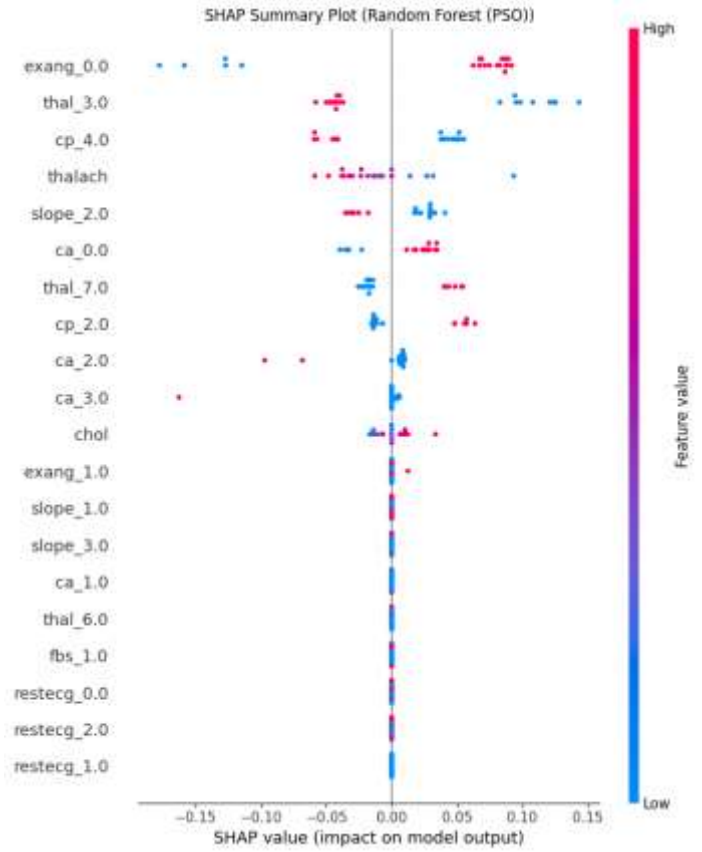

Fig 6. SHAP summary plot for the best heart disease classification model (Random Forest PSO).

Local behaviour of the models was then examined using SHAP dependence plots for the most influential features. For *oldpeak*, the dependence plots show a clear monotonic relationship where higher ST-segment

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15544
e-ISSN : 2541-2019
p-ISSN : 2541-044X

depression values are associated with increasingly positive SHAP values, meaning a higher contribution to the prediction of heart disease. In contrast, *thalach* exhibits an inverse pattern: lower maximum heart rate values tend to produce positive SHAP contributions (increasing the probability of disease), while higher heart rates shift SHAP values towards the negative side, indicating a protective effect. The thallium stress-test categories thal_3.0 and thal_7.0, which correspond to abnormal perfusion patterns, consistently yield positive SHAP values when active, reflecting a strong association with heart disease in both Random Forest and *XGBoost* models. The dummy variables related to the number of affected vessels (ca_*), as well as indicators of exercise-induced angina and specific chest-pain types, also show clearly separated SHAP distributions between their inactive and active states, underlining their importance in shaping the model's decisions.

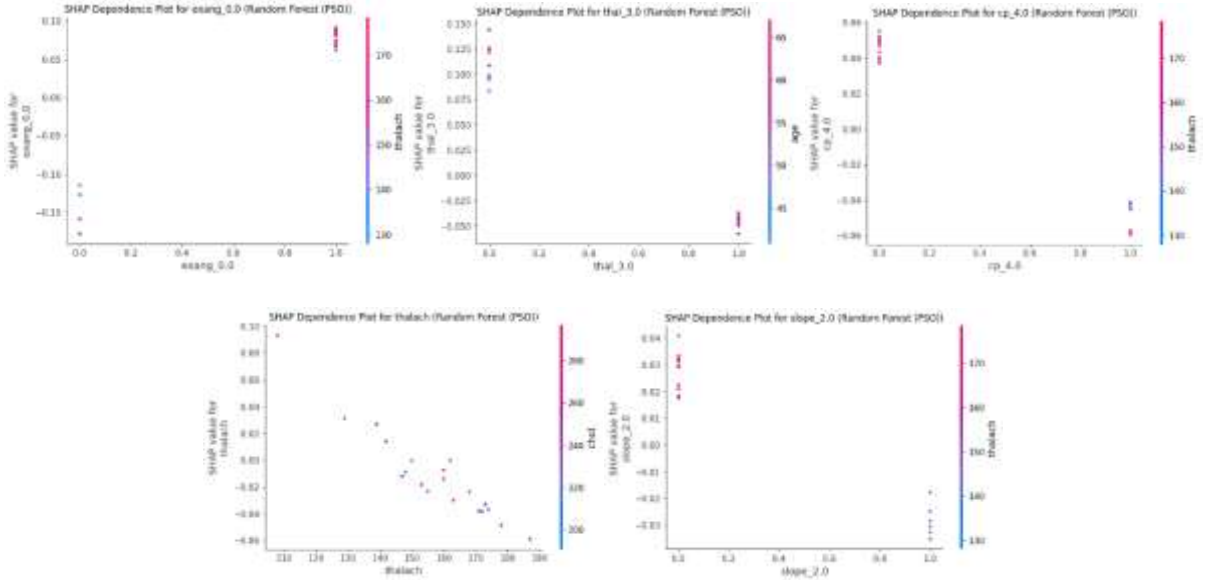

Fig 7. SHAP dependence plots for selected high-impact features in the best heart disease classification models.

When the SHAP results for Random Forest PSO are compared with those for *XGBoost* Bayes and the other optimisation scenarios, the overall ranking and qualitative direction of the main features remain very similar, even though individual SHAP magnitudes differ slightly between algorithms. This indicates that the proposed classification pipeline not only delivers competitive predictive performance but also produces interpretable patterns that are stable across different model families and hyperparameter optimisation strategies. Such stability strengthens the confidence that the learned relationships reflect underlying cardiovascular mechanisms rather than noise specific to a single model configuration.

DISCUSSIONS

The experimental results demonstrate that tree-based ensemble classifiers remain highly competitive for heart-disease classification on the Cleveland dataset. Across all optimisation settings, Random Forest and XGBoost achieved relatively stable performance, with mean ROC–AUC values ranging from 0.8908 to 0.9089 in nested cross-validation. Among all scenarios, Random Forest with PSO optimisation obtained the highest ROC–AUC (0.9089 ± 0.0146), followed very closely by XGBoost tuned with Bayesian optimisation (0.9084 ± 0.0162). The differences between these two models were not statistically significant at the 5% level, indicating that both ensemble approaches are capable of capturing the underlying decision patterns in the clinical attributes of the dataset.

Although the difference in predictive performance is small, the relative behaviours of the optimisation strategies provide additional insight. PSO generated configurations that improved recall without compromising precision drastically, suggesting an advantage in exploring broader hyperparameter spaces in discrete steps. Conversely, Bayesian optimisation tended to produce more conservative but stable parameter combinations, resulting in balanced precision–recall trade-offs. Random search produced reasonably strong baselines but consistently ranked below both PSO and Bayesian optimisation in terms of ROC–AUC. These patterns show that the choice of optimisation method can influence model behaviour even when improvements in overall accuracy appear marginal. In addition, the nested cross-validation protocol ensured strict separation between tuning and evaluation, thereby preventing information leakage a limitation commonly found in earlier studies using single-split or non-nested validation.

*name of corresponding author

The confusion matrix of the best-performing model (Random Forest PSO) shows that most misclassifications occurred in borderline cases, particularly for samples with mild or ambiguous clinical presentations. This aligns with findings from previous studies that models trained on the Cleveland dataset perform well for more distinct cases but may struggle with physiological overlaps between positive and negative classes. However, the overall error distribution remained balanced, indicating that the model did not suffer from extreme bias toward majority classes.

The SHAP analyses provide further explanation of how the ensemble models make decisions. Across all optimised configurations, the most influential features consistently included *oldpeak* (ST depression induced by exercise), *ca* (number of major vessels), *thal* (perfusion defect category), *cp* (chest pain type), *thalach* (maximum heart rate), and *exang* (exercise-induced angina). The prominence of *oldpeak* aligns closely with cardiology literature, as ST-segment abnormalities are well-established markers of myocardial ischemia. Similarly, *ca* and *thal* reflect anatomical and perfusion abnormalities, which are strong indicators of coronary artery disease severity. The importance of *cp* and *exang* further suggests that symptom-triggered indicators remain clinically relevant for risk stratification. The consistency of these findings across Random Forest and XGBoost despite differing optimisation strategies indicates that the physiological interpretation of the model is robust and not dependent on a specific learning algorithm.

An important observation is that SHAP provided coherent global and local explanations. Globally, features representing structural and perfusion-related abnormalities (*ca*, *thal*, *oldpeak*) contributed most to increased risk. Locally, certain patients showed stronger influences from functional attributes such as *thalach* or *cp*, reflecting individual variations in the manifestation of heart disease. This combination of stable global patterns and clinically interpretable local variations strengthens the trustworthiness of the model's decisions. Furthermore, the coherence between SHAP feature rankings and known clinical markers supports the potential utility of the model as a decision-support component, although clinical deployment would require additional external validation.

Comparing the findings with previous work, this study provides three main contributions. First, unlike most prior studies that evaluate only a single optimisation method or perform limited tuning, this work systematically compared random search, Bayesian optimisation, and PSO across both Random Forest and XGBoost using a nested, anti-leakage protocol. Second, the statistical comparison of ROC–AUC values provides a more rigorous assessment of whether observed differences are meaningful rather than incidental. Third, the integration of SHAP across all optimised models enables a cross-model interpretability analysis, which reveals the stability of important features independent of the chosen optimisation strategy. This directly addresses gaps identified in earlier studies, where XAI was often applied to only one model or without examining the consistency of explanations across configurations.

Despite these strengths, several limitations must be acknowledged. The Cleveland dataset remains relatively small and historical, limiting the generalisability of results to contemporary or demographically diverse populations. The models were evaluated solely on internal cross-validation; external validation on multi-centre datasets would be required before clinical deployment. Additionally, SHAP analyses, although consistent, were performed on a limited subset of samples due to computational cost. Larger-scale interpretability analyses may reveal further nuances in feature interactions or subgroup-specific feature influences. Lastly, this study focuses on model performance and interpretability but does not address integration with clinical workflows or real-time diagnostic systems.

Overall, the findings indicate that ensemble classifiers optimised through PSO or Bayesian optimisation can provide highly accurate and interpretable predictions for heart-disease classification using the Cleveland dataset. The consistent SHAP explanations across models reinforce the physiological validity of important features, demonstrating the potential for model-assisted decision support while highlighting the need for broader clinical validation.

CONCLUSION

This study evaluated Random Forest and XGBoost for heart-disease classification using the Cleveland dataset under a nested, anti-leakage validation framework. Three hyperparameter optimisation strategies random search, Bayesian optimisation, and particle swarm optimisation (PSO) were systematically compared to assess their impact on model performance and interpretability. The results showed that ensemble-based models consistently produced strong predictive performance, with ROC–AUC values ranging from 0.8908 to 0.9089 across all scenarios. Random Forest optimised with PSO achieved the highest overall performance (ROC–AUC 0.9089 ± 0.0146), while XGBoost optimised with Bayesian optimisation delivered comparably strong results without significant statistical differences. These findings indicate that both ensemble classifiers are well suited for structured clinical data, and that optimisation strategies can influence model behaviour even when performance differences appear small.

The SHAP analyses further demonstrated that model explanations remained stable across algorithms and optimisation configurations. Features such as *oldpeak*, *ca*, *thal*, *cp*, *thalach*, and *exang* consistently emerged as

*name of corresponding author

key determinants of risk, aligning with established clinical evidence. This consistency highlights the robustness of the interpretability framework and underscores its potential value in supporting clinical decision-making, particularly in identifying physiologically meaningful predictors.

While the study provides a rigorous comparison of optimisation strategies and cross-model interpretability, several limitations must be acknowledged. The Cleveland dataset is relatively small and historical, which restricts the generalisability of the findings to broader, multi-centre populations. SHAP analyses were performed on a limited sample due to computational constraints, and external validation was not conducted. The models also remain untested in real-world clinical workflows. These limitations present opportunities for future work, particularly in validating the models on modern clinical datasets, analysing explanation stability across larger patient cohorts, and integrating model outputs into clinical decision-support systems.

Overall, the study demonstrates that combining ensemble classifiers with systematic optimisation and explainable AI techniques can produce accurate and interpretable heart-disease classification models. The framework presented here addresses key methodological gaps in previous studies by employing rigorous nested validation, comparative optimisation, and consistent SHAP-based explanation analysis. The results contribute to a more reliable and transparent foundation for the development of clinically oriented predictive models and offer a basis for future extensions toward real-world deployment and more comprehensive clinical evaluation.

REFERENCES

Al-Alshaikh, H. A., Prabu, P., Poonia, R. C., Saudagar, A. K. J., Yadav, M., AlSagri, H. S., & AlSanad, A. A. (2024). Comprehensive evaluation and performance analysis of machine learning in heart disease prediction. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-58489-7

Alkhanbouli, R., Matar Abdulla Almadhaani, H., Alhosani, F., & Simsekler, M. C. E. (2025). The role of explainable artificial intelligence in disease prediction: a systematic literature review and future research directions. In *BMC Medical Informatics and Decision Making* (Vol. 25, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s12911-025-02944-6

Alsabhan, W., & Alfadhly, A. (2025). Effectiveness of machine learning models in diagnosis of heart disease: a comparative study. *Scientific Reports*, *15*(1). https://doi.org/10.1038/s41598-025-09423-y

Ansyari, M. R., Mazdadi, M. I., Indriani, F., Kartini, D., & Saragih, T. H. (2023). ShareAlike 4.0 International License (CC BY-SA 4.0). "Implementation of Random Forest and Extreme Gradient Boosting in the Classification of Heart Disease Using Particle Swarm Optimization Implementation of Random Forest and Extreme Gradient Boosting in the Classification of Heart Disease Using Particle Swarm Optimization Feature Selection. *Open Access Journal*, *5*(4), 250–260. https://doi.org/10.35882/jeemi.v5i4.322

Biswas, N., Ali, M. M., Rahaman, M. A., Islam, M., Mia, M. R., Azam, S., Ahmed, K., Bui, F. M., Al-Zahrani, F. A., & Moni, M. A. (2023). Machine Learning-Based Model to Predict Heart Disease in Early Stage Employing Different Feature Selection Techniques. *BioMed Research International*, *2023*. https://doi.org/10.1155/2023/6864343

Bouqentar, M. A., Terrada, O., Hamida, S., Saleh, S., Lamrani, D., Cherradi, B., & Raihani, A. (2024). Early heart disease prediction using feature engineering and machine learning algorithms. *Heliyon*, *10*(19). https://doi.org/10.1016/j.heliyon.2024.e38731

Breiman, L. (2001). *Random Forests* (Vol. 45).

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-August-2016*, 785–794. https://doi.org/10.1145/2939672.2939785

Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J.-J., Sandhu, S., Guppy, K. H., Lee, S., & Froelicher, V. (n.d.). *International Application of a New Probability Algorithm for the Diagnosis of Coronary Artery Disease*.

El-Sofany, H., Bouallegue, B., & El-Latif, Y. M. A. (2024). A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-74656-2

Gnanavelu, A., Venkataramu, C., & Chintakunta, R. (2025). Cardiovascular Disease Prediction Using Machine Learning Metrics. *Journal of Young Pharmacists*, *17*(1), 226–233. https://doi.org/10.5530/jyp.20251231

Hindarto, D. (2024). Case Study: Gradient Boosting Machine vs Light GBM in Potential Landslide Detection. *Journal of Computer Networks, Architecture and High Performance Computing*, *6*(1), 169–178. https://doi.org/10.47709/cnahpc.v6i1.3374

Islam, R. Bin, Akhter, S., Iqbal, F., Saif Ur Rahman, M., & Khan, R. (2023). Deep learning based object detection and surrounding environment description for visually impaired people. *Heliyon*, *9*(6). https://doi.org/10.1016/j.heliyon.2023.e16924

Lundberg, S. M., Allen, P. G., & Lee, S.-I. (n.d.). *A Unified Approach to Interpreting Model Predictions*. https://github.com/slundberg/shap

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Mienye, I. D., & Jere, N. (2024). Optimized Ensemble Learning Approach with Explainable AI for Improved Heart Disease Prediction. *Information (Switzerland)*, *15*(7). https://doi.org/10.3390/info15070394

Rezk, N. G., Alshathri, S., Sayed, A., El-Din Hemdan, E., & El-Behery, H. (2024). XAI-Augmented Voting Ensemble Models for Heart Disease Prediction: A SHAP and LIME-Based Approach. *Bioengineering*, *11*(10). https://doi.org/10.3390/bioengineering11101016

Teja, M. D., & Rayalu, G. M. (2025). Optimizing heart disease diagnosis with advanced machine learning models: a comparison of predictive performance. *BMC Cardiovascular Disorders*, *25*(1). https://doi.org/10.1186/s12872-025-04627-6

*name of corresponding author