Scientific Journal of Informatics
Vol. 13, No. 1, February 2026
p-ISSN 2407-7658 https://journal.unnes.ac.id/journals/sji/index e-ISSN 2460-0040

Hybrid Feature Selection for Effective Heart Disease Detection: A Multi-Algorithm Machine Learning Approach

Syahrani Lonang[1*], Ahmad Fatoni Dwi Putra[2], Fahmi Syuhada[3], Asno Azzawagama Firdaus[4], Alya Masitha[5]
[1]Department of Information Technology, Universitas Qamarul Huda Badaruddin Bagu, Indonesia
[2, 3, 4]Department of Computer Science, Universitas Qamarul Huda Badaruddin Bagu, Indonesia
[5]Department of Software Engineering, Institut Teknologi Statistika dan Bisnis Muhammadiyah Semarang, Indonesia

Abstract.
Purpose: This research aims to develop an effective early detection model for heart disease with data balancing and hybrid feature selection. The study seeks to enhance predictive accuracy and minimize errors, providing a robust model for clinical decision support systems.
Methods: The study used the Heart Failure Prediction dataset derived from Kaggle. A novel hybrid framework was implemented, integrating SMOTEENN (Synthetic Minority Over-sampling Technique + Edited Nearest Neighbors) for data balancing and a Hybrid Feature Selection (HFS) method combining Chi-square and Backward Elimination. Eight machine learning algorithms, including Logistic Regression, Naïve Bayes, Decision Tree, K Nearest Neighbor, Random Forest, Gradient Boosting, Support Vector Machine, and XGBoost. Performance was assessed based on accuracy, precision, recall, f1-score, specificity, AUC Score, fallout and miss rate.
Result: The proposed framework significantly improved classification performance across all algorithms. The Random Forest model emerged as the optimal classifier, achieving an accuracy of 99.44%, AUC Score of 99.98%, and a specific reduction in miss rate to 0.92% (from 10.03% baseline). The HFS method successfully reduced the feature space by 54%, identifying 'ExerciseAngina', 'FastingBS', 'ST_Slope', 'ChestPainType', and 'Sex' as the most critical predictors. The model outperformed standard approaches and recent state-of-the-art benchmarks by over 10% in accuracy.
Novelty: This study introduces a synergistic integration of SMOTEENN with hybrid feature selection. The combination significantly improves model performance in early heart disease detection.

Keywords: Heart disease, Hybrid feature selection, SMOTEENN, Machine learning, Random forest
Received December 2025 / **Revised** January 2026 / **Accepted** January 2026

INTRODUCTION

Cardiovascular disease (CVD) remains one of the leading global health challenges and continues to be recognized as the foremost cause of mortality worldwide [1]. According to the World Health Organization (WHO), CVD accounts for approximately 32% of all global deaths, claiming nearly 17.9 million lives annually [2][3]. The onset of CVD is influenced by a complex interplay of factors, encompassing congenital abnormalities (such as congenital heart disease) [4], metabolic disorders (e.g., hypertension, diabetes, dyslipidemia) [5], and lifestyle-related risks including smoking, obesity, stress, and excessive alcohol consumption [6]. This burden is not only evident in developed countries but also increasingly affects developing nations, including Indonesia, where CVD-related mortality reaches 274 deaths per 100,000 population, underscoring the urgent need for effective prevention and early diagnostic strategies [7][8].

In the pursuit of improved diagnostic precision, Machine Learning (ML) has become powerful tool in the healthcare sector, particularly for the early detection and diagnosis of complex diseases [9]. Numerous studies have demonstrated the successful application of ML-based classification techniques in identifying medical conditions such as breast cancer [10], diabetes [11], brain tumors [12], and heart and liver disease [13], highlighting its versatility in clinical decision support. In disease diagnosis, patient pathology is often represented by datasets containing multiple features, each exerting a unique influence on the prediction

*Corresponding author.
Email addresses: : lonangsyahrani3@gmail.com (Lonang), ahmadfatoni@uniqhba.ac.id (Putra), fahmisy@uniqhba.ac.id (Syuhada), asnofirdaus@uniqhba.ac.id (Firdaus), alya.masitha@itesa.ac.id (Masitha)
DOI: 10.15294/sji.v13i1.38815

outcome. However, only a subset of these features substantially contributes to accurate classification, while others may introduce noise or redundancy. To address this, feature selection methods are widely employed to identify the most informative attributes, thereby enhancing predictive accuracy and reducing computational cost [14]. Furthermore, medical datasets frequently suffer from class imbalance, where negative cases significantly outnumber positive ones. This imbalance can bias the learning process and degrade model performance. Accordingly, class balancing techniques, such as SMOTEENN (a hybri- approach combining synthetic minority oversampling and edited nearest neighbor undersampling), are critical to ensure fair representation of minority classes and to improve the overall validity of predictive models [15][16].

Recent studies have applied diverse machine learning algorithms to heart disease prediction with encouraging results. Biddinika et al. evaluated Random Forest (RF), Naïve Bayes (NB), Support Vector Machine (SVM), and Decision Tree (DT), reporting that RF outperformed others with an accuracy of 88.41% and an AUC of 0.94 [17]. Kirono and Nataliani analyzed heart failure causes using DT, RF, and XGBoost, where XGBoost achieved up to 94% accuracy with hyperparameter tuning [18]. Similarly, Barry et al. combined Particle Swarm Optimization (PSO) with Modified Random Forest, improving accuracy to 88.52% and AUC to 0.93 [19]. Jusia et al. further demonstrated that optimization significantly enhanced performance, with KNN+PSO achieving 89.09% accuracy and AUC 0.935, while C4.5+PSO reached 86.91% [20]. Comparative analysis by Damayunita et al. showed SVM performing best (92%) compared to KNN (91%) and NB (88%) [21]. Moreover, studies on feature selection and resampling revealed that these preprocessing strategies could raise KNN accuracy to 94% [22]. However, despite these promising results, existing studies exhibit several significant limitations that warrant further investigation.

Despite the encouraging results of previous research, several critical limitations remain unaddressed. First, most existing studies employ either basic feature selection methods or no feature selection at all, potentially leading to model overfitting due to the inclusion of redundant and non-informative features [23]. Second, while some studies acknowledge class imbalance issues, they typically utilize simple oversampling or undersampling techniques, which fail to capture the optimal balance between sample diversity and data representativeness. Recent evidence suggests that hybrid resampling approaches combining oversampling and undersampling strategies, such as SMOTEENN, provide superior performance compared to single- strategy methods [16]. Third, many studies focus on evaluating a limited number of algorithms or assessment metrics, making it difficult to determine which approach is most effective across diverse scenarios. Finally, there is a lack of comprehensive studies that systematically integrate hybrid feature selection with resampling techniques and perform comparative analysis across multiple machine learning algorithms using extensive evaluation metrics for heart disease detection. These gaps highlight the need for a more robust and comprehensive approach to machine learning-based heart disease prediction.

To address these research gaps, this study proposes a comprehensive framework for early heart disease detection by integrating three key innovations: hybrid feature selection combining Chi-square test (for statistical significance) and backward elimination (for optimal subset optimization), SMOTEENN resampling to effectively handle class imbalance while preserving data diversity and representativeness, and systematic evaluation of eight machine learning algorithms (Logistic Regression, Naïve Bayes, K-Nearest Neighbor, Support Vector Machine, Decision Tree, Random Forest, Gradient Boosting, and XGBoost) using comprehensive performance metrics. The specific contributions of this research are: integration of hybrid feature selection (Chi-square + backward elimination) with SMOTEENN resampling, an approach not extensively explored in prior heart disease prediction studies; systematic comparative analysis of eight machine learning algorithms using the UCI Machine Learning heart disease datasets (Cleveland, Hungarian, Switzerland, Long Beach VA, Stalog (Heart) Data Set); comprehensive performance evaluation using nine evaluation metrics (accuracy, precision, recall, F1-score, specificity, fallout, miss-rate, g mean and AUC score) to provide deeper insights into model performance. The objectives of this research are: to develop and validate a hybrid feature selection method that optimizes feature relevance and reduces dimensionality, to assess the effectiveness of SMOTEENN in handling imbalanced medical datasets, to conduct a rigorous comparative analysis of eight machine learning algorithms, and to identify the most effective model configuration for early and accurate heart disease detection that has the potential for real-world clinical implementation.

METHODS

This research follows a systematic machine learning pipeline approach to develop an effective heart disease prediction model. The methodology is structured in six sequential phases: data preprocessing and cleaning, data balancing using SMOTEENN, feature selection using hybrid Chi-square and backward elimination, dataset splitting with stratified k-fold cross-validation, model training and testing with eight classification algorithms, and performance evaluation and model comparison. The workflow is designed to ensure reproducibility, minimize overfitting, and provide robust comparative analysis across multiple machine learning algorithms. The research flow is shown in Figure 1.

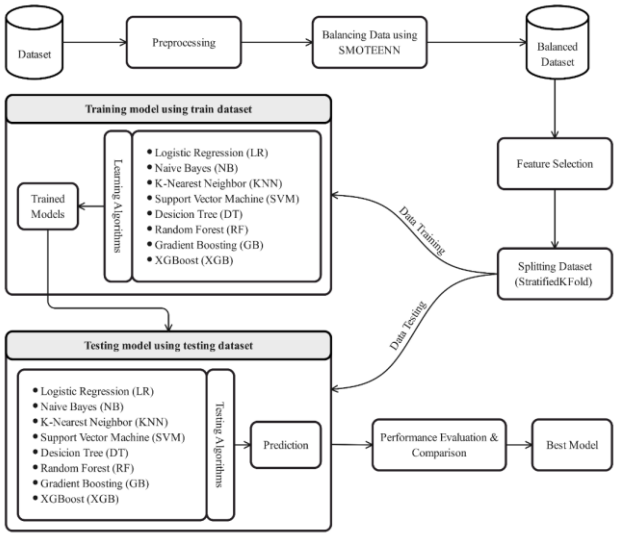

Figure 1. Research flow

Dataset

The data set used in this research was the Heart Failure Prediction Dataset obtained from Kaggle, which represents a comprehensive combination of five publicly available heart disease datasets from the UCI Machine Learning Repository [24]. The dataset comprises 12 attributes, 11 input features and 1 target variable, all of which are clinically relevant indicators of cardiovascular health. The input features including age, sex, chest pain type, resting blood pressure (RestingBP), cholesterol, fasting blood sugar (FastingBS), resting electrocardiogram (RestingECG), maximum heart rate (MaxHR), exercise angina, oldpeak and ST slope. The target variable, HeartDisease, is a binary classification output indicating heart disease (1) or normal (0).

Table 1. Dataset description

Feature	Data type	Detail
Age	Numeric	Representing the patient's age in years
Sex	Binary	Patient gender (1: Male, 0: Female)
ChestPainType	Nominal	Categorizing chest pain into four types (1: Typical Angina, 2: Atypical Angina, 3: Non-Anginal Pain, 4: Asymptomatic)
RestingBP	Numeric	Resting blood pressure in millimeters of mercury (mm Hg)
Cholesterol	Numeric	Amount of cholesterol in patient's blood (mm/dl)
FastingBS	Binary	Fasting blood sugar levels (1: > 120 mg/dl, 0: otherwise)
RestingECG	Nominal	Resting electrocardiogram (0: Normal, 1: Having ST-T wave abnormality, 2: probable or definite left ventricular hypertrophy by Estes' criteria)
MaxHR	Numeric	Maximum heart rate (bpm)
ExerciseAngina	Binary	1: Yes, 0: No
Oldpeak	Numeric	ST depression includes by training to rest
ST_Slope	Nominal	The slope of the peak exercise ST segment in three categories (1: Upsloping, 2: Flat, 3: Downsloping)
Target	Binary	1: Heart diasease, 0: Normal

Data preprocessing

Data preprocessing is a vital stage prior to model training, because the quality and consistency of the input directly affect the reliability of the predictions generated by machine learning models. In this study, the raw Heart Failure Prediction Dataset was first examined for missing values and inconsistent entries. Records with missing or invalid values were identified using descriptive statistics and exploratory data analysis.

After checking missing values, all categorical attributes were transformed into numerical form using label encoding. This step was applied to variables such as Sex, ChestPainType, RestingECG, ExerciseAngina, and ST_Slope. For example, the ChestPainType attribute, originally represented as Typical Angina, Atypical Angina, Non-Anginal Pain, and Asymptomatic, was encoded as 1, 2, 3, and 4, respectively. Label encoding was chosen instead of one-hot encoding to keep the feature space compact and to reduce computational complexity, which is particularly beneficial when training multiple models and performing cross-validation.

Finally, all numerical features were normalized using the min–max normalization method. For each feature x, the normalized value x^* is computed as

$$x^* = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

where $x_{\min}$ and $x_{\max}$ denote the minimum and maximum values of the feature, respectively.

Data balancing using SMOTEENN

In this research, we address the issue of class imbalance using the SMOTEENN method, which combines the Synthetic Minority Over-sampling Technique (SMOTE) with Edited Nearest Neighbors (ENN). Empirical studies have demonstrated that SMOTEENN often outperforms SMOTE alone because it not only increases the representation of the minority class but also cleans the data by removing noisy and borderline samples, leading to clearer decision boundaries [25][26].

SMOTEENN is a hybrid resampling strategy that integrates oversampling and undersampling in a two-step process to enhance the quality of the training data. The first step involves generating synthetic samples for the minority class using SMOTE. This technique interpolates new instances between existing minority samples and their nearest neighbors, effectively increasing the sample size and mitigating overfitting risks associated with simple duplication. However, these synthetic points may inadvertently be placed in overlapping regions or near decision boundaries, potentially confusing the classifier. To resolve this, the second step employs the ENN algorithm as a data cleaning mechanism. ENN evaluates every instance in the dataset including original samples and newly generated synthetic points by checking its k-nearest neighbors. If an instance is misclassified by most of its neighbors, it is identified as noise or a borderline case and is subsequently removed. By eliminating these problematic samples, ENN refines the decision boundaries and ensures a more coherent training set. This combination allows SMOTEENN to balance the class distribution while simultaneously improving data quality, making it particularly effective for datasets with complex or overlapping class structures [26].

Data Leakage Prevention, SMOTEENN was applied exclusively within each training fold during cross-validation, never to test folds. The procedure was: Dataset split into 10 stratified folds. For each fold: - Training set (9 folds) → Apply SMOTEENN → Train model - Test set (1-fold) → Original imbalanced distribution → Evaluate. Report averaged metrics across 10 folds.

Hybrid feature selection: chi-square and backward elimination

Feature selection is a critical preprocessing step that identifies and retains the most informative features while eliminating irrelevant and redundant features. This process improves model interpretability, reduces computational complexity, and enhances predictive accuracy by eliminating noise. This study employs a hybrid feature selection approach that combines two complementary methods: Chi-square test for statistical significance assessment and backward elimination for optimal subset optimization [14][6].

To prevent feature selection bias, feature selection was performed within each cross-validation training fold (90% of data) rather than on the full dataset. This prevents information leakage from test data into feature relevance scoring. The chi-square test was applied to each feature to evaluate statistical association with the target variable. Chi-square scores were computed only on the training fold. Features were ranked by chi-square value, and the top 10 features were selected as candidates for elimination. Selected features were further refined using Sequential Feature Selector (SFS) with backward elimination. The algorithm iteratively removed the feature causing the least decrease in classification accuracy (objective function), until exactly k=5 features remained. The SFS backward elimination was performed using a logistic

regression classifier (max_iter=1000) applied to the training fold data only. The backward elimination process terminated when the feature subset size reached k=5, representing an optimal balance between model interpretability (minimal features) and predictive performance. The selected feature indices from each training fold were applied to the corresponding test fold without modification, ensuring test set independence.

Dataset splitting with stratified k-fold

After feature selection, the balanced and preprocessed dataset is divided into training and testing subsets using Stratified K-Fold cross-validation. Stratified K-Fold is a resampling technique that divides the dataset into k (in this study, k=10) equally sized stratified folds, where each fold maintains the same class distribution (proportion of positive and negative cases) as the original dataset. This stratification is crucial in medical classification tasks to ensure that all folds represent both positive and negative cases appropriately, reducing the variance in performance estimates and preventing bias toward either class [27].

The dataset is divided into 10 equally sized stratified folds. In each iteration, one-fold serves as the test set (10%) while the remaining nine folds form the training set (90%). Stratification ensures each fold maintains the same class distribution as the full dataset, preventing fold bias and ensuring fair performance estimates. Critically, the preprocessing, resampling, and feature selection pipeline operates only within each fold's training data. The test fold is never used during preprocessing, resampling, feature selection, or training. Feature indices selected in the training fold are applied to the test fold without modification or re-selection.

Classification algorithms

This study evaluates eight machine learning classification algorithms to identify the most effective approach for heart disease detection. These algorithms represent diverse learning paradigms including linear models, distance-based models, decision tree-based models, and ensemble methods. The eight algorithms and hyperparameter configurations are described below:

Table 2. Hyperparameter configurations

Algorithms	Hyperparameters
Logistic Regression	solver = 'lbfgs', max_iter = 100, C = 1.0, random_state = 42
Naïve Bayes	-
K-Nearest Neighbor	n_neighbors=7, weights = 'uniform', algorithm = 'auto', metric = ‘euclidean’, p = 2
Decision Tree	criterion='entropy', splitter='best', max_depth = None, min_samples_split = 2, min_samples_leaf = 1, random_state = 42
Random Forest	n_estimators=100, criterion= ‘gini’, max_depth=None, min_samples_split=2, bootstrap = True, random_state=42
Support Vector Machine	kernel = 'rbf', C = 1.0, gamma = 'scale' probability = True, random_state = 42
Gradient Boosting	n_estimators = 100, learning_rate = 0.1, max_depth = 3, min_samples_split = 2, min_samples_leaf = 1, loss = 'log_loss', random_state = 42
XGboost	n_estimators = 100, learning_rate = 0.3, max_depth = 6, gamma = 0 eval_metric = 'logloss', random_state = 42, use_label_encoder = False

Logistic regression (LR)

LR is a linear classification algorithm that models the probability of binary classification using a logistic function. It estimates the relationship between independent features and the binary target variable using the sigmoid function, making it interpretable and computationally efficient. LR is widely used in medical diagnosis due to its simplicity and robustness [23][28].

N**aïve bayes (NB)**

NB is a supervised learning method grounded in Bayes' theorem, widely utilized for classification and predictive tasks. It operates on the assumption that features are conditionally independent given the target class. Known for its computational efficiency, the algorithm scales linearly with the number of attributes, ensuring that training time does not disproportionately impact performance. Additionally, Naïve Bayes supports incremental learning, offers low-variance predictions, and evaluates model accuracy using confusion matrix metrics [29][30].

K-nearest neighbor (KNN)

KNN is one of the most widely used supervised machine learning algorithms due to its ability to solve complex problems [31][32]. KNN's basic working principle is to determine a class based on the class of its nearest neighbours. The KNN algorithm is relatively easy to implement, but it is frequently referred to as a

lazy learning algorithm. because lazy learners postponed the generalization of training data until a task is received by the model [33][34].

Random forest (RF)
RF is a highly efficient supervised learning algorithm widely applied in healthcare to classify disease. It excels at processing feature subsets rapidly and detecting complex, nonlinear relationships, often resulting in minimal prediction errors. Random Forest is an ensemble method that combines multiple decision trees trained on random feature subsets and random data samples. Through ensemble averaging, Random Forest reduces overfitting, improves generalization, and provides robustness against noise. RF can capture complex non-linear relationships and is less sensitive to outliers than individual decision trees [35].

Support vector machine (SVM)
SVM addresses nonlinear problems by mapping data into a high-dimensional space using kernel functions. In this space, the algorithm identifies an optimal hyperplane that maximizes the margin between data classes. Finding this hyperplane, which serves as the best possible separator, is the fundamental objective of the SVM methodology [21] [36].

Decision tree (DT)
DT is a tree-based model that recursively partitions the feature space to create interpretable decision rules. Each node represents a feature, each branch represents a decision outcome, and each leaf represents a class label. Decision trees are highly interpretable, require minimal feature scaling, and provide feature importance rankings valuable for clinical understanding [37].

Gradient boosting (GB)
GB is a versatile machine learning technique applied to both regression and classification tasks. It typically builds predictive models by assembling an ensemble of weak learners, usually decision trees. This method often outperforms Random Forest, particularly in scenarios where individual decision trees struggle to capture underlying patterns. Gradient-boosted models are developed sequentially, like traditional boosting methods, but offer greater flexibility by allowing for the optimization of any differentiable loss function [5].

Extreme gradient boosting (XGB)
XGBoost is an optimized implementation of gradient boosting with enhanced computational efficiency and regularization mechanisms. XGBoost incorporates L1 and L2 regularization, handles missing values automatically, and demonstrates superior performance in many medical classification benchmarks. Recent heart disease studies have highlighted XGBoost as a top performer [38] [39].

Performance evaluation
Comprehensive evaluation of model performance requires multiple metrics that capture different aspects of prediction accuracy. This study employs nine evaluation metrics derived from the confusion matrix to provide holistic assessment of model performance. Each metric reveals different characteristics of model behavior and is particularly relevant for medical diagnosis where both false positives and false negatives carry clinical consequences. The performance metrics used are calculated based on the formula below.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$Specificity = \frac{TN}{TN + FP} \tag{5}$$

$$Fallout = \frac{FP}{FP + TN} \tag{6}$$

$$Missrate = \frac{FN}{FN + TP} \tag{7}$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (8)$$

$$G - Mean = \sqrt{Recall\ x\ Specificity} \quad (9)$$

These metrics become the foundation for evaluating the performance of classifiers, which will be introduced subsequently. Accuracy is the percentage of correct predictions. It measured how well a classified model predicted a condition. Precision is a metric used to evaluate the relevance of the retrieved instances. It is calculated as the ratio of true positive predictions to the set of all true positive values [40]. Recall or often called the true positive rate calculates the percentage of true positives that were correctly predicted. Specificity is the proportion of accurately predicted negative cases to the overall number of true negative cases [41]. F1 score or F measure is precision between recall harmonic mean [42]. Specificity measures the proportion of actual negative cases correctly identified. Fallout indicates false positive rate. Miss Rate indicates false negative rate. G-Mean provides a balanced measure of sensitivity and specificity. and AUC Score represents the area under the ROC curve, indicating the probability that the model ranks a random positive instance higher than a random negative instance.

RESULTS AND DISCUSSIONS

The initial phase of the experiment involved data preprocessing and exploratory data analysis to understand the dataset's underlying structure and prepare it for machine learning. First, the dataset was inspected for missing values and duplicate records, revealing no missing entries or duplicates, which confirmed the high quality of the raw data. Subsequently, categorical variables were transformed using Label Encoding to convert non-numeric attributes into a machine-readable numeric format. This transformation preserves the ordinal information where applicable and ensures compatibility with the chosen classification algorithms without substantially increasing the dataset's dimensionality.

To further investigate the relationships between features and the target variable, a correlation heatmap was shown in Figure 2. The heatmap analysis indicates varying degrees of correlation between input features and the target. Notably, features such as 'ChestPainType', 'ExerciseAngina', 'Oldpeak', and 'ST_Slope' show relatively stronger correlations with the target variable, suggesting their potential importance for classification.

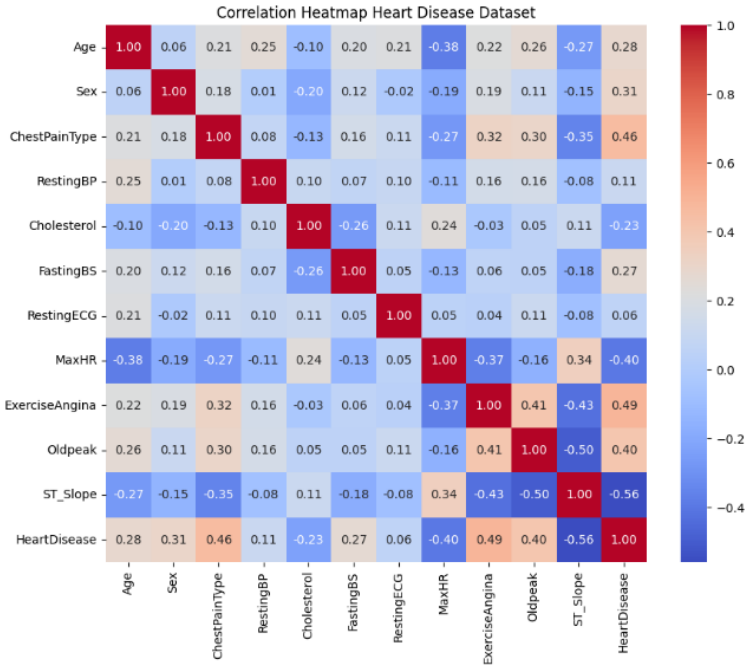

Figure 2. Correlation heatmap

Analysis of the class distribution in the original dataset revealed a relatively balanced overall target distribution, with 410 (44.66%) normal cases and 508 (55.34%) heart disease cases. However, a deeper

demographic breakdown exposed a significant gender imbalance. The dataset is predominantly male, comprising 725 male subjects (456 with heart disease, 267 normal), compared to only 193 female subjects (50 with heart disease, 143 normal). This disparity poses a risk of gender bias in the predictive model, where the algorithm might learn to associate gender more strongly with disease prevalence due to the skewed representation.

To mitigate this issue and improve the model's robustness, the SMOTEENN resampling technique was applied. The application of SMOTEENN resulted in a refined dataset with a total of 696 observations. Post-balancing analysis shows a more equitable distribution: the male population was adjusted to 458 (246 with heart disease, 212 normal), and the female population to 238 (95 with heart disease, 143 normal). This redistribution not only balances the target classes but also significantly reduces gender disparity, ensuring that the model learns from a more representative sample of both genders. The comparison of data distribution before and after SMOTEENN balancing is visualized in Figure 3.

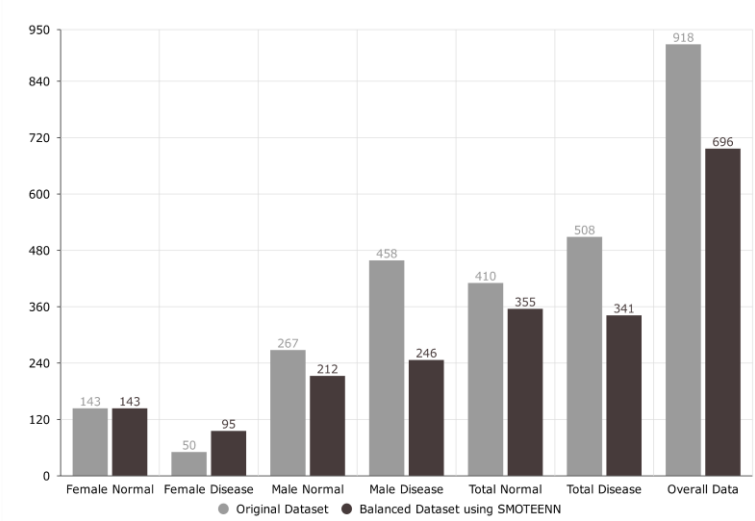

Figure 3. Comparison of class distribution before and after SMOTEENN balancing

Feature selection is a critical step in developing efficient medical diagnostic models, as it removes irrelevant data, reduces computational complexity, and minimizes the risk of overfitting. In this study, a hybrid approach combining the Chi-square test (filter method) and Backward Elimination (wrapper method) was employed to identify the most significant predictors of heart disease.

The first stage utilized the Chi-square statistical test to evaluate the dependency between each input feature and the target variable. Features with higher Chi-square scores exhibit a stronger statistical relationship with heart disease. As presented in Table 3, 'ExerciseAngina' emerged as the most dominant predictor with a score of 224.09, followed significantly by 'FastingBS' (71.67) and 'ST_Slope' (57.50). Conversely, 'RestingBP' obtained a negligible score of 0.13, indicating no significant statistical difference in resting blood pressure between normal and heart disease patients in this dataset. Consequently, 'RestingBP' was discarded, and the top 10 features were retained for the subsequent optimization phase.

Table 3. Top 10 features ranked by chi-square score

Rank	Feature Name	Chi-Square Score
1	ExerciseAngina	224.09
2	FastingBS	71.67
3	ST_Slope	57.50
4	ChestPainType	51.32
5	Sex	23.18
6	MaxHR	10.83
7	Age	8.49
8	Oldpeak	8.02
9	Cholesterol	6.56
10	RestingECG	6.52

Following the initial filtering, Backward Elimination was applied as a wrapper method to further refine the feature subset. This process used Logistic Regression as the base estimator, iteratively removing the least significant feature based on model accuracy. The optimization process is detailed in Table 4. Starting with all 10 features selected by Chi-square, the baseline accuracy was 90.20%. As features with lower contribution were removed, the model's accuracy progressively improved. The highest performance was achieved with a subset of 5 features, yielding an average accuracy of 95.81%. This significant increase demonstrates that features such as 'Cholesterol', 'RestingECG', 'Age', 'MaxHR', and 'Oldpeak', although statistically relevant in the Chi-square test, introduced noise or redundancy when combined in a multi-variate model.

Table 4. Feature subset optimization using backward elimination

Number of Features	Selected Feature Subset	Average Accuracy
10 Features	ExerciseAngina, FastingBS, ST_Slope, ChestPainType, Sex, MaxHR, Age, Oldpeak, Cholesterol, RestingECG	90.20%
9 Features	ExerciseAngina, FastingBS, ST_Slope, ChestPainType, Sex, MaxHR, Age, Oldpeak, Cholesterol	91.18%
8 Features	ExerciseAngina, FastingBS, ST_Slope, ChestPainType, Sex, MaxHR, Age, Oldpeak	92.75%
7 Features	ExerciseAngina, FastingBS, ST_Slope, ChestPainType, Sex, MaxHR, Age	94.37%
6 Features	ExerciseAngina, FastingBS, ST_Slope, ChestPainType, Sex, MaxHR	94.37%
5 Features (Optimal)	ExerciseAngina, FastingBS, ST_Slope, ChestPainType, Sex	95.81%

Based on these results, the final set of 5 optimal features ('ExerciseAngina', 'FastingBS', 'ST_Slope', 'ChestPainType', and 'Sex') was selected for the final model.

To evaluate the effectiveness of the proposed hybrid feature selection (HFS) method, a comprehensive performance comparison was conducted across eight machine learning algorithms: LR, NB, SVM, KNN, DT, RF, GB, and XGBoost. The evaluation utilized stratified 10-fold cross-validation to ensure robust performance estimates. Two experimental scenarios were executed: the first without using feature selection and the second implementing the proposed Hybrid Feature Selection.

The experimental results demonstrate a significant performance improvement across all classifiers when utilizing the Hybrid Feature Selection method. As shown in Table 5, models trained with the selected subset of 5 features consistently outperformed those trained on the full feature set. For instance, the RF algorithm, which achieved the highest overall performance, saw its accuracy increase from 86.32% to 99.44%, and its F1-Score rise from 88.00% to 99.38% after applying HFS. Similarly, XGBoost showed a remarkable improvement, with accuracy jumping from 86.32% to 99.30%. Even simpler models like LR experienced substantial gains, with accuracy improving from 84.64% to 95.81%. This consistent enhancement across metrics such as Precision, Recall, Specificity, and AUC Score confirms that the selected features ('ExerciseAngina', 'FastingBS', 'ST_Slope', 'ChestPainType', 'Sex') hold the most discriminative power for heart disease detection. By removing irrelevant attributes, the models could focus on the most critical patterns, reducing noise and complexity.

Table 5. Performance of classification algorithms with and without hybrid feature selection (mean)

Algorithm	Accuracy	Precision	Recall	F1-Score	Specificity	AUC Score	G-Mean
LR (No HFS)	84.64	85.50	87.22	86.21	81.46	91.19	84.17
LR (HFS)	**95.81**	**95.82**	**95.12**	**95.38**	**96.40**	**99.03**	**95.72**
NB (No HFS)	85.30	87.09	86.64	86.68	83.66	91.36	84.98
NB (HFS)	**93.70**	**92.91**	**93.57**	**93.15**	**93.82**	**98.56**	**93.65**
SVM (No HFS)	86.60	85.72	91.15	88.29	80.98	91.64	85.84
SVM (HFS)	**96.93**	**95.39**	**98.17**	**96.71**	**95.89**	**99.52**	**97.00**
KNN (No HFS)	85.12	85.06	89.08	86.86	80.24	90.34	84.44
KNN (HFS)	**97.07**	**96.18**	**97.56**	**96.83**	**96.66**	**99.57**	**97.09**
DT (No HFS)	79.11	81.75	80.33	80.88	77.32	78.82	78.71
DT (HFS)	**99.30**	**99.41**	**99.08**	**99.24**	**99.48**	**99.28**	**99.28**
RF (No HFS)	86.32	86.27	89.97	88.00	81.95	92.40	85.78
RF (HFS)	**99.44**	**99.71**	**99.08**	**99.38**	**99.74**	**99.98**	**99.41**
GB (No HFS)	87.89	87.88	90.76	89.26	84.15	93.36	87.35
GB (HFS)	**98.75**	**98.25**	**99.08**	**98.65**	**98.45**	**99.81**	**98.76**
XGB (No HFS)	86.32	86.32	89.73	87.94	82.20	92.14	85.79
XGB (HFS)	**99.30**	**99.70**	**98.78**	**99.23**	**99.74**	**99.86**	**99.26**

The proposed HFS method achieved exceptional performance in both dimensions: across all algorithms with HFS, Recall values exceed 93.57% (minimum for NB) and reach 99.08% for multiple algorithms (DT, RF, GB), while Specificity values exceed 93.82% (NB) and reach 99.74% (RF). This balanced excellence in both metrics reflected in G-Mean values consistently exceeding 93.65% demonstrates that the method minimizes both Type I and Type II errors simultaneously, a rare achievement that justifies its clinical applicability. The high Precision values (minimum 92.91% for NB, maximum 99.71% for RF with HFS) further strengthen clinical confidence. When the model predicts "Heart Disease," clinicians can be highly confident that this prediction is reliable. Combined with high Recall, this creates a virtuous diagnostic scenario: the model identifies nearly all positive cases (high Recall) while maintaining low false positive rates (high Precision), enabling confident clinical decision-making. The visual comparison of these performance gains is illustrated in Figure 4 (a) (b).

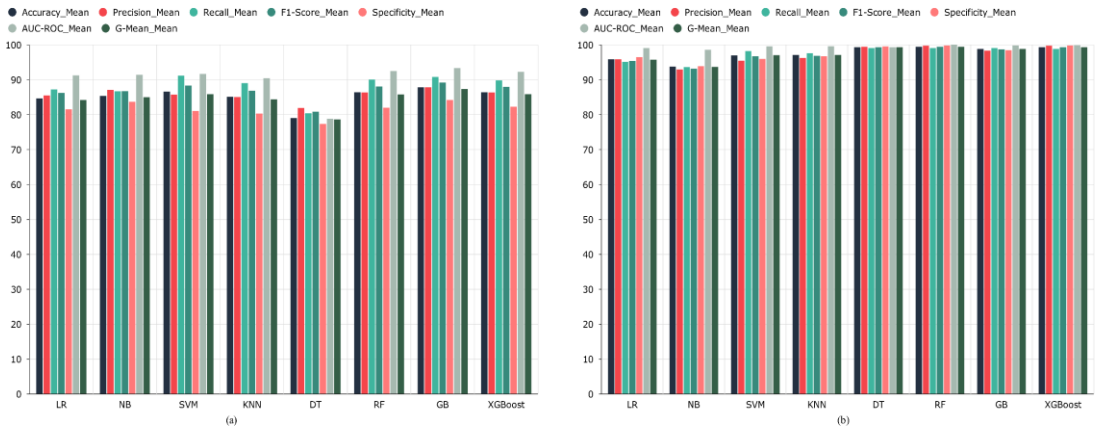

Figure 4: (a) Average performance of classification algorithms without hybrid feature selection, (b) Average performance of classification algorithms using hybrid feature selection

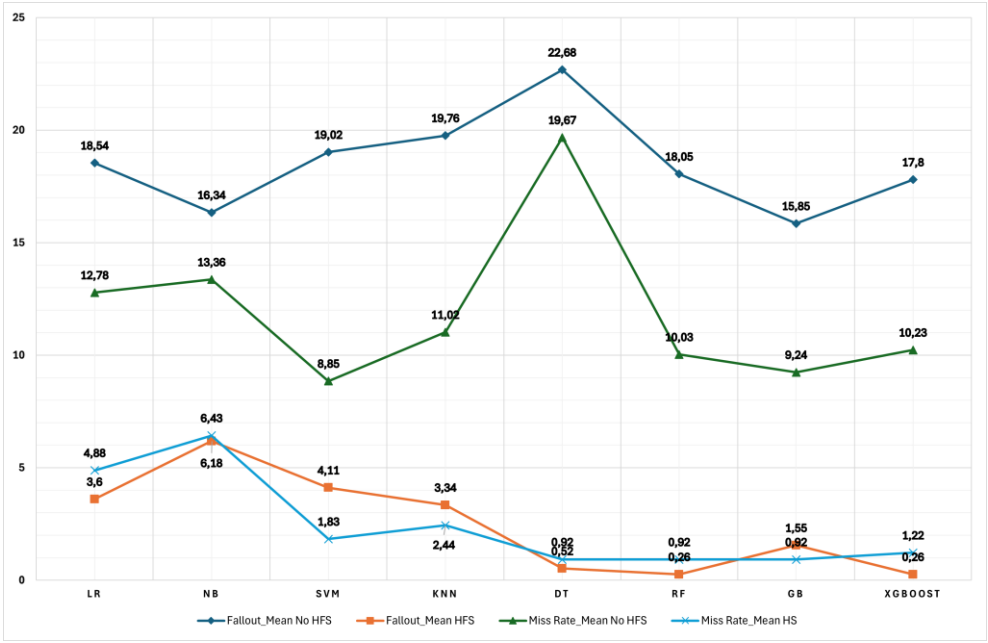

Figure 5. Comparison error (fallout and miss rate) with and without hybrid feature selection

While high accuracy and AUC scores indicate overall model performance, the clinical significance of specific error types of Fallout (False Positive Rate) and Miss Rate (False Negative Rate) warrants detailed analysis. A false negative (missing a heart disease patient) can delay critical treatment and potentially result in serious health complications or mortality. Conversely, a false positive (incorrectly diagnosing a healthy

person as having heart disease) may lead to unnecessary stress, additional testing, and overtreatment. Figure 5 presents the comparison of these critical error rates before and after HFS. The reduction in Miss Rate (False Negative Rate) is particularly noteworthy. Without HFS, Random Forest achieved a 10.03% miss rate; with HFS, this decreased to merely 0.92% a 91.7% reduction. Similarly, DT improved from 19.67% to 0.92% (95.3% reduction), and XGBoost from 10.23% to 1.22% (88.1% reduction). These dramatic reductions mean that the proposed method successfully identifies nearly all patients with heart disease. The Fallout reductions are equally impressive: RF achieved a 0.26% false positive rate (down from 18.05%, a 98.6% reduction), and XGBoost achieved 0.26% (down from 17.80%, a 98.5% reduction). These substantially reduced error rates provide compelling evidence that the hybrid feature selection approach effectively minimizes critical diagnostic errors on the UCI dataset. While these cross-validated results are encouraging, clinical translation requires external validation on independent patient populations with diverse demographic and clinical characteristics. Future prospective studies are essential to establish whether these error reduction benefits generalize to real-world clinical deployment scenarios, where population heterogeneity and data quality variations may differ from the current dataset.

Comprehensive analysis of the experimental results reveals Random Forest (RF) with Hybrid Feature Selection as the superior model for heart disease detection in this research context. RF achieved the highest AUC score of 99.98%, a near-perfect probabilistic ranking that indicates exceptional ability to distinguish between healthy and disease cases across all decision thresholds. It attained 99.44% Accuracy, 99.71% Precision, 99.08% Recall, 99.38% F1-Score, 99.74% Specificity, and 99.41% G-Mean. Most critically, RF achieved a Miss Rate of only 0.92% and a Fallout of 0.26%, ensuring that virtually all heart disease patients are correctly identified while minimizing unnecessary interventions for healthy individuals. RF was selected based on the following rationale: Highest Specificity (99.74%) ensures minimal false alarms; Optimal AUC Score (99.98%) provides maximal discriminative power; Superior Precision (99.71%) reduces patient anxiety from false positives.

To validate the significance of the proposed method, we benchmarked our results against recent related works discussed in the introduction. The effectiveness of the Hybrid Feature Selection combined with SMOTEENN is evident when comparing the obtained performance metrics with existing studies. While Biddinika et al. [17] and Barry et al. [19] achieved accuracies of 88.41% and 88.52% respectively using Random Forest variants, our optimized Random Forest model achieved a remarkable 99.44% accuracy. This represents a substantial improvement of over 10%, indicating that standard Random Forest models significantly benefit from the proposed preprocessing pipeline. Furthermore, our model surpassed the high-performance baselines established by Kirono and Nataliani [18] (94% with XGBoost) and Damayunita et al. [21] (92% with SVM). Even compared to optimized feature selection studies that reached 94% accuracy [22], our approach demonstrated a clear superiority. Jusia et al. [20] reported 89.09% with KNN+PSO, whereas our KNN model with Hybrid Feature Selection achieved 97.07%.

CONCLUSION

This study successfully integrated SMOTEENN for data balancing and a Hybrid Feature Selection (HFS) method combining Chi-square ranking with Backward Elimination. The experimental results conclusively identify the Random Forest (RF) algorithm, optimized with the proposed hybrid approach, as the superior model. It achieved an exceptional Accuracy of 99.44% and an AUC-ROC of 99.98%, representing a substantial improvement over the baseline accuracy of 86.32%. The feature selection process effectively reduced the input space from 11 to 5 critical features proving that a compact subset of clinically relevant variables yields higher diagnostic precision than the comprehensive set. Furthermore, the model demonstrated a profound reduction in critical error rates, lowering the Miss Rate (False Negative Rate) to a mere 0.92%, which is paramount for ensuring patient safety in medical diagnostics. Future research should focus on validating this framework on larger, multi-center datasets to assess its cross-population generalizability. Additionally, developing a real-time clinical decision support system (CDSS) or a mobile application based on this optimized Random Forest model would bridge the gap between theoretical research and practical healthcare application, ultimately contributing to earlier detection and better management of cardiovascular diseases.

ACKNOWLEDGMENT

This research was funded by the Ministry of Higher Education, Science, and Technology through the Directorate of Research and Community Service (DPPM) in the Bima Beginner Lecturer Research Scheme (PDP) program in 2025.

REFERENCES

[1] T. Ullah *et al.*, "Machine Learning-Based Cardiovascular Disease Detection Using Optimal Feature Selection," *IEEE Access*, vol. 12, no. December 2023, pp. 16431–16446, 2024, doi: 10.1109/ACCESS.2024.3359910.

[2] World Health Organization, "Cardiovascular Diseases (CVDs)," *World Health Organization*, 2021.

[3] S. Abbas *et al.*, "Artificial intelligence framework for heart disease classification from audio signals," *Sci. Rep.*, vol. 14, no. 1, p. 3123, Feb. 2024, doi: 10.1038/s41598-024-53778-7.

[4] K. Khan, F. Ullah, I. Syed, and H. Ali, "Accurately assessing congenital heart disease using artificial intelligence," *PeerJ Comput. Sci.*, vol. 10, pp. 1–43, 2024, doi: 10.7717/peerj-cs.2535.

[5] S. L. J M and S. P, "Unveiling the potential of machine learning approaches in predicting the emergence of stroke at its onset: a predicting framework," *Sci. Rep.*, vol. 14, no. 1, p. 20053, Aug. 2024, doi: 10.1038/s41598-024-70354-1.

[6] A. K. Gárate-Escamila, A. Hajjam El Hassani, and E. Andrès, "Classification models for heart disease prediction using feature selection and PCA," *Informatics Med. Unlocked*, vol. 19, p. 100330, 2020, doi: 10.1016/j.imu.2020.100330.

[7] Institute for Health Metrics and Evaluation, "Global Burden of Disease Compare," 2021.

[8] V. Chang, V. R. Bhavani, A. Q. Xu, and M. Hossain, "An artificial intelligence model for heart disease detection using machine learning algorithms," *Healthc. Anal.*, vol. 2, p. 100016, Nov. 2022, doi: 10.1016/j.health.2022.100016.

[9] R. Williams, T. Shongwe, A. N. Hasan, and V. Rameshar, "Heart Disease Prediction using Machine Learning Techniques," in *2021 International Conference on Data Analytics for Business and Industry, ICDABI 2021*, Oct. 2021, pp. 118–123. doi: 10.1109/ICDABI53623.2021.9655783.

[10] Z. Mushtaq, A. Yaqub, S. Sani, and A. Khalid, "Effective K-nearest neighbor classifications for Wisconsin breast cancer data sets," *J. Chinese Inst. Eng. Trans. Chinese Inst. Eng. A*, vol. 43, no. 1, pp. 80–92, 2020, doi: 10.1080/02533839.2019.1676658.

[11] L. A. Al Hak, "Diabetes Prediction Using Binary Grey Wolf Optimization and Decision Tree," *Int. J. Comput.*, vol. 21, no. 4, pp. 489–494, 2022, doi: 10.47839/ijc.21.4.2785.

[12] A. Muis, S. Sunardi, and A. Yudhana, "CNN-based Approach for Enhancing Brain Tumor Image Classification Accuracy," *Int. J. Eng. Trans. B Appl.*, vol. 37, no. 05, pp. 984–996, 2024.

[13] M. P. Behera, A. Sarangi, D. Mishra, and S. K. Sarangi, "A Hybrid Machine Learning algorithm for Heart and Liver Disease Prediction Using Modified Particle Swarm Optimization with Support Vector Machine," *Procedia Comput. Sci.*, vol. 218, no. 2022, pp. 818–827, 2022, doi: 10.1016/j.procs.2023.01.062.

[14] H. A. Al-Alshaikh *et al.*, "Comprehensive evaluation and performance analysis of machine learning in heart disease prediction," *Sci. Rep.*, vol. 14, no. 1, pp. 1–15, 2024, doi: 10.1038/s41598-024-58489-7.

[15] J. Yang and J. Guan, "A Heart Disease Prediction Model Based on Feature Optimization and Smote-Xgboost Algorithm," *Information*, vol. 13, no. 475, 2022, doi: 10.3390/info13100475.

[16] T. Wongvorachan, S. He, and O. Bulut, "A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining," *Inf.*, vol. 14, no. 1, 2023, doi: 10.3390/info14010054.

[17] M. K. Biddinika, A. Masitha, H. Herman, and V. A. N. Fatimah, "Machine Learning Techniques for Heart Disease Prediction Using a Multi-Algorithm Approach," *JUITA J. Inform.*, vol. 12, no. 2, p. 149, Nov. 2024, doi: 10.30595/juita.v12i2.24153.

[18] A. S. Kirono and Y. Nataliani, "Perbandingan Algoritma Machine Learning dalam Analisis Penyebab Penyakit Gagal Jantung," *J. Edukasi dan Penelit. Inform.*, vol. 10, no. 2, pp. 296–301, 2024, doi: 10.26418/jp.v10i2.

[19] K. A. Barry, Y. Manzali, R. Flouchi, and M. Elfar, "Heart disease approach using modified random forest and particle swarm optimization," *IAES Int. J. Artif. Intell.*, vol. 14, no. 2, pp. 1242–1251, 2025, doi: 10.11591/ijai.v14.i2.pp1242-1251.

[20] P. A. Jusia, A. Rahim, H. Yani, and J. Jasmir, "Improving Performance of KNN and C4.5 using Particle Swarm Optimization in Classification of Heart Diseases," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 8, no. 3, pp. 333–339, 2024, doi: 10.29207/resti.v8i3.5710.

[21] A. Damayunita, R. S. Fuadi, and C. Juliane, "Comparative Analysis of Naive Bayes , K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) Algorithms for Classification of Heart Disease Patients," *JOIN (Jurnal Online Inform.*, vol. 7, no. 2, pp. 219–225, 2022, doi: 10.15575/join.v7i2.919.

[22] I. G. A. Gunadi and D. O. Rachmawati, "A Comparative Study on the Impact of Feature Selection

and Dataset Resampling on the Performance of the K-Nearest Neighbors (KNN) Classification Algorithm," *J. Nas. Pendidik. Tek. Inform.*, vol. 13, no. 2, pp. 419–427, 2024, doi: 10.23887/janapati.v13i2.82174.

[23] M. A. Islam, M. Z. H. Majumder, M. S. Miah, and S. Jannaty, "Precision healthcare: A deep dive into machine learning algorithms and feature selection strategies for accurate heart disease prediction," *Comput. Biol. Med.*, vol. 176, no. April, p. 108432, 2024, doi: 10.1016/j.compbiomed.2024.108432.

[24] UCI Machine Learning Repository, "Heart Disease Data Set," *1989*.

[25] J. Srivastava and A. Sharan, "A novel phishing website classification method based on hybrid sampling ABSTRACT," *J. Cyber Secur. Technol.*, vol. 8, no. 1, pp. 1–30, 2024, doi: 10.1080/23742917.2023.2240606.

[26] G. Husain *et al.*, "SMOTE vs. SMOTEENN: A Study on the Performance of Resampling Algorithms for Addressing Class Imbalance in Regression Models," *Algorithms*, vol. 18, no. 1, pp. 1–16, 2025, doi: 10.3390/a18010037.

[27] Ichwan Dwi Nugraha, Triando Hamonangan Saragih, Irwan Budiman, Dwi Kartini, Fatma Indriani, and W. Caesarendra, "Performance Comparison of Extreme Learning Machine (ELM) and Hierarchical Extreme Learning Machine (H-ELM) Methods for Heart Failure Classification on Clinical Health Datasets," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 7, no. 3, pp. 713–728, May 2025, doi: 10.35882/jeeemi.v7i3.904.

[28] M. A. Bouqentar *et al.*, "Early heart disease prediction using feature engineering and machine learning algorithms," *Heliyon*, vol. 10, no. 19, p. e38731, Oct. 2024, doi: 10.1016/j.heliyon.2024.e38731.

[29] R. K. Saroj, P. K. Yadav, R. Singh, and O. N. Chilyabanyama, "Machine learning algorithms for understanding the determinants of under-five mortality," *BioData Min.*, vol. 15, no. 20, pp. 1–22, 2022, doi: 10.1186/s13040-022-00308-8 BioData.

[30] K. Kusrini, E. T. Luthfi, M. Muqorobin, and R. W. Abdullah, "Comparison of naive bayes and K-NN method on tuition fee payment overdue prediction," *2019 4th Int. Conf. Inf. Technol. Inf. Syst. Electr. Eng. ICITISEE 2019*, vol. 6, pp. 125–130, 2019, doi: 10.1109/ICITISEE48480.2019.9003782.

[31] H. Abbad Ur Rehman, C. Y. Lin, and Z. Mushtaq, "Effective K-Nearest Neighbor Algorithms Performance Analysis of Thyroid Disease," *J. Chinese Inst. Eng. Trans. Chinese Inst. Eng. A*, vol. 44, no. 1, pp. 77–87, 2021, doi: 10.1080/02533839.2020.1831967.

[32] A. Yudhana, R. Umar, and S. Saputra, "Fish Freshness Identification Using Machine Learning: Performance Comparison of k-NN and Naïve Bayes Classifier," *J. Comput. Sci. Eng.*, vol. 16, no. 3, pp. 153–164, 2022, doi: 10.5626/JCSE.2022.16.3.153.

[33] P. Cunningham and S. J. Delany, "k-Nearest Neighbour Classifiers - A Tutorial," *ACM Comput. Surv.*, vol. 54, no. 6, pp. 1–25, Jul. 2022, doi: 10.1145/3459665.

[34] J. Maillo, S. Ramírez, I. Triguero, and F. Herrera, "kNN-IS: An Iterative Spark-based design of the k-Nearest Neighbors classifier for big data," *Knowledge-Based Syst.*, vol. 117, pp. 3–15, Feb. 2017, doi: 10.1016/j.knosys.2016.06.012.

[35] P. R. Togatorop, M. Sianturi, D. Simamora, and D. Silaen, "Optimizing Random Forest using Genetic Algorithm for Heart Disease Classification," *Lontar Komput. J. Ilm. Teknol. Inf.*, vol. 13, no. 1, p. 60, 2022, doi: 10.24843/lkjiti.2022.v13.i01.p06.

[36] S. M. Ganie and M. B. Malik, "An ensemble Machine Learning approach for predicting Type-II diabetes mellitus based on lifestyle indicators," *Healthc. Anal.*, vol. 2, no. January, p. 100092, 2022, doi: 10.1016/j.health.2022.100092.

[37] T. A. Assegie, K. K. Napa, T. Thulasi, A. K. Kumar, M. J. T. V. Priya, and V. Dhamodaran, "Scalability and performance of decision tree for cardiovascular disease prediction," *IAES Int. J. Artif. Intell.*, vol. 13, no. 3, pp. 2540–2545, 2024, doi: 10.11591/ijai.v13.i3.pp2540-2545.

[38] K. Aprianto and M. D. Anasanti, "Classifying Heart Disease through Fusion of Multi-Source Datasets: Integration of Feature Selection and Explainable Machine Learning Techniques," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 19, no. 3, pp. 247–258, Jul. 2025, doi: 10.22146/ijccs.92395.

[39] H. F. El-Sofany, "Predicting Heart Diseases Using Machine Learning and Different Data Classification Techniques," *IEEE Access*, vol. 12, no. July, pp. 106146–106160, 2024, doi: 10.1109/ACCESS.2024.3437181.

[40] J. Y. Ho *et al.*, "Towards a time and cost effective approach to water quality index class prediction," *J. Hydrol.*, vol. 575, pp. 148–165, 2019, doi: 10.1016/j.jhydrol.2019.05.016.

[41] A. Tharwat, “Classification assessment methods,” *Appl. Comput. Informatics*, vol. 17, no. 1, pp. 168–192, Jan. 2021, doi: 10.1016/j.aci.2018.08.003.
[42] W. M. Shaban, A. H. Rabie, A. I. Saleh, and M. A. Abo-Elsoud, “A new COVID-19 Patients Detection Strategy (CPDS) based on hybrid feature selection and enhanced KNN classifier,” *Knowledge-Based Syst.*, vol. 205, 2020, doi: 10.1016/j.knosys.2020.106270.