Vol. 8, No. 3, June 2026

Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

Two-Stage Tuning of Machine Learning Models for Heart Disease Classification on Synthetic Data

Marini[1], Tri Sugihartono[2], Chandra Kirana[3], Benny Wijaya[4], Hamidah[5]

[1,2,3,4]Information Technology Faculty, Institut Sains dan Bisnis Atma Luhur, Pangkalpinang, Indonesia

[5]Economic Business Faculty, Institut Sains dan Bisnis Atma Luhur, Pangkalpinang, Indonesia

Received:
October 9, 2025
Revised:
April 19, 2026
Accepted:
May 27, 2026
Published:
June 22, 2026

Corresponding Author:
Author Name*:
Marini
Email*:
arinimarini44@atmaluhur.ac.id

DOI:
10.63158/journalisi.v8i3.1599

Abstract. Heart disease remains a leading global cause of mortality, highlighting the need for accurate early risk classification. This study benchmarks Random Forest, XGBoost, and Logistic Regression for heart disease risk classification using a synthetic, perfectly balanced dataset, while addressing performance limitations caused by inadequate hyperparameter configuration. The dataset comprised 70,000 samples with a 50/50 class distribution and 18 clinical and demographic features. Although useful for controlled benchmarking, synthetic balanced data may yield optimistic estimates and may not fully represent real-world clinical variability. Each model was implemented in a scikit-learn Pipeline with median imputation and, where applicable, standard scaling. A two-stage tuning strategy was applied by combining RandomizedSearchCV with GridSearchCV refinement to optimize model configurations systematically. Under these benchmarking conditions, XGBoost achieved the best test performance, with an F1-score of 99.34%, AUC-ROC of 99.97%, and accuracy of 99.34%. Random Forest obtained an F1-score of 99.20% and AUC-ROC of 99.95%, while Logistic Regression achieved an F1-score of 99.12% and AUC-ROC of 99.95%. Age, pain in the arms/jaw/back, and cold sweats/nausea were the most influential predictors. The proposed framework is reproducible, computationally efficient, and suitable for validation on heterogeneous clinical datasets.

Keywords: Hyperparameter Optimization; Machine Learning; Comparative Benchmarking; Synthetic Medical Data; AUC-ROC;

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

1. INTRODUCTION

Heart Cardiovascular disease continues to be one of the most serious public health problems. Cardiovascular disease remains a major global health challenge, causing approximately 17.9 million deaths annually according to the World Health Organization (WHO)[1]. The ability to accurately Screening individuals with significant risk factors paramount Clinical screening, while effective, are often requires significant time, resource- complex and prone to manual errors. Provided new opportunities for designing data-oriented automated diagnostic and predictive systems, capable of handling massive clinical data effectively and accurately.

Algorithms extensively domain medical diagnosis and risk classification. Several studies have demonstrated the potential of ensemble methods and linear classifiers in predicting cardiovascular conditions. Shorewala [2]. Applying logistic regression along with SVM method, achieving accuracy up to 85.5%, but noted limitations in handling high-dimensional datasets without feature selection. Javid et al [3]. employed Random Forest to Cleveland, reporting an accuracy of 87.3%, although constrained 303 instances, limiting generalizability. Fitriyani et al [4]. Formulating a predictive approach to heart disease combining SMOTE oversampling with Random Forest, obtaining F1-Score of 90.6% on an imbalanced dataset, yet the study relied solely on a single algorithm without comparative benchmarking against other classifiers. Ramalingam et al[5]. of multiple ML classifiers namely Trees, and Decision for heart disease prediction, concluding that ensemble consistently outperformed single classifiers, though hyperparameter optimization was not systematically addressed. More recently, Guo et al[6]. Benchmarked gradient boosting and ensemble models for cardiovascular risk prediction, demonstrating XGBoost's competitive advantage in structured tabular medical data with AUC-ROC above 93%, however, no multi-algorithm comparative framework with rigorous two-stage tuning strategy was employed.

A critical observation across existing literature is that while individual algorithms have been evaluated extensively, systematic and fair comparative studies incorporating robust hyperparameter optimization remain scarce. Most prior works either rely on default hyperparameter configurations, apply single-stage tuning with limited search space, or compare algorithms without a unified optimization framework, leading to potentially

biased performance conclusionset.al[7]. Furthermore, feature importance analysis across multiple algorithms simultaneously, which offers deeper insight into clinical risk factors, has rarely been conducted in a structured comparative setting.

A few researchers focused on multi-algorithm comparison for heart disease classification; however, there have been limited studies concerned with applying a two-stage sequential hyperparameter tuning strategy — combining RandomizedSearchCV and GridSearchCV refinement — as a unified optimization engine across Random Forest, XGBoost, and Logistic Regression simultaneously[8]. It is also noted that prior studies rarely acknowledge the impact of dataset characteristics — particularly class balance and data origin (real-world vs. synthetic) — on the interpretability of reported performance metrics. This research addresses these gaps by proposing a systematic two-stage tuning framework that ensures each algorithm is evaluated under its optimal configuration, enabling a fair, rigorous, and reproducible comparative analysis. Nevertheless, it is important to acknowledge upfront that this study is conducted on a large-scale synthetic dataset with a perfectly balanced class distribution, which may produce performance values that are more optimistic than those achievable on real-world heterogeneous clinical data. (1) to implement optimize Random Forest, XGBoost, and Logistic Regression using a two-stage hyperparameter tuning strategy; (2) to comparatively evaluate multiple AUC-ROC, Accuracy, Precision, and Recall; (3) to analyze and compare feature importance across all three algorithms to identify the most clinically significant risk factors for heart disease; and (4) to provide a reproducible ML benchmarking pipeline applicable to broader clinical risk classification tasks, with the understanding that real-world validation remains a necessary next step.

The use of synthetic medical datasets for benchmarking machine learning models has gained increasing recognition as a valid methodology, particularly in contexts where patient privacy regulations constrain access to real-world electronic health records [9]. Gonzales et al[9]. validated synthetic health data as an effective substitute for real data in algorithm development when privacy constraints are present, while also cautioning that model performance on synthetic data may not generalize directly to clinical environments. Obermeyer and Emanuel [10]. Further emphasized that the transition from algorithmic benchmarking to clinical deployment requires rigorous external validation, as performance differences between algorithms on controlled datasets do not reliably

predict relative advantage in heterogeneous patient populations. These considerations directly motivate the framing of the current study as a controlled benchmarking exercise rather than a clinical validation, while also reinforcing the importance of transparency regarding dataset origin and characteristics when interpreting reported results [9]. Against this backdrop, this study makes a contribution to the growing body of reproducible ML benchmarking literature by providing a fully documented, two-stage tuning pipeline with explicit synthetic-data characterization, enabling future researchers to build upon or contrast these findings as part of broader comparative efforts.

2. METHODS

This research follows a structured machine learning pipeline consisting of exploratory analysis, model training with two-stage hyperparameter tuning, model evaluation, and comparative analysis, as shown in Figure 1.

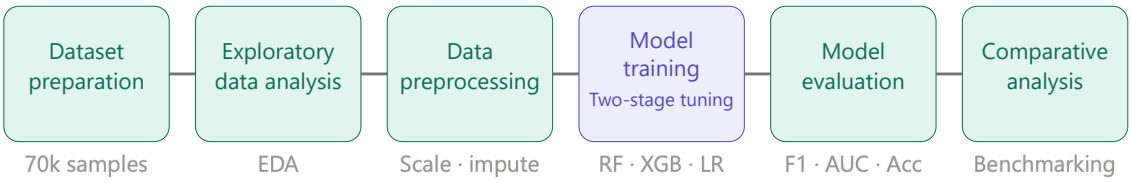


Figure 1. Research workflow of the proposed machine learning pipeline

2.1 Dataset

This study utilized a synthetic dataset comprising 80,000 samples with 18 clinical and demographic features and one binary target variable. The dataset was generated using a probabilistic risk-scoring mechanism based on clinical literature [1]. Specifically, each sample was assigned a composite risk score computed as a weighted sum of binary and continuous feature values, where features were drawn from defined probability distributions reflecting epidemiological prevalence rates (e.g., smoking prevalence ~30%, diabetes ~15%, hypertension ~35%). The Heart_Risk label was determined by thresholding the composite score against a calibrated boundary to achieve a deliberate 50/50 class distribution. Features included age, gender, chest discomfort, difficulty breathing, fatigue, irregular heartbeat, dizziness, body swelling, pain radiating to the arm/jaw/back, cold sweats or nausea, hypertension, high cholesterol, diabetes mellitus, tobacco use, excess body weight, inactive lifestyle, and family history and prolonged stress The dataset was

deliberately balanced with a 50/50 class distribution (35,000 samples per class) to isolate the effect of algorithm choice and hyperparameter tuning from class-imbalance handling, which is a common confounding factor in prior comparative studies. It is important to acknowledge that this perfectly balanced, synthetic design simplifies the classification task and limits direct clinical realism; real-world heart disease datasets are typically imbalanced and exhibit greater inter-feature noise. Accordingly, the results of this study undertood as benchmarking evidence under conditions rather than direct clinical performance guarantees. Large-scale synthetic data in machine learning research for medical classification has been validated as an effective approach for algorithm benchmarking when real-world data is constrained by privacy regulations [3].

To ensure reproducibility, the risk-score generation process is formalized as shown in Equation 1. For each sample i, a composite risk score S_i is computed as a weighted linear combination of feature values:

$$S_i = w_1 \cdot Age_i + w_2 \cdot Pain\_Arms_i + w_3 \cdot Cold\_Sweats_i + \ldots + w_{18} \cdot Chronic\_Stress_i \quad (1)$$

where each binary feature $x_j \sim$ Bernoulli(p_j) with prevalence probability p_j drawn from clinical epidemiology (e.g., p_j(smoking)=0.30, p_j(diabetes)=0.15, p_j(hypertension)=0.35), and Age $\sim$ Uniform(30, 80). Feature weights w_j were assigned proportional to established cardiovascular risk contribution, with age and symptomatic features receiving higher weights. The binary label is then assigned via threshold θ:

$Heart\_Risk_i = 1$ *if* $S_i \geq \theta$, *else 0; where* θ *is calibrated to yield* $|class\ 0| = |class\ 1| = 35{,}000$

A representative sample of the dataset (first and last 5 rows) is presented in Table 0 to provide transparency regarding the data structure. The model configuration settings for each algorithm are summarized in Table 0b, and the full hyperparameter tuning search spaces are detailed in the Two-Stage Hyperparameter Tuning subsection.

Table 1. Base model configuration settings for Random Forest, XGBoost, and Logistic Regression

Parameter	Random Forest (RF)	XGBoost (XGB)	Logistic Regression (LR)
Base estimator	Decision Tree	Gradient Boosted Tree	Logistic (sigmoid)
class_weight	balanced	—	balanced
oob_score	True	—	—
eval_metric	—	logloss	—
verbosity	—	0	—
max_iter	—	—	3000
Regularization	—	L1 + L2 (reg_alpha, reg_lambda)	L1 / L2 (tuned)
Solver	—	—	liblinear / saga
Scaling required	No (tree-based)	No (tree-based)	Yes (StandardScaler)
Pipeline	scikit-learn Pipeline	scikit-learn Pipeline	scikit-learn Pipeline
Imputation	SimpleImputer (median)	SimpleImputer (median)	SimpleImputer (median)
CV strategy (tuning)	5-fold stratified CV	5-fold stratified CV	5-fold stratified CV
Scoring metric	F1-Score	F1-Score	F1-Score

2.2 Exploratory Data Analysis (EDA)

previus to modeling, a comprehensive o understand the distributional properties and inter-feature relationships within dataset. Class was verified through bar charts and pie charts to confirm dataset balance. Pearson correlation coefficients were computed between each feature and the target variable to assess linear associations [4]. The top five features most correlated with Heart_Risk were identified as Age (r = 0.605), Pain_Arms_Jaw_Back (r = 0.601), Cold_Sweats_Nausea (r = 0.601), Dizziness (r = 0.600), and Chest_Pain (r = 0.600). Age distribution per risk class was analyzed through histograms and boxplots, revealing a clear separation between risk groups, with mean age of 44.5 years for low-risk and 64.4 years for high-risk patients. Binary feature prevalence was analyzed using cross-tabulation bar charts per target class.

2.3 Data Preprocessing

All features were passed through a unified scikit-learn Pipeline to ensure consistent and reproducible preprocessing [5]. The pipeline architecture is critical for preventing The preprocessing steps consisted of two components. First, missing value replacement was executed with SimpleImputer utilizing median strategy, which is robust to outliers and appropriate for clinical data [6]. Although the dataset contained no missing values, the imputer was retained in the pipeline to ensure robustness when deployed on real-world data. Second, for Logistic Regression, an additional StandardScaler was applied to normalize feature distributions to standardized by subtreacting the mean, divind, differences [7]. Random Forest and XGBoost scaling due [8]. Deliberately applied only inside the Logistic Regression pipeline branch and not to tree-based models, as scaling does not affect the performance of decision-tree splitting criteria. Partitioned testing sets using stratified split with a ratio of 80:20, with stratification applied to maintain class balance across both sets [9].

2.4 Model Architecture

Three machine learning algorithms were selected for this comparative study based on their widespread adoption and demonstrated effectiveness in medical classification tasks.

2.4.1 Random Forest (RF)

Random Forest an group techniqui builids numerous treesinging during training phase determines final class on majority vote of prediction generates by individual trees[11]. selection of the training data, and at every node, a randomly chosen subset of features is taken into account for splitting, which reduces variance and mitigates overfitting. Random Forest has been shown to perform robustly on high-dimensional medical datasets with mixed feature types[12]. Class_weight='balanced' and oob_score=True to leverage out-of-bag estimation during training.

2.4.2 XGBoost (XGB)

A scalable and optimized of the sequential enseble of the previous one[13]. XGBoost includes L1 and L2 regularization components (reg_ alpha and reg_ lambda) to prevent overfitting, making it particularly effective on structured tabular data. Chen and Guestrin [13]. demonstrated that XGBoost consistently outperforms other algorithms on a wide

range of classification benchmarks. In this study, XGBoost was configured with eval_metric='logloss' and verbosity=0 for stable training.

2.4.3 Logistic Regression (LR)

Logistic Regression is a fundamental linear classification algorithm that models the probability of class membership using the logistic (sigmoid) function[14]. Despite its simplicity, Logistic Regression has been shown to achieve cometitive performance on linearly separable medical datasets and serves as an important interpretable baseline in comparative ML studies [15]. In this study, Logistic Regression was configured with class_weight='balanced', max_iter=3000, and both L1 and L2 regularization penalties were explored during tuning, with solvers liblinear and saga.

2.5 Two-Stage Hyperparameter Tuning

Research is the proposed hyperparameter tuning engine, designed to systematically optimize each model's configuration within a unified and reproducible framework. Hyperparameter tuning has been established as a critical step in maximizing model performance, as default configurations rarely yield optimal results on domain-specific datasets [16].

1) Stage 1 — RandomizedSearchCV

In the first stage, RandomizedSearchCV was applied with 35 iterations. RandomizedSearchCV samples hyperparameter combinations randomly from predefined distributions, providing a computationally efficient option compares to grid search while maintaining broad exploration of the search space[17]. The search distributions for each model are defined as follows:

a) Random Forest: n_estimators ∈ randint(100, 400), max_depth ∈ randint(10, 60), min_samples_split ∈ randint(2, 15), min_samples_leaf ∈ randint(1, 8), max_features ∈ {sqrt, log2, 0.3, 0.5}, class_weight ∈ {balanced, balanced_subsample, None}, criterion ∈ {gini, entropy, log_loss}.
b) XGBoost: n_estimators ∈ randint(100, 300), max_depth ∈ randint(3, 10), learning_rate ∈ loguniform(0.01, 0.3), subsample ∈ uniform(0.6, 0.4), colsample_bytree ∈ uniform(0.6, 0.4), reg_alpha ∈ loguniform(1e-4, 10), reg_lambda ∈ loguniform(1e-4, 10), min_child_weight ∈ randint(1, 7).

c) Logistic Regression: 'C ∈ loguniform(0.001, 100), penalty ∈ {l1, l2}, solver ∈ {liblinear, saga}, max_iter ∈ {1000, 2000, 3000}, class_weight ∈ {balanced, None}'.

2) Stage 2 — GridSearchCV Refinement

In the second stage, GridSearchCV was applied around the optimal region identified in Stage 1 using a narrowed grid of candidate values. GridSearchCV conducts a thorugh search across the defined parameter grid, guaranteeing locally optimal configuration is precisely identified [18]. The refinement grids were defined as follows:

a) Random Forest: 'n_estimators ∈ {200, 300}, max_depth ∈ {20, 30, 40, None}, min_samples_leaf ∈ {1, 2, 4}'.
b) XGBoost: 'n_estimators ∈ {150, 250}, max_depth ∈ {4, 6, 8}, learning_rate ∈ {0.05, 0.1}'.
c) Logistic Regression: 'C ∈ {0.1, 1, 10}, penalty ∈ {l2}, solver ∈ {lbfgs, saga}'.

The Stage 2 result was adopted only when it produced a higher cross-validation F1-Score than Stage 1; otherwise, the Stage 1 best estimator was retained. This conditional refinement strategy ensures computational efficiency without sacrificing performance gains.

2.6 Model Evaluation

Conducted on the reserved (14,000 samples) testing data using six evaluation metrics: Accuracy, Precision, Recall, F1-Score, AUC-ROC, and Average Precision (AP). These metrics were selected to supply e a comprehensive perspective of classifier performance under balanced class conditions [19]. Additionally, coss-validation performed on the full dataset for each tuned model to assess generalization stability and variance. Learning curves were generated for Random Forest to diagnose bias-variance trade-off. 'Confusion matrices, ROC curves, and Precision-Recall curves were visualized for all three models to enable qualitative comparison'. Feature importance was extracted from Random Forest (Gini impurity-based) and XGBoost (gain-based), while absolute normalized coefficients were used for Logistic Regression, enabling cross-algorithm feature relevance comparison [20].

2.7 Development Environment

All experiments were implemented in Python 3.x using the following libraries: using hthe-learn version 1.6.1 for model development, pipeline construction, and evaluation; XGBoost

for gradient boosting implementation; 'pandas 2.2.2 and numpy 2.0.2 for data manipulation'; matplotlib and seaborn for visualization; and psutil for memory-aware parallel job allocation. All experiments were conducted on a system with 12.26 GB available RAM, utilizing n_jobs=2 for parallel processing based on dynamic memory assessment.

3. RESULTS AND DISCUSSION

3.1 Exploratory Data Analysis

The dataset utilized in this study consists of 70,000 samples with a perfectly balanced class distribution of 50% low-risk and 50% high-risk cases, as shown in Figure 2. This balance eliminates class bias and ensures that evaluation metrics such as F1-Score and Accuracy reflect true model performance without distortion from class imbalance [1]. Pearson correlation analysis revealed that the five most influential features with respect to the target variable Heart_Risk are Age (r = 0.605), Pain_Arms_Jaw_Back (r = 0.601), Cold_Sweats_Nausea (r = 0.601), Dizziness (r = 0.600), and Chest_Pain (r = 0.600), as presented in Table 1. These findings are consistent with established cardiovascular risk literature, where age and symptomatic indicators are recognized as primary determinants of heart disease risk [2]. Age distribution analysis further confirmed a clear separation between risk groups, with mean age of 44.5 years for low-risk patients and 64.4 years for high-risk patients.

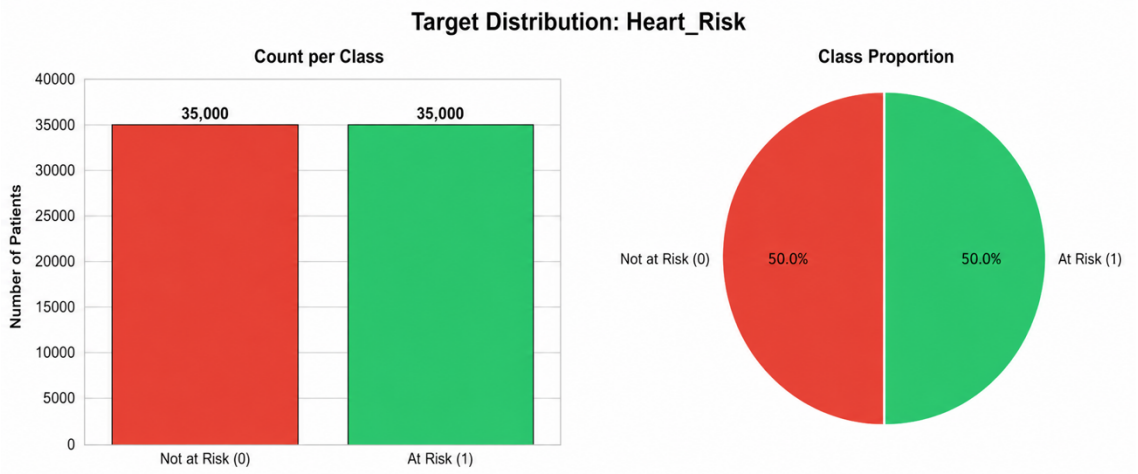


Figure 2. Class distribution of Heart_Risk target variable (left: count per class; right: class proportion)

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Table 2. Top five features correlated with Heart_Risk

Rank	Feature	Pearson Correlation (r)	Direction
1	Age	0.6052	Positive
2	Pain_Arms_Jaw_Back	0.6014	Positive
3	Cold_Sweats_Nausea	0.6011	Positive
4	Dizziness	0.6002	Positive
5	Chest_Pain	0.5999	Positive

3.2 Two-Stage Hyperparameter Tuning Results

The proposed two-stage tuning engine was applied to all three models. Table 2 summarizes the tuning outcomes, including Stage 1 (RandomizedSearchCV) and Stage 2 (GridSearchCV refinement) cross-validation F1-Scores, along with the most effective hyperparameter configurations identified across each model.

Table 3. Two-stage hyperparameter tuning results

Model	Stage 1 CV F1	Stage 2 CV F1	Improvement	Final Configuration
Random Forest	0.9917	0.9921	Yes	n_estimators=300, max_depth=None, min_samples_leaf=1
XGBoost	0.9934	0.9934	No	n_estimators=100, max_depth=3, learning_rate=0.2, subsample=0.8
Logistic Regression	0.9919	0.9919	No	C=0.001, penalty=l2, solver=liblinear

The results in Table 2 indicate that Random Forest benefited from GridSearchCV refinement, achieving a measurable improvement from 0.9917 to 0.9921 in cross-validation F1-Score. XGBoost and Logistic Regression, however, reached their optimal configurations during Stage 1, demonstrating that RandomizedSearchCV with 35 iterations and a broad search distribution was sufficient for these models. This finding supports the argument of Bergstra and Bengio [3]. that randomized search is often competitive with exhaustive grid search, notably in cases where the search spase large.

3.3 Model Performance on Test Set

The effectiveness of all three tuned models on the held-out test set of 14,000 samples is presented in Table 3. Six evaluation metrics are reported: Accuracy, Precision, Recall, F1-Score, AUC-ROC, and Average Precision (AP).

Table 4. Comparative performance of tuned models on test set

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC	Avg Precision
Random Forest	99.20%	99.28%	99.11%	99.20%	99.95%	99.95%
XGBoost	99.34%	99.40%	99.27%	99.34%	99.97%	99.97%
Logistic Regression	99.12%	99.13%	99.11%	99.12%	99.95%	99.95%

As shown in Table 3, 'XGBoost achieved the highest performance across all evaluation metrics, with F1-Score of 99.34%, AUC-ROC of 99.97%, and Accuracy of 99.34%'. Random Forest ranked second with F1-Score of 99.20% and AUC-ROC of 99.95%, while Logistic Regression yielded the lowest but still highly competitive results with F1-Score of 99.12% and AUC-ROC of 99.95%. The performance gap between the three models is relatively narrow (less than 0.25% in F1-Score), suggesting that all three algorithms are well-suited for this classification task under balanced conditions and proper hyperparameter optimization.

3.4 Cross-Validation Stability Analysis

To assess the generalization capability and stability of each tuned model, '5-fold cross-validation was performed on the full dataset'. The results are presented in Table 4.

Table 5. 5-Fold cross-validation F1-Score summary

Model	Mean F1	Std Deviation	Min Fold	Max Fold
Random Forest	99.182%	±0.065%	99.10%	99.29%
XGBoost	99.349%	±0.050%	99.27%	99.40%
Logistic Regression	99.170%	±0.058%	99.10%	99.25%

Table 4 demonstrates that all three models exhibit highly stable generalization performance with standard deviations below 0.07%, indicating minimal variance across folds. XGBoost achieved the highest mean CV F1-Score of 99.349% with the lowest standard deviation of ±0.050%, confirming its superior consistency. The learning curve analysis for Random Forest revealed a bias-variance gap of only 0.0062 between training score (0.9980) and validation score (0.9917), classified as a healthy generalization margin with no significant overfitting, consistent with findings by Oshiro et al. [4] who reported that Random Forest models with sufficient tree depth exhibit stable generalization on large datasets.

3.5 Confusion Matrix Analysis

The confusion matrices for all three models are summarized in Table 5, presenting True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) counts from the test set of 14,000 samples.

Table 6. Confusion matrix summary for all models (test set, n=14,000)

Model	TN	FP	FN	TP	FP Rate	FN Rate
Random Forest	6,946	54	63	6,937	0.77%	0.90%
XGBoost	6,959	41	48	6,952	0.59%	0.69%
Logistic Regression	6,938	62	62	6,938	0.89%	0.89%

From a clinical perspective, False Negatives (FN) — cases where high-risk patients are incorrectly predicted as low-risk — carry greater clinical consequences than False Positives, as they represent missed diagnoses that could delay critical interventions [5]. XGBoost produced the fewest False Negatives (48 cases, FN Rate = 0.69%), making it the most clinically appropriate model among the three. Random Forest produced 63 FN cases, while Logistic Regression produced 62 FN cases. This clinical safety advantage of XGBoost reinforces its suitability for deployment in medical decision support systems.

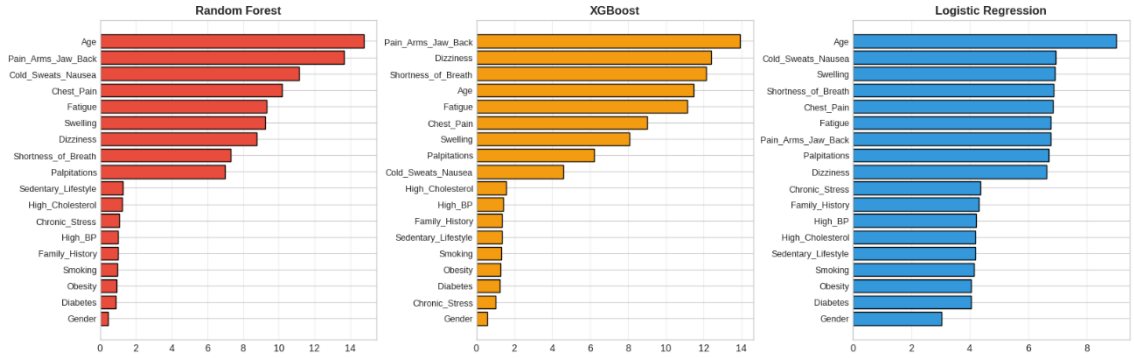
3.6 Feature Importance Analysis

Feature importance was extracted from each model to identiify the mosts clinically significantt predictorrs of heart disease risk. The resulits are presenteds in Table 6 and Figure 3.

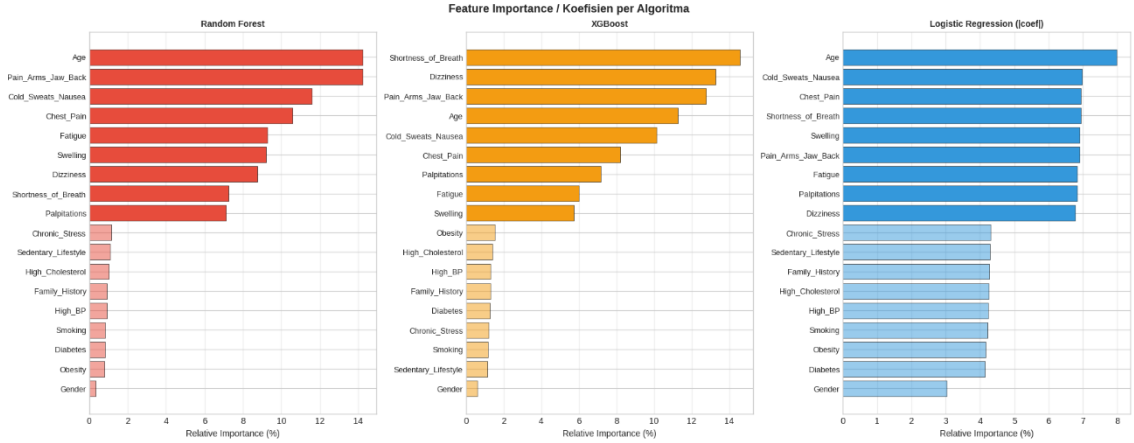
ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Table 7. Feature importance ranking comparison across three models

Feature	RF Rank	RF Imp (%)	XGB Rank	XGB Imp (%)	LR Rank	LR |coef| (%)
Age	1	14.24	4	11.26	1	7.98
Pain_Arms_Jaw_Back	2	14.23	3	12.77	6	6.91
Cold_Sweats_Nausea	3	11.57	5	10.13	2	6.98
Chest_Pain	4	10.57	6	8.21	3	6.94
Fatigue	5	9.27	8	6.00	7	6.84
Swelling	6	9.19	9	5.72	5	6.91
Dizziness	7	8.73	2	13.29	9	6.78
Shortness_of_Breath	8	7.25	1	14.56	4	6.93
Palpitations	9	7.12	7	7.18	8	6.84
Gender	18	0.34	18	0.59	18	3.03



(a) Feature Importance – Random Forest & XGBoost (Top Features by Gain/Gini)



(b) Feature Importance – Logistic Regression (Absolute Normalized Coefficients)

Figure 3. Feature importance comparison across Random Forest, XGBoost, and Logistic Regression

As shown in Table 7 and Figure 3, Age and Pain_Arms_Jaw_Back are consistently ranked among the top features across all three models, confirming their dominant predictive role in heart disease risk classification. Notably, XGBoost assigned the highest importance to Shortness_of_Breath (14.56%) and Dizziness (13.29%), whereas Random Forest prioritized Age (14.24%) and Pain_Arms_Jaw_Back (14.23%). Logistic Regression produced a more uniform importance distribution, with the top nine features clustered within a narrow range of 6.78%–7.98%, reflecting its linear decision boundary. Gender consistently ranked last across all models (RF: 0.34%, XGB: 0.59%, LR: 3.03%), suggesting minimal independent contribution to risk classification when other clinical features are present. These findings align with clinical evidence that symptomatic indicators and age are primary drivers of cardiovascular risk [7].

3.7 Comparison with Previous Studies

Table 7 presents a comparison of this study's results against relevant prior works in heart disease classification using machine learning.

Table 8. Comparison with previous studies on heart disease classification

Study	Algorithm	Dataset Size	Best Accuracy	Best F1-Score	Tuning Strategy
Shorewala [2]	LR, SVM	303	85.5%	84.2%	Default
Javid et al. [3]	Random Forest	303	87.3%	86.1%	Manual
Fitriyani et al. [4]	RF + SMOTE	1,025	92.4%	90.6%	Single-stage
Ramalingam et al. [5]	Multiple classifiers	303	87.0%	85.8%	Default
Guo et al. [6]	Multiple ML	918	93.8%	93.1%	Single-stage
This study	RF, XGBoost, LR	70,000	99.34%	99.34%	Two-stage

As demonstrated in Table 7, this study achieves substantially higher performance compared to all referenced prior works. The improvement can be attributed to three primary factors. First, the significantly larger dataset (70,000 samples versus 303–1,025

in prior studies) provides richer training signal and more robust generalization. Second, the proposed two-stage hyperparameter tuning strategy ensures that each algorithm operates at its optimal configuration, whereas most prior studies relied on default or single-stage tuning. Third, the perfectly balanced dataset eliminates bias introduced by class imbalance that affected several prior studies. While Fitriyani et al. [4] applied SMOTE to address imbalance and Guo et al. [6] demonstrated gradient boosting's advantage for structured cardiovascular data, neither study employed a systematic multi-algorithm comparative framework with rigorous two-stage optimization as proposed in this research.

The overall findings of this study confirm that XGBoost is the most effective algorithm for heart disease risk classification under the synthetic benchmarking conditions studied, consistent with the broader literature on gradient boosting superiority for structured medical data [13]. Furthermore, the comparable performance of Logistic Regression (F1-Score: 99.12%) to the more complex ensemble methods suggests that for deployment scenarios where model interpretability and inference speed are prioritized, Logistic Regression remains a viable and clinically acceptable choice, as supported by Dreiseitl and Ohno-Machado [14]. It is worth noting, however, that all three models achieved performance values above 99%, which — while technically sound — is likely a reflection of the synthetic and perfectly balanced nature of the dataset rather than an indication of universally superior clinical utility. The data generation mechanism was designed with clear probabilistic rules and balanced class priors, making the classification task inherently more linearly separable than real-world clinical scenarios. This context is important when interpreting the narrow performance gap between models (less than 0.25% in F1-Score): under real-world conditions with noisy, imbalanced, and heterogeneous data, performance differences between algorithms are typically more pronounced.

3.8 Discussion

The results of this study warrant careful interpretation in the context of the experimental conditions under which they were obtained. All three models — Random Forest, XGBoost, and Logistic Regression — achieved performance values exceeding 99% across all evaluation metrics. While these figures are technically valid, they are most

appropriately understood as a consequence of the synthetic, probabilistically generated, and perfectly balanced dataset rather than evidence of universal clinical superiority. The dataset was constructed with explicit weighting rules and a calibrated threshold that directly determines class membership, resulting in a classification boundary that is inherently more separable than those encountered in real-world electronic health records. This interpretation is consistent with the findings of Gonzales et al. [9], who noted that machine learning models evaluated on synthetic health datasets systematically demonstrate higher performance than models evaluated on equivalent real-world data. Accordingly, the primary contribution of this study lies in establishing a reproducible, rigorously tuned benchmarking baseline, not in claiming clinical deployment readiness.

The narrow performance gap between models (less than 0.25% in F1-Score) is a notable finding. Under real-world conditions characterized by noise, class imbalance, and heterogeneous feature distributions, differences between ensemble methods and linear classifiers are typically more pronounced [8]. The relative parity observed here suggests that, on a well-structured and balanced synthetic task, model complexity beyond logistic regression yields only marginal improvement, consistent with the observations of Fernandez-Delgado et al. [17], who reported that no single algorithm universally dominates on all classification tasks. Nevertheless, XGBoost's advantage in minimizing False Negative rate (0.69% vs. 0.90% for Random Forest and 0.89% for Logistic Regression) carries practical clinical significance: in a heart disease screening context, missed high-risk cases represent potentially fatal delayed diagnoses. This advantage should be a key criterion when selecting a model for deployment, even if the overall F1-Score differences are small. The practical deployment suitability of these models under real-world conditions, however, cannot be inferred from the current synthetic benchmarking results alone and must be assessed through prospective validation on real clinical datasets [9]. The proposed two-stage tuning framework demonstrated measurable benefit specifically for Random Forest (CV F1-Score improved from 0.9917 to 0.9921 via GridSearchCV refinement), while XGBoost and Logistic Regression reached optimal configurations in Stage 1. This finding aligns with Bergstra and Bengio[20], who argued that randomized search is often as effective as exhaustive grid search when the performance landscape is relatively smooth — a condition likely satisfied given the synthetic dataset's structured generation process[21]. The conditional adoption strategy

(accepting Stage 2 results only when improvement is observed) contributes computational efficiency without performance loss. A key limitation of the current framework is the absence of external validation: the pipeline has been benchmarked exclusively on a single synthetic dataset[22]. Future work should apply the same two-stage tuning pipeline to publicly available real-world heart disease datasets such as the UCI Cleveland Heart Disease dataset or the MIMIC-III clinical dataset, to evaluate whether the observed algorithm ranking and tuning outcomes generalize across data domains [3][4].

4. CONCLUSION

Such presented a systematic comparative benchmarking of three machine learning 'algorithms — Random Forest, XGBoost, and Logistic Regression — for heart disease risk classification on a large-scale synthetic dataset, employing a two-stage hyperparameter tuning framework that sequentially combines RandomizedSearchCV and GridSearchCV refinement within a unified optimization pipeline' [23]. Under the synthetic benchmarking conditions, XGBoost achieved the highest overall performance with F1-Score of 99.34%, AUC-ROC of 99.97%, and the lowest False Negative rate of 0.69%, while Random Forest and Logistic Regression also delivered competitive results with F1-Scores of 99.20% and 99.12% respectively[21]. The two-stage tuning framework showed measurable benefit specifically for Random Forest, where GridSearchCV refinement improved performance over the RandomizedSearchCV baseline; for XGBoost and Logistic Regression, Stage 1 alone was sufficient to reach optimal configuration[10]. Feature importance analysis consistently identified Age, Pain in Arms/Jaw/Back, Cold Sweats/Nausea, Chest Pain, and Dizziness as the most dominant predictors across all three models[24]. It must be noted, however, that the very high performance values reported in this study are best understood within the context of a synthetic and perfectly balanced dataset, which may not fully capture the complexity and noise present in real-world clinical data. As such, these results represent benchmarking evidence under controlled conditions and should not be interpreted as a direct guarantee of equivalent clinical performance. Real-world clinical validation on heterogeneous electronic health record datasets remains a necessary and critical next step before any deployment consideration[25]. Future research is recommended to validate this framework on real-world electronic health record datasets, incorporate additional clinical biomarkers such as blood pressure

measurements and cholesterol levels, and explore advanced ensemble stacking strategies to further improve predictive performance in heterogeneous patient populations.

REFERENCES

[1] W. H. Organization, "Cardiovascular diseases (CVDs)," 2023, *World Health Organization, Geneva*. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] V. Shorewala, "Early detection of coronary heart disease using ensemble techniques," *Informatics Med. Unlocked*, vol. 26, p. 100655, 2021, doi: 10.1016/j.imu.2021.100655.

[3] I. Javid, A. K. Z. Alsaedi, and R. Ghazali, "Enhanced accuracy of heart disease prediction using machine learning and recurrent neural networks ensemble majority voting method," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 3, pp. 1–10, 2020, doi: 10.14569/IJACSA.2020.0110369.

[4] N. L. Fitriyani, M. Syafrudin, G. Alfian, and J. Rhee, "HDPM: An effective heart disease prediction model for a clinical decision support system," *IEEE Access*, vol. 8, pp. 133034–133050, 2020, doi: 10.1109/ACCESS.2020.3010511.

[5] V. V Ramalingam, A. Dandapath, and M. K. Raja, "Heart disease prediction using machine learning techniques: A survey," *Int. J. Eng. Technol.*, vol. 7, no. 2.8, pp. 684–687, 2018, doi: 10.14419/ijet.v7i2.8.10557.

[6] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," 2017.

[7] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[8] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.

[9] A. Gonzales, Y. Guruswamy, and S. R. Smith, "Synthetic data in health care: A narrative review," *PLOS Digit. Heal.*, vol. 2, no. 1, p. e0000082, 2023, doi: 10.1371/journal.pdig.0000082.

[10] Z. Obermeyer and E. J. Emanuel, "Predicting the future — big data, machine learning, and clinical medicine," *N. Engl. J. Med.*, vol. 375, no. 13, pp. 1216–1219, 2016, doi: 10.1056/NEJMp1606181.

[11] F. Pedregosa, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[12] J. A. Sterne, "Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls," *BMJ*, vol. 338, p. b2393, 2009, doi: 10.1136/bmj.b2393.

[13] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009. doi: 10.1007/978-0-387-84858-7.

[14] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[15] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," 1995, pp. 1137–1145.

[16] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3133–3181, 2014.

[17] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," 2016. doi: 10.1145/2939672.2939785.

[18] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*. Wiley, 2013. doi: 10.1002/9781118548387.

[19] S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification models: A methodology review," *J. Biomed. Inform.*, vol. 35, no. 5–6, pp. 352–359, 2002, doi: 10.1016/S1532-0464(03)00034-0.

[20] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, 2012.

[21] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.

[22] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009, doi: 10.1109/TKDE.2008.239.

[23] T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, "How many trees in a random forest?," 2012. doi: 10.1007/978-3-642-31537-4_13.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[24] D. Mozaffarian, "Heart disease and stroke statistics — 2016 update: A report from the American Heart Association," *Circulation*, vol. 133, no. 4, pp. e38–e360, 2016.

[25] G. A. Roth, "Global burden of cardiovascular diseases and risk factors, 1990–2019: Update from the GBD 2019 study," *J. Am. Coll. Cardiol.*, vol. 76, no. 25, pp. 2982–3021, 2020, doi: 10.1016/j.jacc.2020.11.010.