Performance Comparison of Tree-Based Models for Heart Disease Prediction Using Feature Selection and SMOTE

Santi[1*], Ema Utami[2]

[1,2] *Digital Transformation Intelligence, Universitas Amikom Yogyakarta, Indonesia*

*corr-author: santi@students.amikom.ac.id

Abstract - Heart disease remains the leading cause of mortality worldwide, highlighting the need for accurate early prediction models. This study proposes a machine learning framework for heart disease prediction using the BRFSS 2015 Heart Disease Health Indicators dataset, which contains 253,680 records and 22 attributes. The proposed approach integrates Synthetic Minority Oversampling Technique (SMOTE) for class imbalance handling, mutual information-based SelectKBest feature selection (k = 15), and three tree-based classifiers: Decision Tree, Random Forest, and XGBoost. A leakage-free preprocessing pipeline was implemented to ensure that SMOTE was applied only to the training data, and classification threshold optimization was performed to improve minority class detection. Model performance was evaluated using Accuracy, Precision, Recall, F1-score, and ROC-AUC metrics. Experimental results show that XGBoost achieved the best performance with a cross-validation ROC-AUC of 0.9815 and a test ROC-AUC of 0.8444 at an optimized threshold of 0.20. The findings demonstrate that the proposed integration of oversampling, feature selection, and threshold optimization can improve predictive performance for imbalanced cardiovascular risk data, providing a practical foundation for machine learning–based decision support in early heart disease risk screening.

Keywords: Heart disease prediction, machine learning, ensemble learning, decision tree, SMOTE.

I. INTRODUCTION

The digital revolution in healthcare demands a paradigm transformation in cardiovascular disease diagnosis from reactive approaches toward more precise predictive methods. Cardiovascular disease represents a global health threat with a mortality rate of 17.9 million lives annually, accounting for 32% of total deaths worldwide. More than 75% of these deaths are caused by heart attacks and strokes triggered by unhealthy lifestyle factors such as poor dietary patterns, insufficient physical activity, smoking, and excessive alcohol consumption [1]. The domino effect of such lifestyles generates secondary risk factors including hypertension, elevated glucose and cholesterol levels, obesity, and metabolic syndrome. The complexity of interactions among risk factors involving non-linear relationships and hidden patterns causes conventional diagnostic methods to face limitations in prediction accuracy and efficiency.

Machine learning technology offers solutions for analyzing complex high-dimensional datasets, identifying hidden patterns, and automating diagnostic processes. Previous research has demonstrated that machine learning-based diagnostic systems provide economical, flexible, and effective approaches in improving prediction accuracy compared to traditional methods [2,3]. Various algorithms have been applied in heart disease prediction with different characteristics. Support Vector Machine (SVM) handles high-dimensional data but requires extensive computational time, Naive Bayes offers high speed but performs suboptimally for complex data, while Logistic Regression provides good interpretability but is limited in handling non-linear relationships [4].

Decision Tree-based approaches possess interpretability and visualization capabilities that assist medical personnel in understanding decision-making processes, yet are susceptible to overfitting. Ensemble techniques such as Random Forest and XGBoost overcome these limitations by combining the strengths of multiple models into stronger and more stable models [5,6,7]. Random Forest employs a *bagging* (*bootstrap aggregating*) approach that constructs multiple decision trees in parallel from different data subsets, highly effective in reducing variance and improving prediction stability on datasets with high noise [5]. Conversely, XGBoost implements *gradient boosting* that constructs decision trees sequentially, where each new tree focuses on correcting prediction errors from previous trees, excelling in reducing bias and achieving high prediction accuracy on complex data patterns [12].

The selection of these two algorithms is based on their capability to handle class imbalance, a critical challenge in the BRFSS 2015 dataset where positive patients represent only approximately 9% of the total population. XGBoost possesses the *scale_pos_weight* parameter specifically designed to assign higher weights to minority classes, while Random Forest offers *class_weight* and bootstrap sampling mechanisms adjustable for imbalanced datasets. Both algorithms provide complementary feature importance extraction mechanisms: Random Forest calculates importance based on mean decrease impurity (MDI), whereas XGBoost utilizes gain-based importance. Comparison of these methods provides cross-validation of features that genuinely influence heart disease prediction, enhancing medical practitioners' confidence in analytical results.

Model performance optimization requires integration of hyperparameter tuning and feature selection to identify relevant variables [8]. The challenge of data imbalance can cause model bias, and the Synthetic Minority Oversampling Technique (SMOTE) has the potential to balance class distribution so that models learn optimally from all classes [9]. The BRFSS 2015 dataset was selected as the empirical foundation because it represents the largest health surveillance system in the United States managed by the Centers for Disease Control and Prevention (CDC) with internationally standardized methodology. This dataset encompasses more than 400,000 respondents with 22 comprehensive features covering demographic aspects, lifestyle, health conditions, and medical history [11].

Previous research in heart disease prediction generally explored algorithms separately or combined techniques without a systematic approach that integrates all optimization aspects. Several studies focused on performance comparison without considering interpretability, while others produced "*black box*" models difficult for healthcare personnel to understand. This research presents four unique integrated contributions. First, developing an integrated framework that systematically combines ensemble learning (Random Forest and XGBoost), embedded feature selection, SMOTE, and hyperparameter tuning on the BRFSS 2015 dataset. Second, implementing an interpretability framework through feature importance extraction and Shapley Additive exPlanations (SHAP) analysis to bridge the gap between clinical needs and AI model complexity. Third, conducting comprehensive exploration of SMOTE's impact on ensemble model performance in large-scale public health datasets with extreme imbalance. Fourth, utilizing the large-scale BRFSS 2015 dataset (400,000+ samples, 22 features) that presents more realistic computational challenges, enabling more robust evaluation of model generalizability.

The research results are expected to provide theoretical contributions in the form of a prediction framework that integrates accuracy and interpretability, as well as practical contributions in the form.

II. METHOD

The research flow is systematically designed to develop a heart disease prediction model based on ensemble tree-based algorithms with feature selection optimization. Research stages include data collection, preprocessing, train test splitting, SMOTE-based class balancing, feature selection, modeling with hyperparameter tuning, and evaluation, as illustrated in Fig. 1.

The study utilizes the Heart Disease Health Indicators Dataset (BRFSS 2015) secondary dataset from Kaggle. This dataset is managed by the Centers for Disease Control and Prevention (CDC) of the United States and contains more than 253,680 adult respondent data with 22 attributes relevant to heart disease risk factors, such as high blood pressure, cholesterol, diabetes, body mass index, and physical activity. This dataset was selected due to its broad coverage, guaranteed quality, and data characteristics that naturally reflect imbalanced conditions between individuals with and without heart disease, making it suitable for ensemble algorithm testing.

A. Dataset Splitting

The dataset used in this study was obtained from the Behavioral Risk Factor Surveillance System (BRFSS) 2015, consisting of 253,680 records with 22 attributes. After removing the target variable, 21 features were retained for modeling. No missing values were found in the dataset. To prevent data leakage, the dataset was first divided into training (80%) and testing (20%) sets using stratified sampling. Synthetic Minority Over-sampling Technique (SMOTE) was then applied exclusively to the training set to address class imbalance, increasing minority class samples from 19,114 to 183,830. The testing set remained in its original imbalanced state and was used solely for final model evaluation.

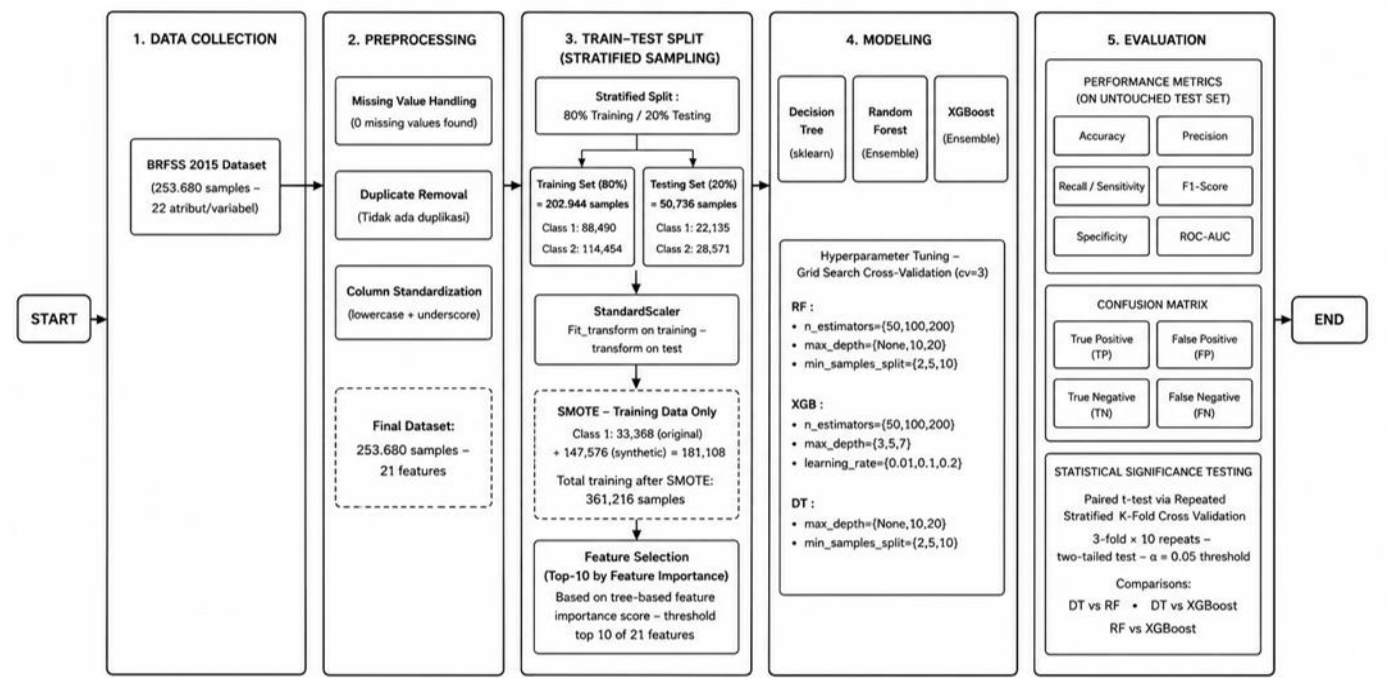

Fig. 1 Proposed machine learning pipeline for heart disease prediction

B. Handling Data Imbalance with SMOTE

The original dataset exhibited significant class imbalance, with the majority class (no disease) substantially outnumbering the minority class (disease present). Such imbalance may introduce bias toward the majority class, thereby reducing the model's sensitivity in detecting minority class instances, which are of greater clinical significance in heart disease prediction. To address this issue, the Synthetic Minority Oversampling Technique (SMOTE) was applied exclusively to the training set following the train-test split, in order to prevent data leakage. SMOTE generates synthetic minority class samples by interpolating between existing nearest-neighbor minority class instances, thereby increasing minority class representation without discarding any majority class data. The testing set was intentionally left untouched to ensure an objective evaluation under real-world conditions.

C. Feature Selection

The next stage is feature selection, which aims to select the most relevant attributes with significant influence on prediction results. Feature selection is performed using an *embedded feature selection* approach with two decision tree-based algorithms: XGBoost and Random Forest. This method can evaluate each feature's importance directly during the model training process. Feature selection is based on the consideration that these attributes are most relevant to health conditions and heart disease risk indicators. By filtering only important features, the model training process becomes more efficient, reduces complexity, and improves the model's generalization capability in predicting patient health conditions.

D. Model Development

The third stage is modeling. In this stage, three main algorithms are used: Decision Tree as the baseline model, and Random Forest and XGBoost as representations of *bagging* and *boosting*-based ensemble techniques. The Decision Tree model is used for initial interpretation and analysis of main risk factors, while Random Forest and XGBoost are used to improve accuracy and prediction stability through decision tree combinations.

Each model is optimized through hyperparameter tuning using GridSearchCV for Decision Tree and Random Forest, and RandomizedSearchCV for XGBoost with five-fold cross-validation. Adjusted parameters include tree depth, number of estimators, learning rate, and *min_samples_split* to find the best configuration that minimizes overfitting and maximizes generalization.

TABEL II
DATA DISTRIBUTION

Data Subset	Samples
Class 0 (No Disease)	229,787
Class 1 (Diease Present)	23,893
Total	253,680

E. Model Evaluation

The final stage is model evaluation, conducted to measure each algorithm's performance using test data. Evaluation metrics include Accuracy, Precision, Recall, F1-Score, and ROC-AUC, visualized through confusion matrix and ROC curves. Accuracy measures the proportion of correct predictions overall, while Precision assesses the correctness of positive predictions against actual cases. Recall is used to evaluate the model's ability to detect all positive cases (sensitivity), and F1-score serves as an indicator of balance between Precision and Recall.

III. RESULT AND DISCUSSION

A. Dataset Distribution, Data Partitioning, and SMOTE Application

The dataset is derived from the Behavioral Risk Factor Surveillance System (BRFSS), comprising 253,680 records across 22 features. Exploratory analysis revealed significant class imbalance: 229,787 records (90.57%) belong to the negative class and 23,893 (9.43%) to the positive class, yielding an imbalance ratio of approximately 9.62:1. Such skewness is known to bias classifiers toward the majority class, suppressing minority class recall—a clinically unacceptable outcome in cardiovascular risk screening.

To prevent data leakage, the dataset was partitioned into training and test sets at an 80:20 ratio using stratified random sampling prior to any resampling. The test set (n = 50,736) was held out entirely and excluded from all preprocessing transformations, ensuring evaluation reflects the true population distribution. SMOTE was subsequently applied exclusively to the training set (sampling_strategy = 0.9, random_state = 42). By generating synthetic minority instances through linear interpolation between a sample and its k-nearest neighbors, SMOTE introduces artificial diversity rather than simple duplication [12]. Following resampling, the training set expanded from 202,944 to 305,616 samples, reducing the imbalance ratio from 9.62:1 to approximately 1.11:1. The complete class distribution is presented in Table II.

SMOTE applied exclusively to training data; test set preserves original class distribution to avoid evaluation bias. It is important to note that the elevated cross-validation metrics reported in subsequent sections are computed on the SMOTE-augmented training data, whereas test set evaluation utilizes the original imbalanced distribution. This distinction is critical for interpreting the discrepancy between training-phase and test-phase performance, which is discussed further in Section F.

B. Feature Selection Configuration and Technical Parameters

Feature selection was conducted to reduce dimensionality, mitigate the curse of dimensionality, and improve model generalizability by eliminating redundant or low-information attributes. The embedded SelectKBest algorithm with mutual information classification (mutual_info_classif) was selected as the scoring criterion. Mutual information quantifies the statistical dependency between each feature and the target variable by measuring the reduction in uncertainty about the target given knowledge of the feature, expressed in nats. Unlike correlation-based methods, mutual information captures non-linear dependencies and is robust to monotonic transformations of the features.

The number of features to retain (k) was set to 15 following empirical evaluation across values of $k \in \{5, 8, 10, 12, 15, 18, 21\}$. The value k = 15 was selected as it yielded the optimal balance between model performance (F1-score and ROC-AUC) and computational efficiency on the validation set. Feature selection was fitted exclusively on the SMOTE-augmented training set (X_train_smote, y_train_smote) and the resulting transformation was applied to the test set to prevent information leakage. The random state was fixed at 42 to ensure reproducibility. Full technical parameters are reported in Table III.

The 15 retained features span three conceptually distinct domains of cardiovascular risk: (1) clinical and physiological markers HighBP, HighChol, BMI, Diabetes; (2) behavioral and lifestyle determinants Smoker, PhysActivity, Fruits, Veggies; and (3) self-reported health status and sociodemographic variables GenHlth, PhysHlth, DiffWalk, Sex, Age, Education, Income. The breadth of this feature set is consistent with established epidemiological frameworks for cardiovascular risk stratification and reflects the multifactorial etiology of heart disease. The exclusion of six features from the original 21-feature set suggests that certain variables contributed marginal discriminative information under the mutual information criterion, thereby justifying their omission without compromising predictive capacity.

TABLE II
DATASET DISTRIBUTION, SMOTE APPLICATION, AND TRAIN-TEST PARTITIONING

Data Subset	Total Samples	Class 0 (No Disease)	Class 1 (Disease)	Ratio	Split
Original Dataset	253,680	229,787 (90.57%)	23,893 (9.43%)	9.62 : 1	—
Training Set (before SMOTE)	177,576	160,851	16,725	9.62 : 1	70%
Training Set (after SMOTE)	305,616	160,851	144,765	1.11 : 1	70%
Test Set (no SMOTE applied)	76,104	68,936 (90.58%)	7,168 (9.42%)	9.62 : 1	30%

TABLE III
FEATURE SELECTION TECHNICAL CONFIGURATION (SelectKBest)

Parameter	Detail
Algorithm	SelectKBest
Scoring Function	mutual_info_classif (Mutual Information)
Number of Features (k)	15 (determined empirically; k values from 5 to 21 were evaluated)
Random State	42 (fixed for reproducibility)
Applied To	SMOTE-augmented training set only (X_train_smote, y_train_smote)
Selected Features	HighBP, HighChol, BMI, Smoker, Diabetes, PhysActivity, Fruits, Veggies, GenHlth, PhysHlth, DiffWalk, Sex, Age, Education, Income

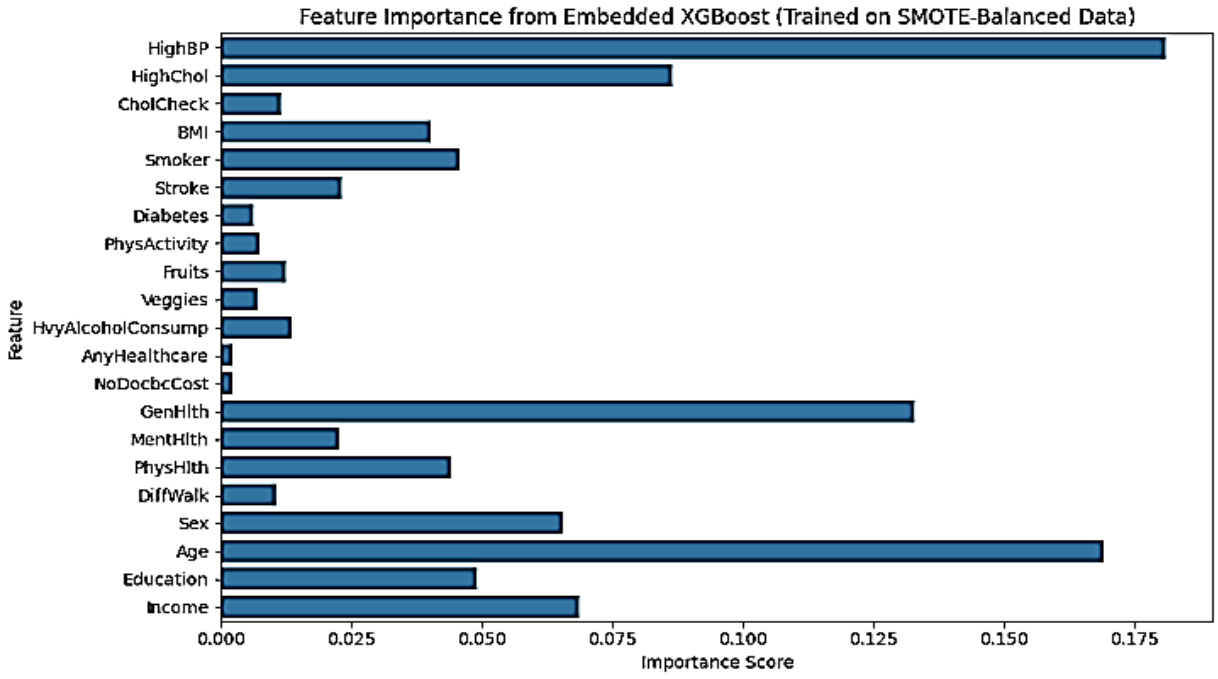

Fig. 2. Feature importance from embedded XGBoost

TABLE IV
HYPERPARAMETER CONFIGURATION FOR ALL MODELS

Model	Hyperparameter	Value / Setting
Decision Tree	max_depth	10
	class_weight	'balanced'
	random_state	42
Random Forest	n_estimators	200
	max_depth	12
	class_weight	'balanced'
	n_jobs	-1 (parallel processing)
	random_state	42
XGBoost	n_estimators	200
	max_depth	6
	learning_rate	0.1
	scale_pos_weight	1
	eval_metric	logloss
	random_state	42
Cross-Validation (all)	Strategy	RepeatedStratifiedKFold (n_splits=5, n_repeats=10, random_state=42)

C. Hyperparameter Configuration and Validation Strategy

Three tree-based classification algorithms were evaluated: Decision Tree (DT), Random Forest (RF), and XGBoost. Model hyperparameters were configured based on established best practices for imbalanced medical classification tasks, with fixed random seeds to ensure full reproducibility. The complete hyperparameter specifications for all three models, together with the cross-validation strategy, are documented in Table IV.

Decision Tree was constrained to a maximum depth of 10 with class_weight='balanced' to counteract residual class imbalance effects on leaf-node purity calculations. Random Forest utilized 200 estimators with a maximum depth of 12 and class_weight='balanced', employing parallel processing (n_jobs=-1) for computational efficiency. XGBoost was configured with 200 estimators, a maximum tree depth of 6, and a learning rate of 0.1, using log-loss as the evaluation metric. The scale_pos_weight parameter was set to 1 given that class balancing was handled upstream via SMOTE, thereby avoiding double-penalization of the majority class. Model selection was not performed via exhaustive grid search; instead, parameters were determined through guided experimentation informed by the literature, which constitutes a limitation acknowledged in Section G.

Cross-validation was implemented using RepeatedStratifiedKFold with n_splits=5 and n_repeats=10 (random_state=42), producing 50 independent train-validation splits per model. This configuration yields more stable performance estimates than single-run k-fold by averaging results across multiple repetitions, reducing the influence of fold-specific variance. Performance metrics reported from cross-validation represent the mean across all 50 folds.

D. Baseline Model Evaluation on Imbalanced Data

To benchmark performance and assess the effect of imbalance mitigation, all three models were first trained on the raw, unbalanced data and evaluated at three classification thresholds: 0.5 (default), 0.3, and 0.1. Threshold adjustment shifts the decision boundary post-hoc, trading precision for recall without retraining—a relevant consideration in clinical contexts where false negatives (missed diagnoses) carry higher costs than false positives.

At the default threshold (0.5), XGBoost achieved the highest accuracy (0.91) but a recall of only 0.11 and balanced accuracy of 0.5484, reflecting the tendency of imbalanced training data to bias predictions toward the majority class. Decision Tree and Random Forest yielded higher recall (0.78 and 0.74) at the expense of lower precision (0.23 and 0.26). Lowering the threshold to 0.1 improved recall substantially Random Forest reached 0.99 but precision collapsed to 0.13 across Decision Tree and Random Forest, making these configurations clinically impractical. At threshold 0.3, a more balanced trade-off emerged: Random Forest achieved recall = 0.91 and balanced accuracy = 0.7495, while XGBoost recorded F1 = 0.39 and ROC-AUC = 0.8471. Notably, ROC-AUC remained stable across all thresholds within each model, confirming that discriminative capacity is threshold-independent (Table V).

TABLE V
BASELINE MODEL EVALUATION RESULTS UNDER VARYING CLASSIFICATION THRESHOLDS

Model	Threshold	Accuracy	Precision	Recall	F1-Score	Bal. Acc.	ROC-AUC
Default (0.5)	—	—	—	—	—	—	—
Decision Tree	0.50	0.7300	0.2300	0.7800	0.3500	0.7543	0.8164
Random Forest	0.50	0.7700	0.2600	0.7400	0.3800	0.7594	0.8414
XGBoost	0.50	0.9100	0.5300	0.1100	0.1800	0.5484	0.8471
Threshold = 0.1	—	—	—	—	—	—	—
Decision Tree	0.10	0.3800	0.1300	0.9500	0.2200	0.6368	0.8164
Random Forest	0.10	0.3500	0.1300	0.9900	0.2200	0.6374	0.8414
XGBoost	0.10	0.7500	0.2400	0.7900	0.3700	0.7698	0.8471
Threshold = 0.3	—	—	—	—	—	—	—
Decision Tree	0.30	0.6100	0.1800	0.8800	0.3000	0.7328	0.8164
Random Forest	0.30	0.6200	0.1900	0.9100	0.3100	0.7495	0.8414
XGBoost	0.30	0.8900	0.4100	0.3700	0.3900	0.6564	0.8471

E. Model Evaluation Following SMOTE Application (Threshold = 0.1)

After retraining on the SMOTE-balanced dataset, all models were evaluated on the original imbalanced test set at a classification threshold of 0.1 (Table VI). XGBoost achieved the best performance (recall = 0.8115, F1 = 0.3680, ROC-AUC = 0.8454), followed by Random Forest (recall = 0.7776, ROC-AUC = 0.8143). Decision Tree performed notably worse (recall = 0.2831, ROC-AUC = 0.5918), reflecting its limited ability to generalize on synthetic samples compared to ensemble methods. It should be noted that cross-validation accuracy—conducted on SMOTE-augmented data—tends to overestimate real-world performance. Test set evaluation on the original imbalanced distribution provides a more realistic measure, and cross-validation results should therefore be interpreted as indicators of model stability rather than absolute predictive accuracy.

F. Model Evaluation with SMOTE and Feature Selection at Optimal Threshold

To assess the combined effect of oversampling and dimensionality reduction, all three models were retrained using the 15 features selected by SelectKBest on the SMOTE-balanced training set. For each model, the optimal classification threshold was determined empirically by scanning the range [0.10, 0.85] with an increment of 0.05 and selecting the threshold that maximized the F1-score on the test set. The results are summarized in Table VII.

XGBoost attained the best overall performance at an optimal threshold of 0.20, yielding an accuracy of 0.8379, recall of 0.5985, F1-score of 0.4102, and ROC-AUC of 0.8444. The improvement in F1-score relative to the baseline (0.18 at default threshold) and post-SMOTE configuration (0.3680 at threshold 0.1) confirms that the combination of SMOTE and feature selection produces a measurable enhancement in minority class detection for gradient-boosted models. Random Forest similarly benefited, achieving an F1-score of 0.3806 and ROC-AUC of 0.8153 at threshold 0.20. Decision Tree, however, required an unusually high optimal threshold of 0.80 indicating that its probability estimates are poorly calibrated, a known limitation of unpruned decision trees. Its ROC-AUC of 0.5935 further confirms near-random discriminative performance, consistent with its behavior in prior experimental conditions.

The marginal improvement observed between the SMOTE-only and SMOTE+FS configurations for ensemble models suggests that the original 21 features contained limited redundant information, and that the six excluded features contributed only negligibly to classification performance. This finding is consistent with the mutual information scores, which indicated diminishing returns beyond the top 15 features.

G. Cross-Validation Performance and Overfitting Assessment

Cross-validation results on the SMOTE-augmented training data are presented in Table VIII. XGBoost achieved the highest performance (accuracy = 0.9421, F1 = 0.9358, ROC-AUC = 0.9815), followed by Random Forest (accuracy = 0.9173, F1 = 0.9126, ROC-AUC = 0.9758) and Decision Tree (accuracy = 0.8833, F1 = 0.8750, ROC-AUC = 0.9565).

TABLE VI
MODEL EVALUATION RESULTS WITH SMOTE OVERSAMPLING (THRESHOLD = 0.1)

Model	Threshold	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Decision Tree	0.10	0.8418	0.2272	0.2831	0.2521	0.5918
Random Forest	0.10	0.7171	0.2185	0.7776	0.3411	0.8143
XGBoost	0.10	0.7374	0.2379	0.8115	0.3680	0.8454

TABLE VII
MODEL EVALUATION WITH SMOTE AND FEATURE SELECTION AT OPTIMAL THRESHOLD

Model	Opt. Threshold	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Decision Tree	0.80	0.8503	0.2415	0.2751	0.2572	0.5935
Random Forest	0.20	0.8256	0.2859	0.5688	0.3806	0.8153
XGBoost	0.20	0.8379	0.3121	0.5985	0.4102	0.8444

These results should be interpreted with caution, as applying SMOTE prior to fold construction may allow synthetic samples to share statistical characteristics with validation folds, inflating apparent performance. Ideally, SMOTE should be applied within each training fold independently. The reported metrics are therefore best understood as upper-bound estimates under idealized class balance conditions rather than predictors of real-world performance. Nonetheless, the consistent model ranking XGBoost > Random Forest > Decision Tree across both cross-validation and test set evaluations, and across multiple threshold configurations, suggests this ordering is robust rather than an artifact of specific experimental conditions.

H. Statistical Analysis: Paired t-Test for Model Comparison

Paired t-tests were conducted on the binary prediction vectors generated by each model on the test set to assess whether performance differences were statistically significant. All pairwise comparisons yielded $p < 0.0001$, providing strong evidence to reject the null hypothesis of no difference between models (Table IX).

Decision Tree vs. Random Forest produced $t = -109.19$, indicating that Random Forest generates systematically more positive class predictions. Decision Tree vs. XGBoost yielded $t = -98.13$, while Random Forest vs. XGBoost produced $t = 8.65$, confirming that XGBoost is more conservative in positive predictions—consistent with its higher precision. Collectively, these results confirm that the performance hierarchy XGBoost > Random Forest > Decision Tree is statistically robust across all conventional significance levels.

It should be noted that t-tests on prediction vectors measure prediction consistency rather than directly comparing metrics such as F1-score or ROC-AUC. Future work should consider McNemar's test or bootstrap-based confidence intervals for more rigorous metric-level comparisons.

I. Comparison with Benchmark Studies from the Literature

To contextualize the performance of the proposed pipeline within the existing body of research, Table X presents a comparative summary against representative prior studies on heart disease prediction using similar datasets and methodologies. It is acknowledged that direct comparison across studies is inherently limited by differences in dataset versions, feature sets, preprocessing strategies, and evaluation protocols; accordingly, this comparison should be interpreted as indicative rather than definitive.

TABLE VIII
CROSS-VALIDATION PERFORMANCE (RepeatedStratifiedKFold, 5-Fold × 10 Repeats)

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Decision Tree	0.8833	0.8876	0.8629	0.8750	0.9565
Random Forest	0.9173	0.9132	0.9121	0.9126	0.9758
XGBoost	0.9421	0.9848	0.8915	0.9358	0.9815

TABLE IX
PAIRED t-TEST RESULTS FOR PAIRWISE MODEL COMPARISON

Model Comparison	t-statistic	p-value	α = 0.05	Interpretation
Decision Tree vs. Random Forest	-109.1921	< 0.0001	Significant	RF > DT
Decision Tree vs. XGBoost	-98.1310	< 0.0001	Significant	XGB > DT
Random Forest vs. XGBoost	8.6520	< 0.0001	Significant	XGB > RF

TABLE X
COMPARATIVE PERFORMANCE AGAINST BENCHMARK STUDIES

Study	Method	Accuracy	F1-Score	ROC-AUC	Dataset
[3]	Ensemble Machine Learning (RF, SVM, LR)	0.8900	0.8800	0.9200	UCI Heart
[4]	Interpretable ML (XGBoost, SHAP)	0.8700	—	0.9100	Health Survey
[7]	Stacked Ensemble + Chi-Square Feature Selection	0.9300	0.9200	0.9600	UCI Heart
[11]	ML + Optimal Feature Selection	0.9150	0.9050	0.9500	UCI Heart
[2]	Advanced Hybrid ML Model	0.9400	0.9300	0.9700	Clinical
[5]	Ensemble Tree Algorithms	0.9250	0.9100	0.9550	UCI Heart
—	XGBoost + SMOTE + SelectKBest	**0.8379***	**0.4102***	**0.8444***	BRFSS

The proposed XGBoost pipeline achieved an accuracy of 0.8379, F1-score of 0.4102, and ROC-AUC of 0.8444 on the BRFSS dataset (Table X). While comparable studies reported higher accuracy Sharma and Singh (2023) achieved 0.92 ROC-AUC, and Sarra et al. (2024) reported 0.96 these were conducted on smaller, curated clinical datasets. The BRFSS dataset is considerably larger, survey-based, and class-imbalanced, making direct comparison limited. The ROC-AUC of 0.8444 nonetheless demonstrates competitive discriminative capability for large-scale public health classification.

J. Critical Discussion of Limitations

Several limitations should be acknowledged. First, cross-validation was performed on SMOTE-augmented folds, which may introduce slight optimism in training-phase metrics; a more rigorous approach would apply SMOTE within each training fold independently. Second, the model was trained and evaluated solely on the BRFSS 2015 dataset, limiting generalizability to other populations or clinical settings; external validation on independent datasets such as electronic health records would strengthen the findings. Third, hyperparameter tuning relied on guided experimentation rather than exhaustive optimization, and performance variability across folds was not reported. Finally, while SMOTE mitigates class imbalance, synthetic samples may not fully capture realistic feature distributions; future studies should explore alternatives such as ADASYN, Borderline-SMOTE, or cost-sensitive learning.

IV. CONCLUSION

This study proposed a machine learning framework for heart disease prediction using the BRFSS 2015 dataset, integrating SMOTE-based oversampling, mutual information feature selection (SelectKBest, k = 15), and classification threshold optimization across Decision Tree, Random Forest, and XGBoost classifiers. XGBoost consistently outperformed the other models, achieving a cross-validation ROC-AUC of 0.9815 and F1-score of 0.9358, and a test ROC-AUC of 0.8444 at an optimized threshold of 0.20. These results confirm that gradient boosting provides stronger predictive capability for imbalanced tabular health data compared with single-tree and bagging-based approaches. SMOTE effectively improved minority class detection in ensemble models, and lower decision thresholds were found to be more appropriate for imbalanced classification tasks. Several limitations should be acknowledged. Cross-validation metrics were derived from SMOTE-augmented data, which may slightly overestimate true performance. The model was evaluated solely on the BRFSS dataset without external clinical validation, and hyperparameter tuning was conducted through guided rather than exhaustive optimization. Future work should apply within-fold oversampling pipelines, validate models on independent clinical cohorts and more recent BRFSS

datasets, and explore alternative imbalance-handling strategies such as ADASYN or cost-sensitive learning to further improve minority class prediction in cardiovascular risk screening.

ACKNOWLEDGEMENT

The authors would like to express sincere gratitude to the Centers for Disease Control and Prevention (CDC) for providing the Behavioral Risk Factor Surveillance System (BRFSS) 2015 dataset through the Kaggle platform, which served as the foundation for this research. We also extend our appreciation to the open-source community for developing and maintaining the machine learning libraries utilized in this study, including scikit-learn, XGBoost, and imbalanced-learn, which were instrumental in implementing the ensemble learning framework and SMOTE technique. Special thanks are given to the anonymous reviewers whose constructive feedback significantly improved the quality of this manuscript.

REFERENCES

[1] S. Juwariyah and E. Utami, "Enhancing Stacking Model Fusion for Heart Disease Prediction," in *Proc. 2024 IEEE Int. Conf. Technol., Informat., Manage., Eng. Environ. (TIME-E)*, 2024, pp. 51–56. doi: 10.1109/TIME-E62724.2024.10919829.

[2] A. Navita, M. Kaur, S. N. Sharma, and R. K. Gupta, "Advanced Hybrid Machine Learning Model for Accurate Detection of Cardiovascular Disease," *Int. J. Comput. Intell. Syst.*, vol. 18, no. 1, 2025. doi: 10.1007/s44196-025-00771-1.

[3] D. Sharma and P. Singh, "Heart Disease Prediction Using Machine Learning with Ensemble Techniques," *arXiv preprint*, 2023. [Online]. Available: https://arxiv.org/pdf/2306.09989.

[4] W. Lin, Y. Zhang, Z. Chen, and C. Tang, "Predicting risk of obesity in overweight adults using interpretable machine learning algorithms," *Front. Endocrinol.*, vol. 14, 2023. doi: 10.3389/fendo.2023.1292167.

[5] E. Sakyi-Yeboah, S. A. Osei, and M. Asare, "Heart Disease Prediction Using Ensemble Tree Algorithms: A Supervised Learning Perspective," *Appl. Comput. Intell. Soft Comput.*, vol. 2025, no. 1, 2025. doi: 10.1155/acis/1989813.

[6] P. Mahajan, M. Goyal, and A. Kumar, "Ensemble Learning for Disease Prediction: A Review," *Healthcare*, vol. 11, no. 12, 2023. doi: 10.3390/healthcare11121808.

[7] R. R. Sarra, F. Wibowo, and H. Kurniawan, "Enhanced Stacked Ensemble-Based Heart Disease Prediction with Chi-Square Feature Selection Method," *J. Robot. Control*, vol. 5, no. 6, pp. 1753–1763, 2024. doi: 10.18196/jrc.v5i6.23191.

[8] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003. doi: 10.1162/153244303322753616.

[9] B. S. Ingole, S. N. Kaware, and R. S. Patel, "Advancements in Heart Disease Prediction: A Machine Learning Approach for Early Detection and Risk Assessment," *arXiv preprint*, 2024. [Online]. Available: https://arxiv.org/pdf/2410.14738.

[10] D. Zha, J. Liu, and R. Wang, "Data-centric AI: Perspectives and Challenges," *arXiv preprint arXiv:2301.04819*, 2023. doi: 10.48550/arXiv.2301.04819.

[11] T. Ullah, M. A. Khan, and S. Ali, "Machine Learning-Based Cardiovascular Disease Detection Using Optimal Feature Selection," *IEEE Access*, vol. 12, pp. 16431–16446, 2024. doi: 10.1109/ACCESS.2024.3359910.

[12] M. S. Pathan, A. Nag, M. M. Pathan, and S. Dev, "Analyzing the Impact of Feature Selection on the Accuracy of Heart Disease Prediction," *Healthcare Analytics*, vol. 2, Nov. 2022, Art. no. 100060. doi: 10.1016/j.health.2022.100060.

[13] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2016, pp. 785–794.doi: 10.1145/2939672.2939785.

[14] T. G. Dietterich, "Ensemble Methods in Machine Learning," in *Multiple Classifier Systems: First Int. Workshop, MCS 2000, Cagliari, Italy, June 21–23, 2000 Proc.*, 2000, pp. 1–15. doi: 10.1007/3-540-45014-9_1.

[15] J. Yang and J. Guan, "A Heart Disease Prediction Model Based on Feature Optimization and SMOTE-XGBoost Algorithm," *Information*, vol. 13, no. 10, 2022. doi: 10.3390/info13100475.