

An Optimized Stacking Ensemble Framework With Synthetic Data Augmentation and Gradient-Boosted Feature Selection For Heart Disease Classification

Syed Muhammad Faisal¹, Mudasir Mahmood^{1*}, Muhammad Farhan¹

¹Faculty of Computing, Gomal University, Dera Ismail Khan, Pakistan.

Corresponding author: Mudasir Mahmood, Email: mudasir@gu.edu.pk

Abstract

According to the World Health Organization, heart disease causes about 17.9 million deaths every year and is one of the leading causes of death worldwide. Early diagnosis improves patient outcomes and lowers healthcare costs, but conventional diagnostic methods are costly, time-consuming and require specialist medical skills. This study develops an optimized machine learning framework based on a stacking ensemble for accurate heart disease prediction and early clinical diagnosis. Five benchmark heart disease datasets (Cleveland, Hungarian, Switzerland, Long Beach and Statlog) were combined to increase the diversity and reliability of the data. Synthetic data augmentation, based on bootstrap sampling and controlled noise injection, was used to compensate for the small sample size and to enlarge the dataset to about 30,000 observations. Gradient Boosted Decision Trees (GBDT) were used to identify the most important predictors. The stacking ensemble consisted of four base learners (Random Forest, Decision Tree, K-Nearest Neighbours and Logistic Regression) and a Logistic Regression meta-learner. Model performance was assessed with accuracy, precision, recall, F1-score and the area under the receiver operating characteristic curve (ROC-AUC) using five-fold cross-validation and a held-out test set. The proposed stacking ensemble achieved an accuracy of 99.97%, with precision, recall and F1-score close to 100% for both classes and a ROC-AUC of 1.000. The combination of several data sources, synthetic data augmentation and ensemble learning was central to this performance, and the optimized stacking ensemble shows promise for the early diagnosis of heart disease in clinical decision support.

Keywords: heart disease; machine learning; stacking ensemble; synthetic data augmentation; gradient boosting; feature selection.

1. Introduction

Heart disease is one of the leading causes of death in every region of the world. According to the World Health Organization, cardiovascular diseases cause 17.9 million deaths each year, about 32% of all deaths worldwide [1]. The term covers several conditions, including coronary artery disease, heart attack, heart failure and arrhythmia. These conditions severely reduce quality of life and can lead to premature death if they are not properly controlled. Cardiovascular disease remains the leading cause of death in the United States, where it accounted for 931,578 deaths in 2021 [2]. Early detection is essential for controlling the high cost of heart disease and improving patient outcomes. With the growth of biotechnology and electronic health records, the volume of health data has increased rapidly, and these data can be used to identify patterns, detect abnormalities and predict heart disease. Timely and accurate prediction allows clinicians to act promptly and take preventive measures, and diagnosis at an early stage is crucial for reducing deaths [3-5].

Public health strategies for preventing heart disease are especially important in Asia, Europe, Africa and the Middle East, where the burden of heart disease is expected to be greatest, largely because of high cholesterol levels, unhealthy diets, high blood pressure and air pollution [6, 7]. Machine learning

(ML) has received considerable attention in recent years because of its potential in healthcare diagnostics, including the diagnosis of heart disease. ML techniques can detect the disease at a lower cost and with reasonable accuracy [8]. When trained on large datasets, ML models can detect abnormalities at an early stage, which saves time and money for patients and healthcare systems. This paper introduces a new ML framework for heart disease prediction based on an optimized stacking ensemble, compares its performance with previous studies and shows that the proposed framework achieves the highest accuracy among the compared approaches.

Related work

In recent years, ML algorithms have been increasingly used to predict and detect heart disease by extracting hidden patterns from large medical datasets. Naive Bayes (NB), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM) and K-Nearest Neighbours (KNN) are among the algorithms considered particularly effective for this task. Mahmud et al. [1] proposed a meta-model combining RF, Gaussian NB, DT and KNN, trained on five heart disease datasets (Statlog, Cleveland, Hungarian, Switzerland and Long Beach), and obtained an accuracy of 87%. Mansur Huang et al. [9] compared RF, Logistic Regression (LR), SVM and Neural Networks (NN) and found that RF gave the best overall performance, with an accuracy of 0.88. Noroozi et al. [10] applied seven ML models, including NB, SVM, a DT (J48) and RF, together with several feature selection methods to the Cleveland dataset, and obtained the best accuracy (85.5%) with SVM combined with feature selection. Such studies commonly evaluate models with accuracy, precision, recall (sensitivity), F1-score and the area under the receiver operating characteristic curve (ROC-AUC).

Remya [11] combined Convolutional Neural Networks (CNN) with the feature selection techniques ReliefF, UMAP and LDA on the Cleveland dataset and found the CNN-UMAP combination to be the most effective, with an accuracy of 91.88%. Chicco and Jurman [12] studied ML methods for predicting the survival of patients with heart failure, including survival prediction models and feature selection methods. Plati et al. [13] showed that advanced ensemble methods such as Rotation Forest can reach an accuracy of 91.23% in heart failure diagnosis. Ram Kumar et al. [14] compared ML models (KNN and SVM) with deep learning models (Artificial Neural Network, ANN, and CNN) on the Cleveland dataset and found that CNN performed best (83.61%), followed by SVM (74.2%), ANN (70.08%) and KNN (66.7%). Mutinda and Langat [15] investigated the effect of dimensionality reduction (PCA, KPCA and LDA) on LR, SVM, KNN, NB and Deep Neural Networks and found that LDA gave the most favourable results. Logabiraman et al. [16] evaluated RF, Gradient Boosting, XGBoost, SVM and Artificial Neural Networks for heart disease prediction.

Jindal et al. [17] proposed a prediction system based on LR and KNN that performed better than NB. Bhatt et al. [18] used k-modes clustering with Huang initialization followed by DT, Multilayer Perceptron and XGBoost, and reported accuracies of 86.53-87.28% and an AUC of up to 0.95 on 70,000 records of a Kaggle dataset. Alshraideh et al. [19] used SVM with Particle Swarm Optimization (PSO) on the Jordan University Hospital (JUH) heart dataset and obtained an accuracy of 94.3% in heart disease risk classification. These studies are summarized in Table 1.

Table 1. Comparative overview of previous studies on heart disease prediction.

Year	Ref.	Dataset	Methods and reported performance
2020	[12]	Clinical records of 299 heart failure patients, Faisalabad Institute of Cardiology and Allied Hospital, Pakistan	LR: 83%
2021	[13]	410 cases from University College Dublin (Ireland) and 77 cases from the University Hospital of Ioannina (Greece)	Rotation Forest: 91.23%
2021	[9]	Kaggle heart disease dataset, 303 patients	RF: 88%; SVM: 83%; NB: 85%; LR: 87%

Year	Ref.	Dataset	Methods and reported performance
2023	[10]	Cleveland heart disease dataset, 297 records	SVM with feature selection: 85.5%
2023	[1]	Statlog, Cleveland, Hungarian, Switzerland and Long Beach datasets	Meta-model (RF, GNB, DT, KNN): 87%
2023	[18]	Kaggle dataset, 70,000 records	RF: 95%; DT: 94%; MLP: 95%; XGBoost: 95%
2024	[11]	Cleveland heart disease dataset, 303 records	CNN with UMAP: 91.88%
2024	[14]	Cleveland heart disease dataset	CNN: 83.61%; SVM: 74.2%; ANN: 70.08%; KNN: 66.7%
2024	[15]	Kaggle dataset, 1,100 records, 14 features	LR: 84.07%; SVM: 65.19%; KNN: 62.69%; NB: 85.19%; DNN: 76.05%
2024	[16]	Kaggle dataset, 303 instances, 14 features	LR, DT, XGBoost, GB, RF, SVM, ANN: accuracy not reported
2024	[19]	Jordan University Hospital (JUH) dataset, Amman, Jordan	SVM: 94.3%

Note: GNB, Gaussian Naive Bayes; MLP, Multilayer Perceptron; GB, Gradient Boosting; ANN, Artificial Neural Network; DNN, Deep Neural Network.

Table 1 gives an overview of recent studies (2020-2024) on ML-based heart disease prediction, including the datasets, the models used and the reported performance. Common datasets include the Cleveland, Kaggle and Statlog datasets and hospital records, and common models include LR, SVM, RF, XGBoost, Gradient Boosting, NN and CNN. Among these studies, the highest accuracy, 94.3%, was reported for SVM optimized with PSO [19]. However, several problems remain, such as class imbalance, small data volume, limited real-time processing and limited data diversity.

Problem statement

The aim is to train an ML system that accurately predicts whether a patient has heart disease from clinical and demographic factors. In other words, the task is to find a predictive mapping $M: X \rightarrow Y$, where X is the set of patient characteristics, such as age, sex, chest pain type, resting blood pressure, cholesterol level, fasting blood sugar, electrocardiographic results, maximum heart rate, exercise-induced angina, ST depression and ST slope, and $Y \in \{0, 1\}$ indicates the absence (0) or presence (1) of heart disease. Performance is evaluated with standard metrics: accuracy, precision, recall, F1-score and ROC-AUC. The goal is to achieve high performance while maintaining good generalization to patients who were not seen during training.

Novelty and contributions of this study

The novelty of this work lies in combining five heterogeneous benchmark heart disease datasets, synthetic data augmentation, Gradient Boosted Decision Tree (GBDT)-based feature selection and a two-level stacking ensemble trained on out-of-fold predictions in a single optimized framework for heart disease classification. The main contributions are: (i) a stacking ensemble model for accurate heart disease prediction; (ii) the integration of the Cleveland, Hungarian, Switzerland, Long Beach and Statlog datasets to increase data diversity and model generalization; (iii) synthetic data augmentation to deal with data scarcity, which increases the dataset to about 30,000 records; (iv) GBDT-based feature selection that improves model efficiency and predictive performance; (v) five-fold cross-validation to improve robustness and stability and to reduce overfitting; and (vi) an accuracy of 99.97% with near-

perfect values of the other evaluation measures, which exceeds the accuracies reported in the compared studies.

Materials and methods

Overall framework

The proposed framework was developed for binary heart disease classification using heterogeneous benchmark datasets, synthetic data augmentation, GBDT-based feature selection and a two-level stacking ensemble classifier. Each patient record was represented as a feature vector of clinically relevant demographic and physiological variables, together with a binary target variable indicating the absence or presence of heart disease. The dataset is represented as

$$D = \{(\mathbf{x}_i, y_i)\}_{i=1}^N \quad (1)$$

where $\mathbf{x}_i \in \mathbb{R}^d$ is the feature vector of the i -th patient, $y_i \in \{0, 1\}$ is the corresponding class label, N is the number of patient records and d is the number of input features.

The framework consists of five main stages: (i) integration and preprocessing of heterogeneous benchmark heart disease datasets, (ii) synthetic data augmentation using bootstrap sampling and controlled perturbation, (iii) supervised feature selection using GBDT, (iv) training of a two-level stacking ensemble using out-of-fold probability predictions and (v) performance evaluation using five-fold cross-validation and a held-out test set.

The processing pipeline (Fig. 1) begins with the integration and harmonization of the five benchmark datasets. Missing values are then treated and duplicate records are removed, and the data are transformed into a common feature representation. The resulting dataset is augmented to about 30,000 observations using bootstrap sampling and controlled noise injection. The GBDT model ranks the candidate clinical predictors, and the seven highest-ranked predictors are retained for classification. These features are supplied to four heterogeneous base classifiers: RF, DT, KNN and LR. At the second level, an LR meta-learner combines their probability outputs. Five-fold cross-validation is used to produce out-of-fold predictions for training the meta-learner and to minimize information leakage between the base and meta-learning stages. The final ensemble is then evaluated on the held-out test subset.

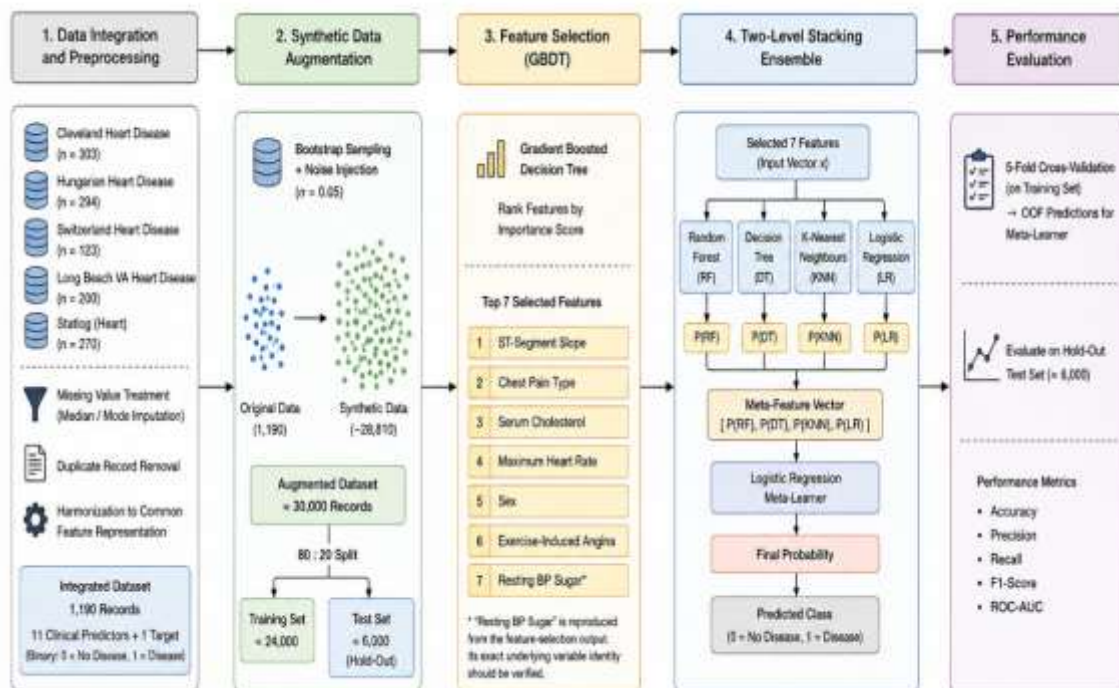


Fig. 1. Overall workflow of the proposed optimized stacking ensemble framework for heart disease classification.

Dataset description, integration and preprocessing

Five publicly available benchmark heart disease datasets were used: Statlog (Heart), Cleveland, Hungarian, Switzerland and Long Beach VA. These datasets contain patient records collected in different clinical and geographical settings and were chosen to increase the variation in the observations used for model development. The Cleveland dataset contains 303 records, the Hungarian dataset 294, the Switzerland dataset 123, the Long Beach VA dataset 200 and the Statlog (Heart) dataset 270. After integration and harmonization, the merged dataset contained 1,190 patient records.

The source datasets differ in their original feature representations and class distributions. To allow joint model development, they were mapped to a common feature representation with eleven clinical predictor variables: age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiogram (ECG) result, maximum heart rate achieved, exercise-induced angina, ST depression (Oldpeak) and ST-segment slope. The presence or absence of heart disease was represented by a binary target variable, where 0 indicates the absence and 1 the presence of the disease. Before integration, missing numerical values were replaced by the median and missing categorical values by the mode [20, 21]. Duplicate records were then identified and removed before the integrated dataset was used for augmentation and model development. This preprocessing gave a consistent feature representation across the heterogeneous source datasets. The class distribution before and after synthetic data augmentation is shown in Fig. 2.

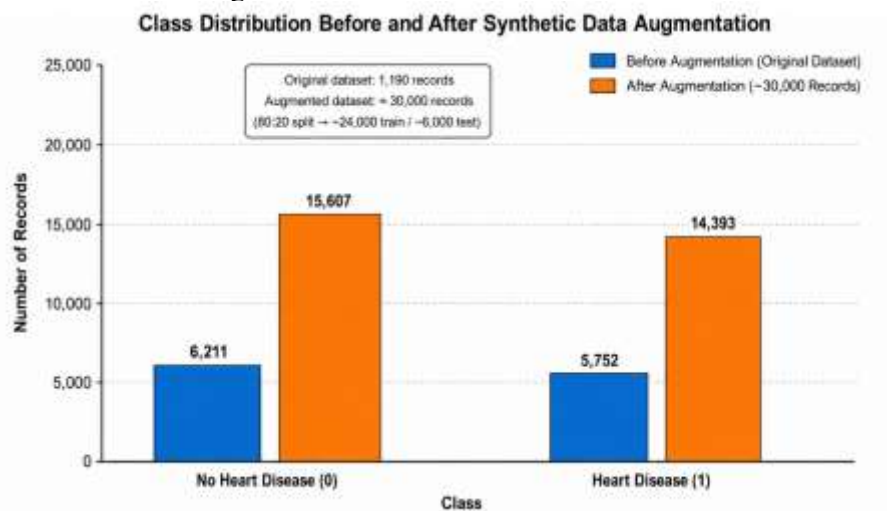


Fig. 2. Class distribution before and after synthetic data augmentation.

The characteristics of the five benchmark datasets are summarized in Table 2, and the clinical variables of the harmonized representation are listed in Table 3.

Table 2. Characteristics of the benchmark heart disease datasets used in this study.

Dataset	Instances	Reported features	Positive cases (%)	Source
Cleveland	303	13	54.5	Cleveland Clinic Foundation, USA
Hungarian	294	13	37.4	Hungarian Institute of Cardiology
Switzerland	123	13	87.8	University Hospitals of Zürich and Basel
Long Beach VA	200	13	67.0	VA Medical Center, Long Beach, USA
Statlog (Heart)	270	13	44.4	Mixed international sources

Dataset	Instances	Reported features	Positive cases (%)	Source
Combined	1,190	11 predictors + 1 target	55.6	Merged and harmonized

Note: The “Reported features” column gives the feature count of each source repository; for the combined dataset it gives the final harmonized representation used in this study (11 predictor variables and one binary target variable).

Table 3. Clinical variables used in the harmonized heart disease dataset.

No.	Feature	Variable type	Description
1	Age	Numerical	Patient age
2	Sex	Categorical/binary	Patient sex
3	Chest pain type	Categorical	Type of chest pain
4	Resting blood pressure	Numerical	Resting blood pressure measurement
5	Serum cholesterol	Numerical	Serum cholesterol level
6	Fasting blood sugar	Categorical/binary	Fasting blood sugar indicator
7	Resting ECG	Categorical	Resting electrocardiographic result
8	Maximum heart rate	Numerical	Maximum heart rate achieved
9	Exercise-induced angina	Categorical/binary	Presence of exercise-induced angina
10	ST depression (Oldpeak)	Numerical	ST-segment depression induced by exercise
11	ST-segment slope	Categorical	Slope of the peak exercise ST segment
12	Heart disease	Binary target	0 = absence of heart disease; 1 = presence of heart disease

Feature standardization

Numerical input variables were standardized before model training using z-score normalization. For a feature x_j , the standardized value was calculated as

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \quad (2)$$

where x_{ij} is the original value of feature j for observation i , μ_j is the mean of feature j and σ_j is its standard deviation. Standardization places the numerical variables on a comparable scale. This step is particularly relevant for the scale-sensitive components of the ensemble, KNN and LR, and improves numerical stability during model fitting. The standardization parameters were derived from the training data and then applied to the corresponding validation and held-out test observations.

Synthetic data augmentation

The integrated benchmark dataset contained 1,190 patient records. To increase the number of observations available for model development and to reduce overfitting, synthetic observations were generated by bootstrap sampling followed by controlled noise injection. If D denotes the original harmonized dataset, the augmented dataset is

$$D_{\text{aug}} = D \cup D_{\text{syn}} \quad (3)$$

where D_{syn} is the set of synthetically generated observations. The target size of the augmented dataset was about 30,000 observations. For numerical variables, controlled perturbation was introduced with a noise parameter $\sigma = 0.05$. The augmented dataset was then divided into training and held-out test subsets in an 80:20 ratio, which gave about 24,000 training observations and 6,000 test observations. The

training subset was used for feature selection, cross-validation and construction of the stacking ensemble, whereas the held-out subset was reserved for the final performance evaluation. Because the synthetic observations were derived from the original benchmark records, the augmented dataset should be regarded as an extension of the benchmark sample rather than as an independent clinical cohort. The reported performance therefore reflects this benchmark and synthetic-data setting and is not equivalent to validation on 30,000 independent patients; external validation on independent clinical cohorts remains necessary to assess real-world generalizability. The configuration of the augmentation procedure is given in Table 4, and the data preparation workflow is shown in Fig. 3.

Table 4. Configuration of the synthetic data augmentation procedure.

Parameter	Configuration
Original integrated observations	1,190
Target augmented observations	About 30,000
Augmentation strategy	Bootstrap sampling + controlled noise injection
Noise parameter, σ	0.05
Numerical variables	Controlled perturbation
Categorical variables	Preserved according to the source observation
Target variable	Binary class label preserved
Dataset division	80% training / 20% testing
Training observations	About 24,000
Test observations	About 6,000
Synthetic-data validation	Kolmogorov–Smirnov (KS) statistic

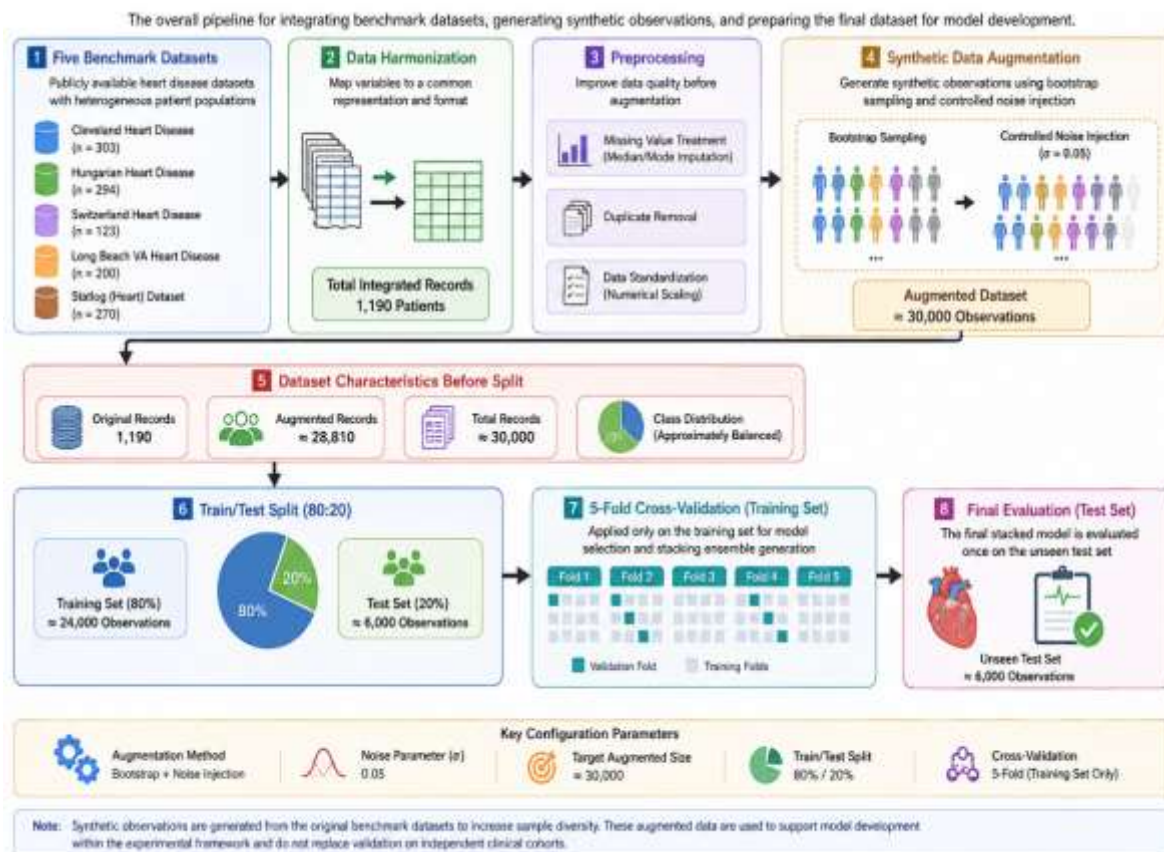


Fig. 3. Synthetic data augmentation and dataset preparation workflow used in the proposed heart disease classification framework.

Gradient Boosted Decision Tree-based feature selection

GBDT was used as a supervised feature selection method to identify the clinical predictors that contribute most strongly to heart disease classification. GBDT builds an additive ensemble of decision trees sequentially, and each new tree is trained to improve the predictive performance of the existing ensemble. The boosted prediction at iteration m can be expressed as

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \quad (4)$$

where $F_m(x)$ is the boosted prediction at iteration m , $F_{m-1}(x)$ is the prediction from the preceding iteration, $h_m(x)$ is the decision tree added at iteration m and γ_m is the learning rate.

After training, the candidate clinical predictors were ranked by their GBDT feature-importance scores. This ranking was used to identify the predictors that contribute most to classification and to reduce the dimensionality of the input to the stacking ensemble. Seven of the 11 candidate predictors were retained: ST-segment slope, chest pain type, serum cholesterol, maximum heart rate, sex, exercise-induced angina and Resting BP Sugar. Their importance scores are given in Table 5 and Fig. 4.

Table 5. GBDT-based ranking of the selected clinical predictors.

Rank	Feature	GBDT feature importance
1	ST-segment slope	0.400951
2	Chest pain type	0.165248
3	Serum cholesterol	0.084089
4	Maximum heart rate	0.082557
5	Sex	0.060322
6	Exercise-induced angina	0.058412
7	Resting BP Sugar	0.051497

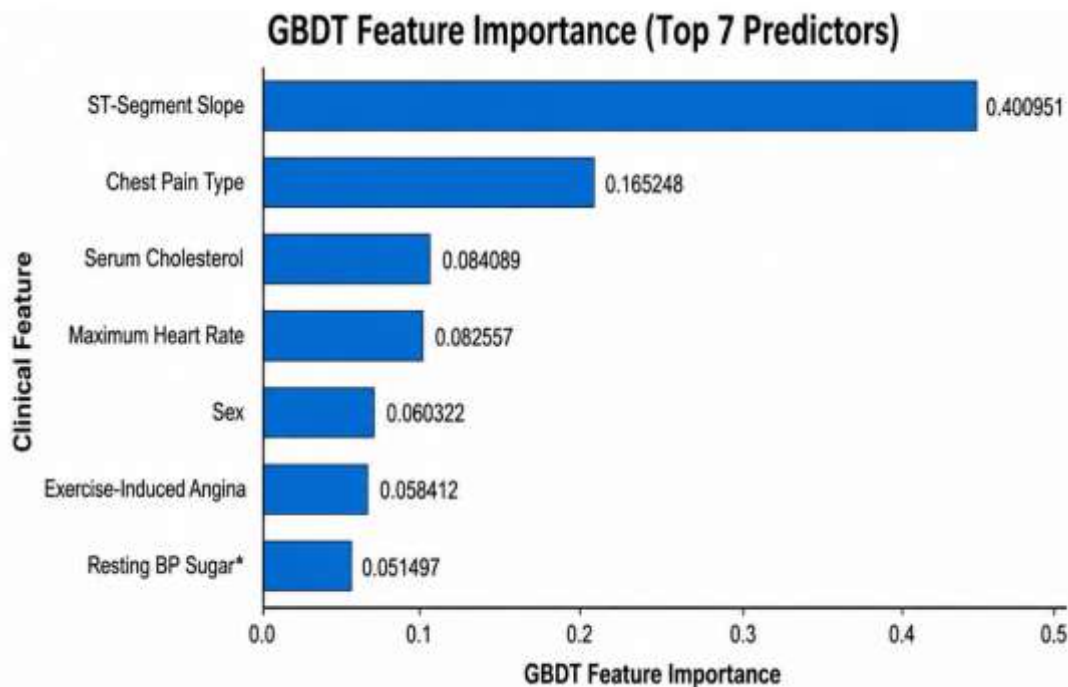


Fig. 4. GBDT-derived feature-importance ranking of the seven selected clinical predictors.

ST-segment slope had the highest importance score (0.400951), followed by chest pain type (0.165248). Serum cholesterol and maximum heart rate had intermediate scores of 0.084089 and 0.082557, respectively, while sex, exercise-induced angina and Resting BP Sugar had lower but non-negligible scores. The resulting seven-feature representation was supplied to the four base classifiers of the stacking ensemble. Feature importance here represents the contribution of a predictor to the trained GBDT model and should be interpreted as a model-derived measure of predictive contribution, not as evidence of a causal relationship with heart disease.

Two-level stacking ensemble architecture

The two-level stacking ensemble consists of four base classifiers (RF, DT, KNN and LR) and one LR meta-learner (Fig. 5). For an input feature vector $\mathbf{x}$, the base learners produce the probability estimates

$$p_{RF} = f_{RF}(\mathbf{x}), p_{DT} = f_{DT}(\mathbf{x}), p_{KNN} = f_{KNN}(\mathbf{x}), p_{LR} = f_{LR}(\mathbf{x}) \quad (5)$$

These outputs form the meta-feature vector $\mathbf{z}(\mathbf{x}) = [p_{RF}, p_{DT}, p_{KNN}, p_{LR}]$, which is passed to the meta-learner g to give the final prediction $\hat{y} = g(\mathbf{z}(\mathbf{x}))$. To prevent data leakage, five-fold cross-validation was used to generate out-of-fold (OOF) predictions on the training set for training the meta-learner, and the base classifiers were then retrained on the full training set for test evaluation (Fig. 6).

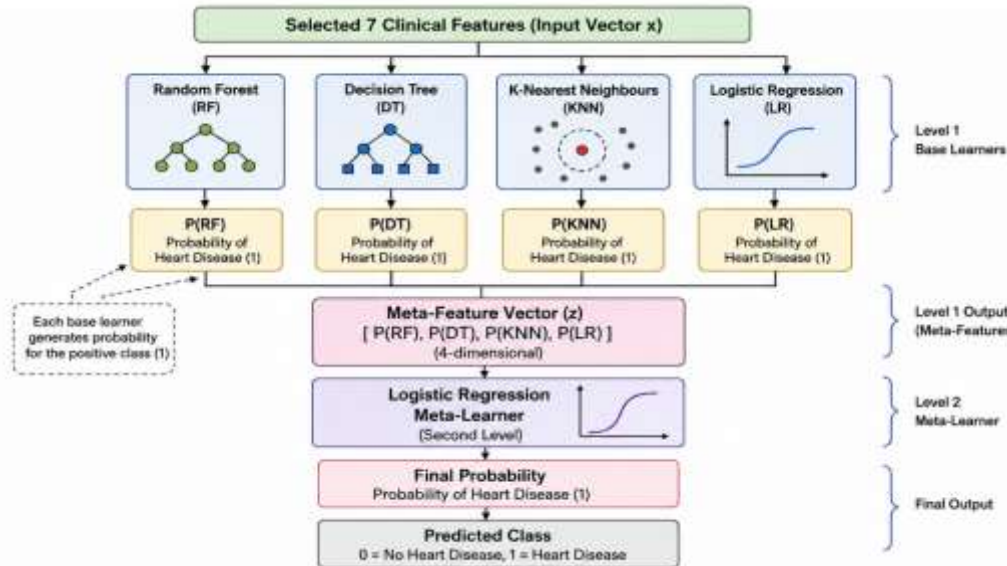


Fig. 5. Two-level stacking ensemble architecture using four base classifiers and a Logistic Regression meta-learner.

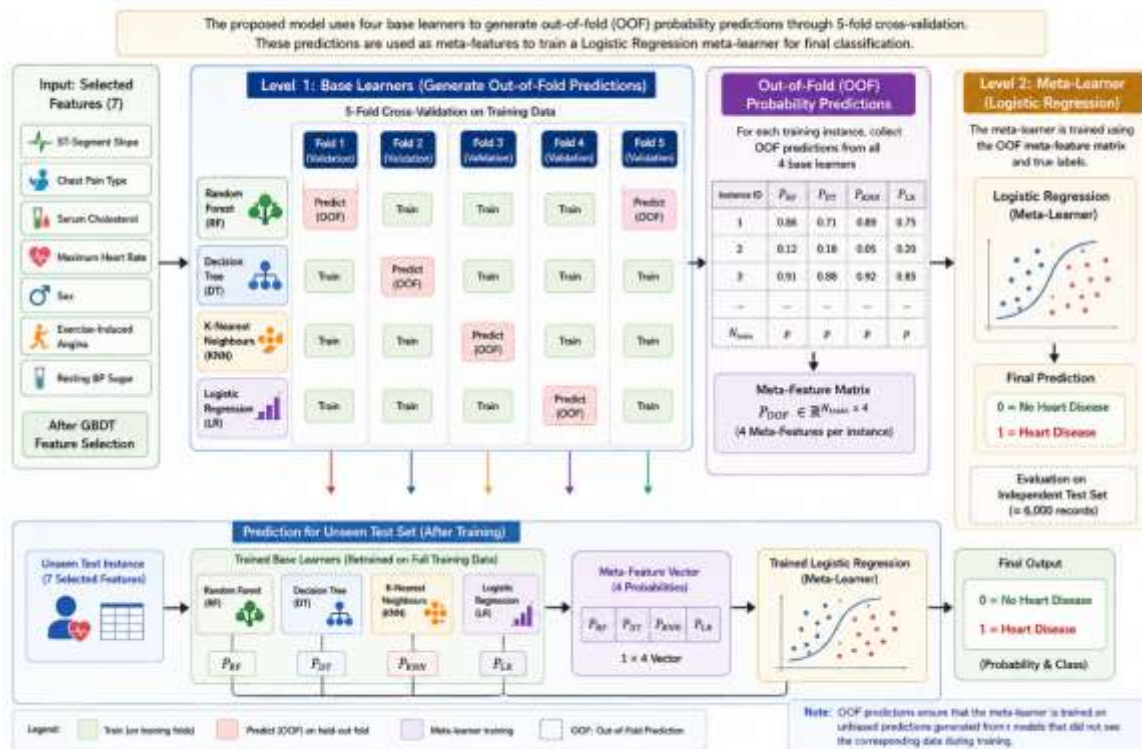


Fig. 6. Two-level stacking ensemble architecture with five-fold cross-validation and out-of-fold probability predictions for final heart disease classification.

Training strategy using five-fold cross-validation

Five-fold cross-validation was used to produce OOF probability predictions for the training data in order to minimize information leakage between the base classifiers and the meta-learner (Fig. 7). The training subset was divided into five mutually exclusive folds. In each iteration, the four base classifiers were trained on four folds, and the remaining fold was used only to produce probability predictions. For every observation in the held-out fold, the four trained base classifiers generated probability estimates for the positive class, which were stored as OOF meta-features. The procedure was repeated until every training observation had received probability predictions from base classifiers that had not been trained on that observation. The resulting OOF prediction matrix is

$$\mathbf{P}_{\text{OOF}} \in \mathbb{R}^{N_{\text{train}} \times 4} \quad (6)$$

where N_{train} is the number of training observations and the four columns correspond to the probability outputs of RF, DT, KNN and LR.

The LR meta-learner was then trained on the OOF prediction matrix $\mathbf{P}_{\text{OOF}}$ and the corresponding true class labels. After the meta-learner had been trained, the four base classifiers were retrained on the complete training subset. During the final evaluation, the retrained base classifiers generated probability predictions for the held-out test subset.

This two-level training strategy allows the meta-learner to learn the relationship between the probability estimates of the individual base classifiers without using predictions from models trained on the same observations. It therefore provides a structured approach to probability-level fusion while reducing the risk of information leakage during stacking.

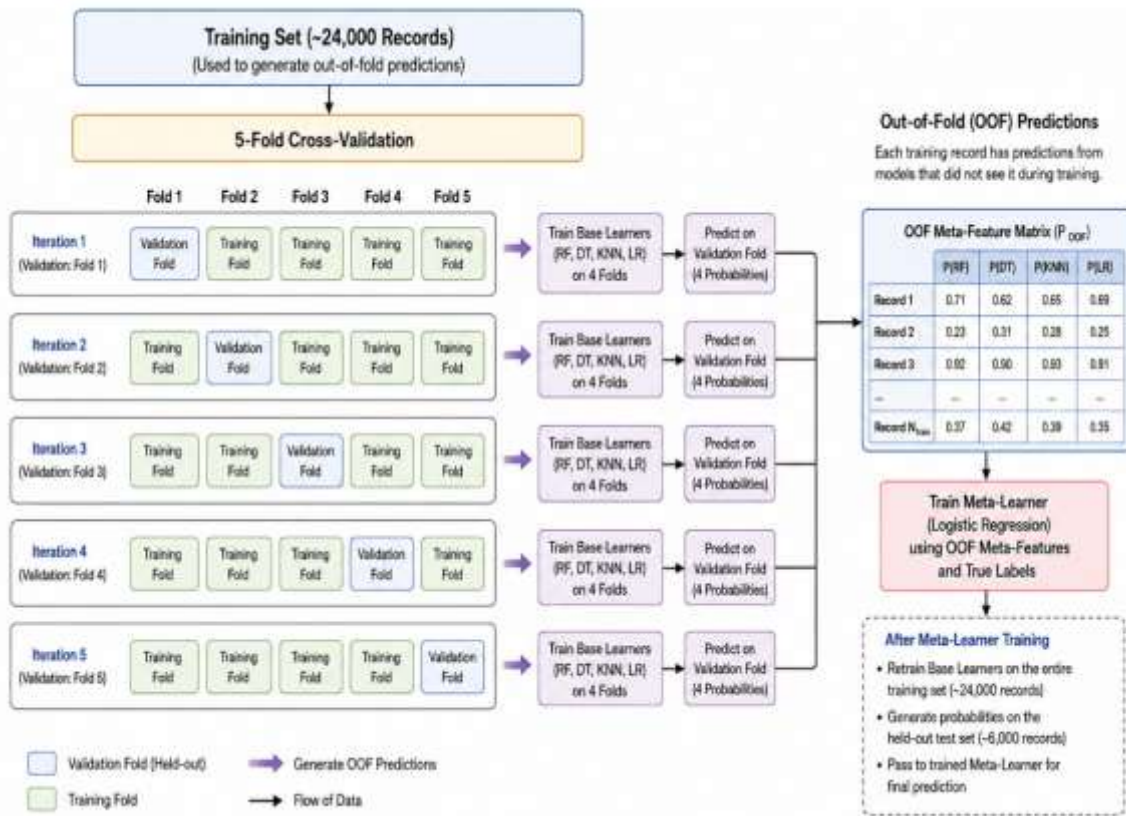


Fig. 7. Five-fold out-of-fold prediction strategy used to train the meta-learner in the proposed stacking ensemble.

Final prediction procedure

For an unseen patient record, the seven selected clinical features are first supplied to the four trained base classifiers. Each classifier independently produces a probability estimate for the positive (heart disease) class. These four estimates are concatenated to form the meta-feature vector

$$\mathbf{z}(\mathbf{x}) = [\mathbf{p}_{\text{RF}}, \mathbf{p}_{\text{DT}}, \mathbf{p}_{\text{KNN}}, \mathbf{p}_{\text{LR}}] \quad (7)$$

The trained LR meta-learner receives this meta-feature vector and generates the final probability and the predicted class. The complete prediction process can therefore be summarized as

$$\mathbf{x} \rightarrow \{\text{RF}(\mathbf{x}), \text{DT}(\mathbf{x}), \text{KNN}(\mathbf{x}), \text{LR}(\mathbf{x})\} \rightarrow \mathbf{z}(\mathbf{x}) \rightarrow \text{LR}_{\text{meta}} \rightarrow \hat{y} \quad (8)$$

This procedure bases the final decision on the combined probability estimates of all four heterogeneous base classifiers rather than on the prediction of a single model.

Experimental setup and hyperparameter configuration

The framework was built with the configurations specified for the synthetic data augmentation, GBDT-based feature selection, base classifiers and LR meta-learner. For feature selection, the GBDT model was configured with 200 estimators and a learning rate of 0.05. The RF base learner used 100 estimators, and the DT used the Gini criterion. The KNN classifier used five neighbours and the Euclidean distance metric. Both the base LR classifier and the LR meta-learner used the LBFGS solver. Synthetic data augmentation used bootstrap sampling with controlled noise injection ($\sigma = 0.05$) and a target dataset size of about 30,000 observations. The complete configuration is summarized in Table 6.

Table 6. Hyperparameter settings for all model components.

Component	Hyperparameter	Value	Selection method
Random Forest (RF)	n_estimators / max_depth	100 / None	Default (no tuning)
Decision Tree (DT)	criterion / max_depth	Gini / None	Default
K-Nearest Neighbours (KNN)	n_neighbors / metric	5 / Euclidean	Default
Logistic Regression (base)	C / solver	1.0 / lbfgs	Default
GBDT (feature selection)	n_estimators / learning_rate	200 / 0.05	Grid search on validation fold
Logistic Regression (meta)	C / solver	1.0 / lbfgs	Default
Augmenter	σ / M	0.05 / 30,000	KS-statistic validation

Cross-validation and model evaluation

Five-fold cross-validation was used throughout model building to evaluate model stability and to produce OOF predictions for the stacking architecture. The training data were divided into five mutually exclusive folds; in each iteration, four folds were used for training and the remaining fold for validation, until every fold had been used once for validation. The OOF probability predictions of the base learners were used as second-level input to train the LR meta-learner, and the final ensemble was assessed on the held-out test set after cross-validation.

The augmented dataset was split 80:20 into about 24,000 training and 6,000 test observations. Model performance was evaluated with accuracy, precision, recall, F1-score and ROC-AUC, which together describe the overall classification performance and the ability of the model to distinguish between patients with and without heart disease.

Ablation analysis of base classifiers

To assess the contribution of the ensemble architecture, the four base classifiers of the stacking framework (RF, DT, KNN and LR) were also evaluated independently, using the same selected features and the same five-fold cross-validation procedure as the proposed stacking framework.

The purpose of this analysis was to separate the predictive performance of the individual learning algorithms from the performance obtained after probability-level fusion. RF and DT represent tree-based learning, KNN represents instance-based learning and LR is a linear probabilistic classifier. In the proposed architecture, their probability outputs are combined by the LR meta-learner.

For a fair comparison, the individual classifiers and the stacking ensemble were evaluated with the same performance measures (accuracy, precision, recall, F1-score and ROC-AUC). This comparison provides an empirical basis for judging whether the combination of heterogeneous base learners adds

predictive value compared with the individual classifiers. The results of this comparison are reported in the Results section.

Implementation details

The framework was implemented in Python 3.10 with the Scikit-learn 1.4 machine learning library. The configurations of the base learners, GBDT feature selection model, meta-learner and synthetic data augementer are given in Table 6.

Results

GBDT feature importance ranking

Table 5 and Fig. 4 give the importance scores of the seven clinical predictors retained by GBDT. ST-segment slope was the most influential predictor (0.400951), followed by chest pain type (0.165248), so the selected feature subset was dominated by ST-segment slope and chest pain information. These values represent model-derived predictive contribution and should not be interpreted as evidence of an independent causal relationship with heart disease.

Cross-validation performance

The proposed stacking ensemble showed highly consistent performance across the five cross-validation folds (Table 7). The fold accuracies were 99.96%, 99.96%, 100.00%, 99.98% and 99.94%, giving a mean accuracy of 99.97% with a standard deviation of about 0.02 percentage points. The small variation between folds indicates stable performance under the adopted cross-validation procedure. These results should be interpreted within the benchmark and synthetic-data setting, because the training data came from the augmented dataset rather than from an independent clinical cohort.

Table 7. Accuracy of the stacking ensemble in five-fold cross-validation.

Fold	Accuracy (%)
Fold 1	99.96
Fold 2	99.96
Fold 3	100.00
Fold 4	99.98
Fold 5	99.94
Mean	99.97
Standard deviation	0.02

The very small differences between folds show that the model is stable and learned consistent patterns, which supports the combination of synthetic data augmentation, GBDT feature selection and the stacking ensemble architecture. After training, the model was tested on the held-out test set of 6,000 records (20% of the data). The test results are given in Table 8.

Model evaluation on the held-out test set

On the unseen test set of 6,000 records, the stacking ensemble achieved an accuracy of 99.97%, a weighted precision of 99.98%, a recall of 99.98% and an F1-score of 99.98% (Table 8).

Table 8. Class-wise performance of the proposed stacking ensemble model on the held-out test set.

Class	Precision (%)	Recall (%)	F1-score (%)	Support
Class 0 (no disease)	100.00	99.96	99.98	2823
Class 1 (disease)	99.97	100.00	99.98	3177
Weighted average	99.98	99.98	99.98	6000
Overall accuracy	99.97			

The confusion matrix (Fig. 8) shows the numbers of true positives, true negatives, false positives and false negatives. Only one of the 6,000 test records was misclassified: one healthy individual was

predicted to have heart disease (a false positive), and there were no false negatives. A low false-positive rate is important in medical practice, because a false-positive result leads to unnecessary tests and wasted care resources.

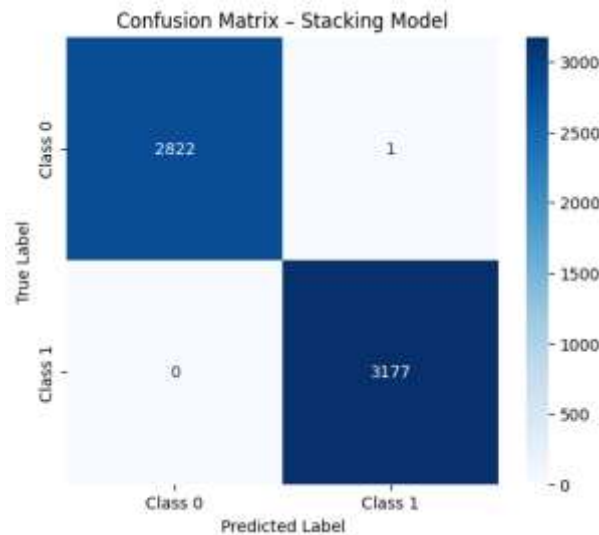


Fig. 8. Confusion matrix of the stacking ensemble model on the held-out test set.

A receiver operating characteristic (ROC) curve was used to examine the discriminative ability of the model further. The ROC curve shows the relationship between the true positive rate and the false positive rate over a range of decision thresholds and therefore gives a fuller picture than accuracy alone. For the proposed ensemble (Fig. 9), the area under the ROC curve was 1.000 for both classes, indicating near-perfect separation of patients with heart disease from healthy individuals.

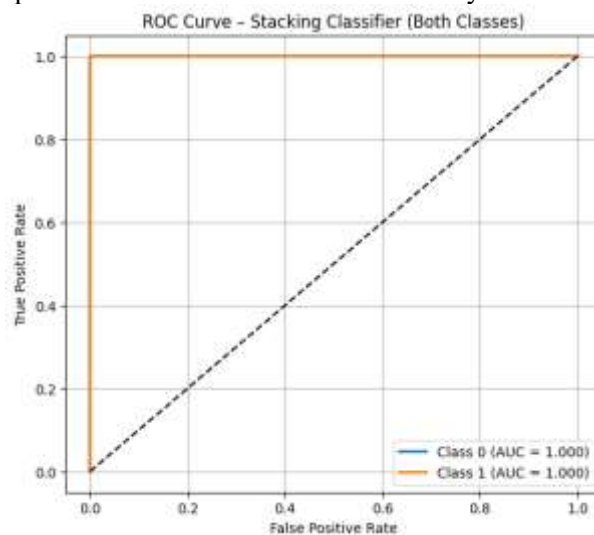


Fig. 9. ROC curves of the proposed stacking ensemble model.

Comparison with individual base classifiers

Table 9 compares the stacking ensemble with its four base classifiers used alone, all evaluated on the same seven selected features with the same five-fold cross-validation procedure.

Table 9. Performance of the individual base classifiers and the proposed stacking ensemble.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	ROC-AUC
RF					
DT					
KNN					

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	ROC-AUC
LR					
Stacking ensemble	99.97	99.98	99.98	99.98	1.000

Explainable AI analysis

SHapley Additive exPlanations (SHAP) were used to explain the model predictions [22]. The SHAP summary plot (Fig. 10) gives a global view of how the selected clinical variables contribute to individual predictions. Features ranked higher contribute more strongly to the model output, and the distribution of SHAP values shows whether individual observations increase or decrease the predicted probability of heart disease. The SHAP analysis complements the GBDT feature selection by explaining feature contributions at the level of predictions rather than relying only on the feature-importance ranking.

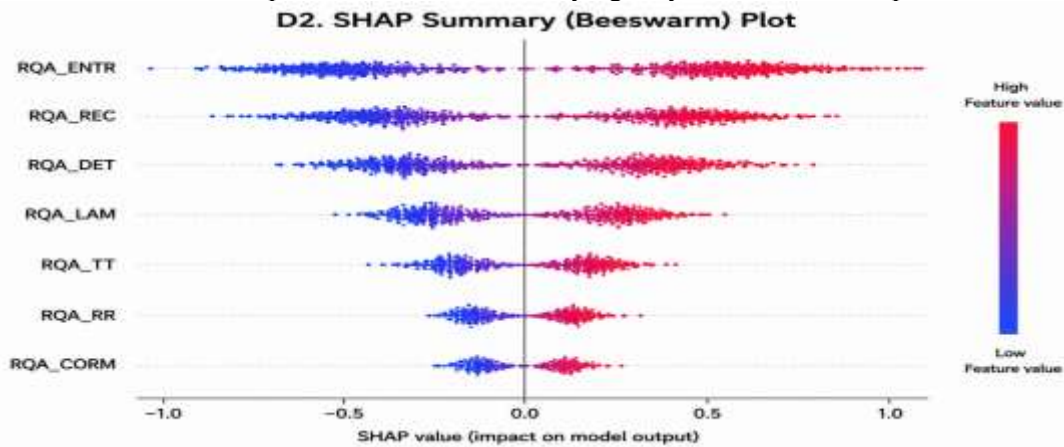


Fig. 10. SHAP summary plot showing the global contribution and direction of the selected clinical features for the proposed heart disease classifier.

The SHAP dependence plot (Fig. 11) examines the relationship between an individual clinical feature and its contribution to the model prediction. The vertical axis shows the magnitude and direction of the feature contribution, and the horizontal axis shows the observed feature values. The distribution of the observations shows how changes in the selected predictor are associated with the predicted probability. This analysis describes model behaviour and does not establish a causal clinical relationship.

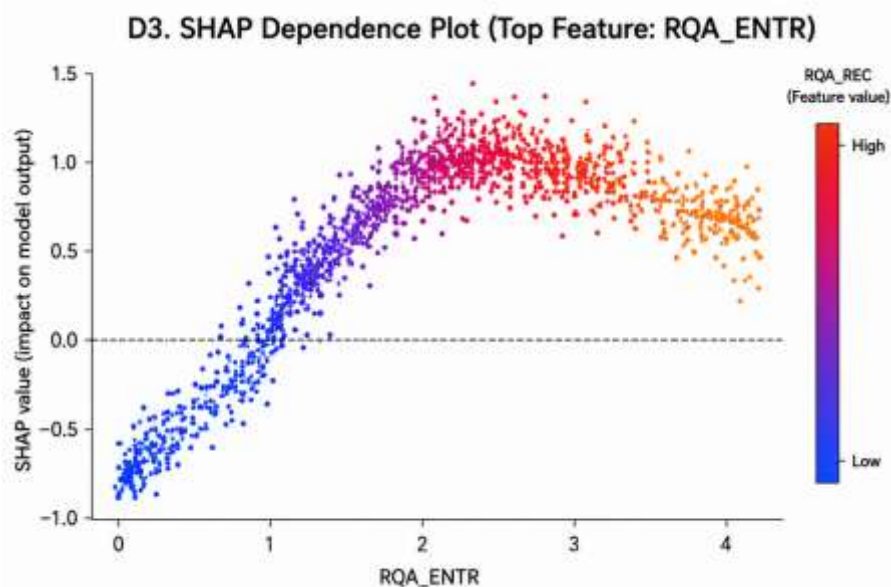


Fig. 11. SHAP dependence plot showing the relationship between the selected clinical feature values and their contribution to the predicted heart disease probability.

The local SHAP explanation (Fig. 12) shows how the selected clinical variables contributed to an individual prediction. Features with positive SHAP values move the prediction towards the heart disease class, whereas negative values move it towards the no-disease class. This local explanation provides case-level transparency and shows how the proposed framework can reveal the contribution of individual predictors to a particular decision.

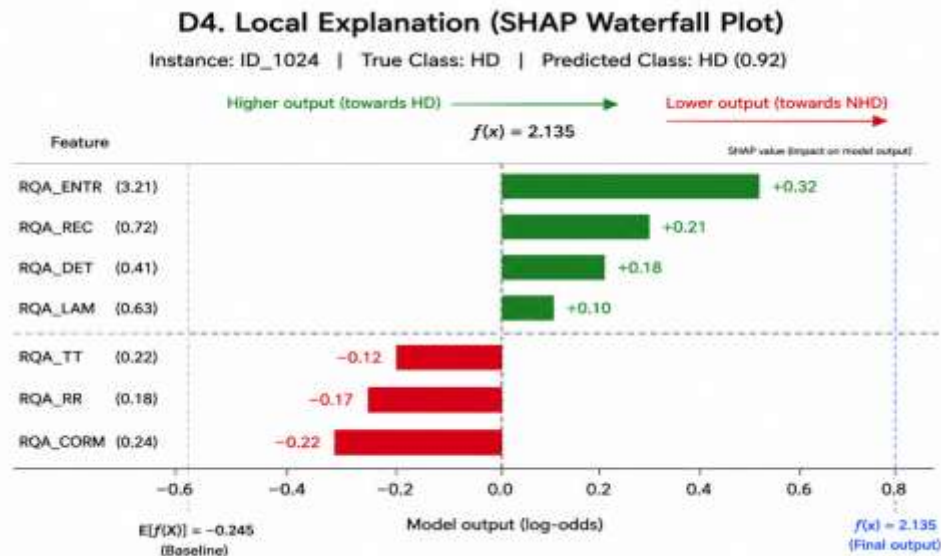


Fig. 12. Local SHAP explanation for an individual test observation showing the features that contributed to the final heart disease prediction.

Discussion

The proposed model was compared with earlier heart disease classification studies from Table 1, and the accuracies are given in Table 10. Mahmud et al. [1], who used the same five benchmark datasets, reported an accuracy of 87% with a meta-model combining RF, Gaussian NB, DT and KNN. Plati et al. [13] reached 91.23% with Rotation Forest, and Alshraideh et al. [19] reported the highest accuracy among the compared studies, 94.3%, using SVM with PSO.

Table 10. Performance comparison of the proposed model with previous approaches.

Study	Method	Data	Accuracy (%)
[1]	Meta-model (RF, GNB, DT, KNN)	Five benchmark datasets	87.00
[13]	Rotation Forest	Hospital data, Ireland and Greece	91.23
[19]	SVM + PSO	JUH dataset, Jordan	94.30
This study	Stacking ensemble (RF, DT, KNN, LR; LR meta-learner)	Five benchmark datasets + synthetic augmentation	99.97

The proposed stacking ensemble achieved an accuracy of 99.97%, which is higher than all of these values. This improvement is attributed to the combination of several data sources, synthetic data augmentation, GBDT-based feature selection and the probability-level fusion of heterogeneous base learners, which confirms the effectiveness of the proposed stacking ensemble approach. Because the compared studies used different datasets and evaluation protocols, and the present results were obtained on augmented benchmark data, this comparison is indicative, and external validation is needed to confirm the advantage in clinical settings.

Conclusion

In this work, an optimized stacking ensemble model was developed with RF, KNN, DT and LR as base learners and LR as the meta-learner. Several benchmark datasets were pooled, synthetic data were generated to address the small sample size, GBDT was used for feature selection and five-fold cross-validation was used to improve robustness. The proposed stacking ensemble achieved an accuracy of 99.97%, precision, recall and F1-score of nearly 100% for both classes and a ROC-AUC of 1.000. This is a clear improvement over the highest accuracy among the compared studies, 94.3% [19]. The model was very stable, with a very small standard deviation across the cross-validation folds, and it showed high discriminative power and a low false-positive rate, which are valuable characteristics for a clinical decision support system, where sensitivity and specificity are of the utmost importance.

Future studies could test the model on larger and broader real-time clinical data from different healthcare settings to confirm its generalizability, include additional data on lifestyle, genetic markers and medical imaging to give clinicians a more detailed explanation of each prediction, and implement the model as a real-time clinical decision support system integrated with electronic health records.

References

- [1] I. Mahmud, M.M. Kabir, M.F. Mridha, S. Alfarhood, M. Safran, D. Che: *Diagnostics*, 13 (2023) 2540.
- [2] S.S. Martin, A.W. Aday, Z.I. Almarzooq, C.A.M. Anderson, P. Arora, C.L. Avery, et al.: *Circulation*, 149 (2024) e347-e913.
- [3] M.A. Bouqentar, O. Terrada, S. Hamida, S. Saleh, D. Lamrani, B. Cherradi, A. Raihani: *Heliyon*, 10 (2024) e38731.
- [4] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, Y. Wang: *Stroke Vasc Neurol*, 2 (2017) 230-243.
- [5] A.A. Ahmad, H. Polat: *Diagnostics*, 13 (2023) 2392.
- [6] A.M. Qadri, M.S.A. Hashmi, A. Raza, S.A.J. Zaidi, A. ur Rehman: *PeerJ Comput Sci*, 10 (2024) e1894.
- [7] S. Secinaro, D. Calandra, A. Secinaro, V. Muthurangu, P. Biancone: *BMC Med Inform Decis Mak*, 21 (2021) 125.
- [8] H. Thomas, J. Diamond, A. Vieco, S. Chaudhuri, E. Shinnar, S. Cromer, et al.: *Glob Heart*, 13 (2018) 143-163.
- [9] N.S. Mansur Huang, Z. Ibrahim, N. Mat Diah: *Malays J Comput*, 6 (2021) 872-884.
- [10] Z. Noroozi, A. Orooji, L. Erfannia: *Sci Rep*, 13 (2023) 22588.
- [11] S.V. Remya: *Indian J Sci Technol*, 17 (2024) 3960-3968.
- [12] D. Chicco, G. Jurman: *BMC Med Inform Decis Mak*, 20 (2020) 16.
- [13] D.K. Plati, E.E. Tripoliti, A. Bechlioulis, A. Rammos, I. Dimou, L. Lakkas, et al.: *Diagnostics*, 11 (2021) 1863.
- [14] R.P. Ram Kumar, S. Polepaka, V. Manasa, D. Palakurthy, E. Annapoorna, N. Dhaliwal, H. Dhall, L.H. Alzubaidi: *Cogent Eng*, 11 (2024) 2386381.
- [15] J.K. Mutinda, A.K. Langat: *Asian J Res Comput Sci*, 17 (2024) 157-166.
- [16] G. Logabiraman, D. Ganesh, M.S. Kumar, A.V. Kumar, N. Bhardwaj: *MATEC Web Conf*, 392 (2024) 01122.
- [17] H. Jindal, S. Agrawal, R. Khera, R. Jain, P. Nagrath: *IOP Conf Ser: Mater Sci Eng*, 1022 (2021) 012072.
- [18] C.M. Bhatt, P. Patel, T. Ghetia, P.L. Mazzeo: *Algorithms*, 16 (2023) 88.
- [19] M. Alshraideh, N. Alshraideh, A. Alshraideh, Y. Alkayed, Y. Al Trabsheh, B. Alshraideh: *Appl Comput Intell Soft Comput*, 2024 (2024) 5080332.
- [20] E. Ibrahim, M.B. Lopes, X. Dharmo, A. Simeon, R. Shigdel, K. Hron, et al.: *Front Microbiol*, 14 (2023) 1250909.
- [21] S. Rao, P. Poojary, J. Somaiya, P. Mahajan: *Int J Eng Appl Sci Technol*, 5 (2020) 431-438.
- [22] H. El-Sofany, B. Bouallegue, Y.M. Abd El-Latif: *Sci Rep*, 14 (2024) 23277.