Proceedings of the Second International Conference on Multi-Agent Systems for Collaborative Intelligence (ICMSCI-2026)
IEEE Xplore Part Number: CFP26VH0-ART; ISBN: 979-8-3315-6898-6

A Cost-Aware Stacking Ensemble Framework for Early Heart Disease Detection

Jhansi B
Department of Computer Science, School of Computing Sciences, Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai, India
jhansisathish.bcn@gmail.com
ORCID: 0009-0004-5243-4422

Dr. T. Kamalakannan
Department of DACE, School of Computing Sciences, Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai, India
kkannan.scs@velsuniv.ac.in
ORCID: 0000-0002-5982-8983

2026 Second International Conference on Multi-Agent Systems for Collaborative Intelligence (ICMSCI) | 979-8-3315-6898-6/26/$31.00 ©2026 IEEE | DOI: 10.1109/ICMSCI67830.2026.11469762

***Abstract*—Heart disease remains a leading cause of mortality worldwide, emphasizing the need for reliable and early screening mechanisms. Although machine learning techniques have been widely explored for heart disease prediction, many existing approaches primarily optimize overall accuracy and provide limited attention to misclassification cost asymmetry and interpretability, which are critical for clinical screening applications. This paper proposes a cost-aware stacking ensemble framework for early heart disease detection that integrates heterogeneous base classifiers, optimized decision thresholding, and explainable artificial intelligence techniques within a unified pipeline. Logistic regression, support vector machine, and random forest models are employed as base learners, and their outputs are combined using a logistic regression–based meta-learner. Cost-sensitive learning is incorporated to prioritize reducing false negatives, while the final classification threshold is optimized using receiver operating characteristic analysis. Model transparency is enhanced through Shapley value–based local explanations and permutation-based global feature importance analysis. Experimental evaluation on the UCI Heart Disease dataset using stratified cross-validation demonstrates that the proposed framework achieves an accuracy of 87.17%, a sensitivity of 0.92, a specificity of 0.84, and an area under the ROC curve of 0.87. The results indicate that the proposed approach provides a robust and interpretable decision-support system suitable for early heart disease screening.**

Keywords—clinical decision support, cost-sensitive learning, explainable artificial intelligence, Heart disease prediction, stacking ensemble.

I. INTRODUCTION

Cardiovascular diseases remain one of the leading causes of mortality worldwide, accounting for a significant proportion of premature deaths and long-term disability. Among these, heart disease poses a major public health challenge due to its multifactorial nature and often asymptomatic progression during early stages. Timely identification of individuals at risk is therefore critical for initiating preventive interventions and reducing adverse clinical outcomes. Conventional diagnostic approaches rely heavily on clinical expertise, laboratory investigations, and imaging procedures, which can be costly, invasive, and time-consuming, particularly in large-scale screening scenarios [1].

The increasing availability of electronic health records and structured clinical datasets has enabled the application of machine learning techniques to assist in heart disease prediction. By learning patterns from historical patient data, machine learning models can help clinicians more efficiently and consistently identify high-risk individuals. Early studies employing traditional classifiers such as logistic regression, decision trees, and support vector machines demonstrated the feasibility of data-driven heart disease prediction [2][3].

However, the predictive capability of single models is often constrained by limited generalizability, sensitivity to data characteristics, and an inability to capture the complex nonlinear relationships inherent in clinical data. To address these limitations, ensemble learning methods have been increasingly adopted in cardiovascular disease prediction. Techniques such as random forests and boosting algorithms combine multiple learners to improve robustness and classification performance [4]. While ensemble approaches generally outperform individual classifiers, many existing studies focus predominantly on maximizing overall accuracy. In medical screening contexts, such an objective is insufficient, as false negatives—cases where diseased patients are incorrectly classified as healthy—can have serious clinical consequences. Consequently, models optimized solely for accuracy may fail to meet the practical requirements of early disease detection.

Another important limitation of many existing machine learning–based diagnostic systems is the lack of interpretability. Complex ensemble and nonlinear models are often treated as black boxes, providing limited insight into the factors driving their predictions. This lack of transparency poses a significant barrier to clinical adoption, where trust, accountability, and alignment with medical knowledge are essential [5]. Recent advances in explainable artificial intelligence have sought to address this issue by introducing methods such as Shapley value–based explanations, which quantify feature contributions to individual predictions [6]. However, explainability is often used as a post hoc analysis rather than integrated into the model development process.

More recently, stacking ensemble learning has emerged as a powerful technique that combines the strengths of multiple base classifiers through a meta-learning framework. By learning how to optimally fuse model outputs, stacking ensembles have demonstrated improved generalization performance in various healthcare applications [7][8]. Despite their potential, existing stacking-based heart disease prediction models often neglect clinically relevant considerations such as misclassification cost asymmetry and decision threshold optimization. Furthermore, the integration of stacking ensembles with explainability mechanisms remains limited in current literature.

Motivated by these observations, this work proposes a cost-aware and explainable stacking ensemble framework for early heart disease screening. The proposed approach

explicitly prioritizes sensitivity through cost-sensitive learning, optimizes the decision threshold using receiver operating characteristic analysis, and incorporates model-agnostic explainability to enhance transparency. By systematically combining performance optimization and interpretability within a unified framework, the proposed method aims to bridge the gap between predictive accuracy and clinical applicability.

The main contributions of this work are summarized as follows:

1. A cost-aware stacking ensemble framework is proposed for early heart disease screening by integrating heterogeneous machine learning classifiers through a meta-learning strategy.
2. Cost-sensitive learning and ROC-guided decision threshold optimization are incorporated to prioritize sensitivity and reduce false-negative diagnoses in clinical screening scenarios.
3. An explainability-driven analysis is integrated using Shapley value–based local explanations and permutation-based global feature importance to enhance model transparency and clinical interpretability.

The remainder of this paper is organized as follows. Section 2 reviews related work on heart disease prediction, ensemble learning, and explainable artificial intelligence. Section 3 describes the proposed methodology in detail. Section 4 outlines the experimental setup and evaluation protocol. Section 5 presents and discusses the experimental results. Finally, Section 6 concludes the paper and outlines directions for future research.

II. Literature Review

Heart disease remains one of the leading causes of morbidity and mortality worldwide, necessitating reliable and early diagnostic mechanisms. Traditional clinical diagnostic procedures rely heavily on physician expertise, invasive tests, and subjective interpretation of patient data, which can be time-consuming and prone to variability. With the rapid growth of digital healthcare records and computational power, machine learning (ML) techniques have emerged as effective tools for assisting clinicians in the early detection of heart disease by learning complex patterns from historical medical data.

Initial research efforts in heart disease prediction primarily focused on conventional machine learning classifiers, including logistic regression, decision trees, k-nearest neighbors, and naive Bayes. Logistic regression has been widely adopted due to its simplicity, interpretability, and statistical grounding, particularly in binary medical classification tasks [2]. However, its linear decision boundary restricts its ability to capture complex nonlinear relationships present in clinical data.

Decision tree–based models gained popularity as they provide rule-based explanations that align with human reasoning [4]. Despite their interpretability, decision trees are highly sensitive to data variations and often suffer from overfitting, leading to unstable predictions. Similarly, k-nearest neighbor classifiers were explored for heart disease diagnosis due to their nonparametric nature, but their performance heavily depends on distance metrics and feature scaling, limiting their robustness.

Support vector machines (SVMs) were introduced to address nonlinear separability via kernel functions and have been shown to improve classification accuracy compared to linear models [9]. Nevertheless, SVMs require careful parameter tuning and provide limited transparency, which poses challenges for clinical adoption.

While single-model approaches demonstrated moderate success, their limited generalization capability and sensitivity to data characteristics highlighted the need for more robust predictive frameworks. To overcome the shortcomings of individual classifiers, ensemble learning techniques were introduced to combine multiple learners, improving predictive stability and accuracy. Bagging-based approaches, particularly random forests, became widely used in medical prediction tasks due to their ability to reduce variance and handle feature interactions effectively [10]. Random forests have consistently outperformed single classifiers on the UCI Heart Disease dataset, making them a strong baseline in cardiovascular risk prediction studies.

Boosting algorithms such as AdaBoost and Gradient Boosting Machines further improved predictive performance by sequentially focusing on misclassified samples [11]. These methods demonstrated higher accuracy but increased sensitivity to noise and class imbalance, which are common in medical datasets. Moreover, boosting models often lack interpretability, which limits their adoption in clinical decision-making.

Voting-based ensemble methods that combine heterogeneous classifiers, such as logistic regression, SVMs, and decision trees, were proposed to leverage the complementary strengths of different models [12]. While majority-voting and weighted-voting ensembles improved accuracy, most studies relied on heuristic weight assignment and optimized only overall accuracy. This approach neglects the asymmetric costs associated with medical misclassification, particularly the severe consequences of false negatives in heart disease screening.

Although ensemble learning improved performance compared to single classifiers, many ensemble-based studies remained focused on accuracy metrics without addressing clinical priorities such as sensitivity and interpretability. Medical diagnosis inherently involves asymmetric risk, where false negatives can lead to delayed treatment and severe outcomes. Cost-sensitive learning addresses this issue by assigning different penalties to different types of misclassifications [3]. Several studies have applied cost-sensitive techniques to cardiovascular disease prediction, demonstrating improved sensitivity at the expense of a marginal loss in accuracy. However, existing cost-sensitive approaches are often used to single classifiers or simple ensembles. The integration of cost sensitivity within stacking ensemble frameworks remains underexplored. Moreover, many studies adopt fixed probability thresholds, which may not be optimal under varying cost conditions.

Most ML-based heart disease prediction systems employ a default decision threshold of 0.5 for binary classification. This assumption is often inappropriate for medical screening applications, where maximizing sensitivity is more critical than maximizing accuracy [13][14]. Threshold optimization techniques based on ROC analysis, such as the Youden index,

provide a principled approach for balancing sensitivity and specificity. Despite its relevance, threshold optimization has received limited attention in ensemble-based studies of heart disease prediction. Many works report ROC curves and AUC values, but do not explicitly optimize the operating point for clinical screening scenarios.

From the existing literature, it is evident that:

1. Single classifiers lack robustness and generalization.
2. Ensemble methods improve performance but often ignore clinical cost asymmetry.
3. Stacking ensembles enhances predictive capability but lacks cost awareness and threshold optimization.
4. Explainability is typically treated as an isolated component rather than an integral part of the predictive framework.

There is a clear research gap in developing a unified framework that integrates stacking ensemble learning, cost-sensitive optimization, decision-threshold tuning, and explainability for early heart disease screening. This gap forms the basis for the proposed work.

III. Methodology

This section describes the proposed heart disease prediction framework, which is designed through systematic experimentation, optimization, and validation to achieve reliable screening performance. The methodology integrates data preprocessing, heterogeneous base learning, stacking-based ensemble fusion, cost-sensitive learning, optimized decision thresholding, and explainability analysis within a unified framework. The heart disease prediction problem is formulated as a binary classification task, in which the objective is to predict the presence or absence of heart disease from a set of clinical and demographic attributes. Let the dataset be represented as

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

where $\mathbf{x}_i \in \mathbb{R}^d$ denotes the d-dimensional feature vector of the i-th subject and $y_i \in \{0,1\}$ represents the class label, with 0 indicating no heart disease and 1 indicating the presence of disease. The original target values are binarized by grouping all non-zero severity levels into the positive class, which is consistent with standard early screening practices.

A. Data Preprocessing and Base Classifier Modeling

Before model training, the dataset undergoes a structured preprocessing pipeline to ensure robustness and numerical stability across classifiers. Categorical attributes are converted into numeric representations to enable compatibility with machine learning models. Missing values, which are common in clinical datasets, are handled using column-wise median imputation. Median imputation is preferred over mean imputation as it is less sensitive to outliers and skewed distributions typically observed in medical variables. To avoid dominance of features with larger numeric ranges and to ensure fair contribution across models, z-score normalization is applied to each feature as

$$x_{ij}^{\text{norm}} = \frac{x_{ij} - \mu_j}{\sigma_j},$$

where μ_j and σ_j denote the mean and standard deviation of the j -th feature, respectively. This normalization is particularly important for distance-based and margin-based classifiers such as support vector machines.

To capture diverse decision characteristics, three heterogeneous base classifiers are employed. Logistic regression is used as a linear probabilistic model that provides stable baseline performance and interpretability. Support vector machines with a radial basis function kernel are incorporated to model nonlinear decision boundaries in the feature space. Random forests are used as a tree-based ensemble to capture complex feature interactions and nonlinear relationships. Each base classifier produces an estimated posterior probability or decision score for the positive class. Let the outputs of the base classifiers for an instance $\mathbf{x}_i$ be denoted as

$$\mathbf{z}_i = [z_i^{(1)}, z_i^{(2)}, z_i^{(3)}],$$

corresponding to logistic regression, support vector machine, and random forest predictions, respectively.

B. Cost-Sensitive Learning Strategy and Stacking Ensemble Architecture

In medical screening applications, misclassification costs are inherently asymmetric. A false negative, where a diseased patient is classified as healthy, poses a significantly higher clinical risk than a false positive. To reflect this asymmetry, cost-sensitive learning is incorporated during model training. A misclassification cost matrix is defined as

$$\mathbf{C} = \begin{bmatrix} 0 & C_{FP} \\ C_{FN} & 0 \end{bmatrix}, \text{with } C_{FN} > C_{FP},$$

where C_{FN} and C_{FP} denote the costs associated with false negatives and false positives, respectively. This cost matrix is integrated into the training of applicable base classifiers, biasing the learning process toward minimizing false negatives and thereby improving sensitivity.

Rather than relying on fixed or heuristic voting schemes, a stacking ensemble architecture is adopted to learn an optimal combination of base classifier outputs. In the stacking framework, predictions from the base classifiers form a new feature space for a meta-level classifier. Logistic regression is employed as the meta-learner due to its stability and ability to provide calibrated probability estimates.

Formally, the meta-level prediction is given by

$$\hat{y}_i = \sigma(\mathbf{w}^\top \mathbf{z}_i + b),$$

where $\mathbf{w}$ and b are the meta-learner parameters and $\sigma(\cdot)$ denotes the logistic sigmoid function. The stacking model is trained using out-of-fold predictions generated through stratified cross-validation, ensuring that the meta-learner does not observe predictions from base models trained on the same data instances.

C. Cross-Validation Protocol and Decision Threshold Optimization

To obtain unbiased performance estimates and to ensure robustness, stratified 10-fold cross-validation is employed. The dataset is partitioned into ten mutually exclusive folds while preserving the class distribution in each fold. In each iteration, 9 folds are used for training, and 1 fold is reserved for testing.

Base classifier predictions are generated fold-wise and aggregated to train the stacking meta-learner. This protocol ensures strict separation between training and evaluation data at all stages of the ensemble construction. Most classification systems employ a fixed decision threshold of 0.5 to convert predicted probabilities into class labels. However, such a threshold may not be optimal for medical screening tasks. Therefore, the decision threshold is optimized using receiver operating characteristic (ROC) analysis. Let $\mathrm{TPR}(\tau)$ and $\mathrm{FPR}(\tau)$ denote the true positive rate and false positive rate at threshold τ. The optimal threshold $\tau^{\backslash *}$ is selected by maximizing the Youden index:

$$\tau^{\backslash *} = \arg\max_{\tau} \left(\mathrm{TPR}(\tau) - \mathrm{FPR}(\tau)\right).$$

This criterion ensures a balanced trade-off between sensitivity and specificity while prioritizing the correct detection of diseased cases.

D. Explainability and Feature Contribution Analysis

To enhance transparency and clinical trust, explainability is incorporated into the proposed framework. A model-agnostic Shapley value–based explanation technique is employed to analyze the contribution of individual features to model predictions. For a given prediction, the Shapley value of a feature represents its average marginal contribution across all possible feature subsets, providing a principled explanation grounded in cooperative game theory.

In addition to local explanations, global feature importance is estimated using permutation-based importance derived from the random forest model. This analysis identifies clinically relevant attributes that consistently influence predictions, supporting alignment between the model's behavior and established medical knowledge.

The overall architecture of the proposed framework is illustrated in Fig. 1. The input clinical dataset is first preprocessed, including encoding, imputation, and normalization. The processed data are then fed in parallel to the base classifiers—logistic regression, support vector machine, and random forest. The prediction scores generated by these classifiers are combined at the stacking layer, where a logistic regression meta-learner produces the final probability estimate. Cost-sensitive learning influences the training of base models, while threshold optimization refines the final decision boundary. An explainability module operates alongside the prediction pipeline to provide local and global interpretability of the results.

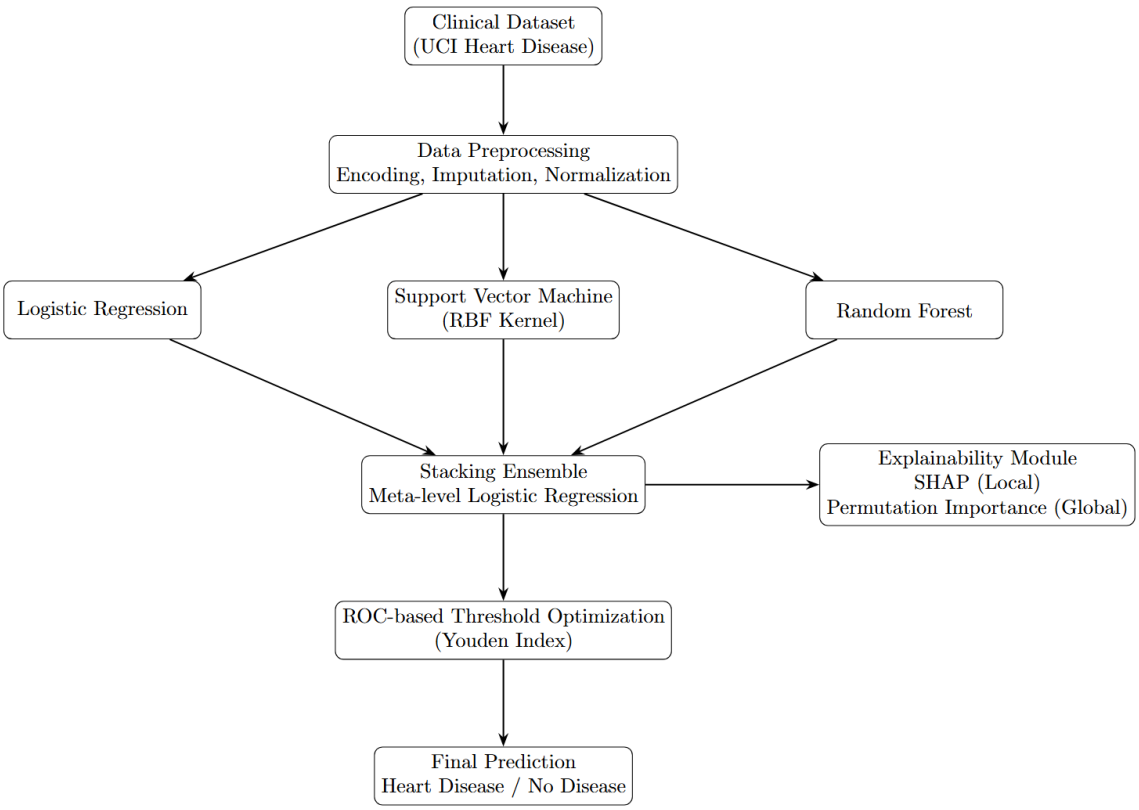

Fig. 1. Block Diagram

The proposed methodology systematically integrates stacking ensemble learning, cost-sensitive optimization, ROC-based threshold selection, and explainable artificial intelligence into a single framework. Each component is motivated by empirical evaluation and iterative refinement, resulting in a robust and clinically oriented heart disease prediction system.

IV. Experimental Setup

The experimental evaluation of the proposed framework is designed to ensure reproducibility, robustness, and clinical relevance. All experiments are conducted using the UCI Heart Disease dataset, which is a widely accepted benchmark in cardiovascular disease prediction studies and allows direct comparison with existing methods reported in the literature [4][5].

A. Dataset and Class Distribution

The UCI Heart Disease dataset consists of multiple clinical and demographic attributes, including age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, electrocardiographic results, maximum heart rate achieved, exercise-induced angina, and ST depression. The original multi-class target variable is transformed into a binary classification problem by grouping all non-zero disease severity levels into a single positive class. This binarization aligns with the objective of early screening, where the primary concern is identifying the presence of heart disease rather than estimating severity.

The resulting dataset exhibits mild class imbalance, which is typical of medical screening datasets. This imbalance further motivates the use of cost-sensitive learning and sensitivity-focused evaluation metrics.

B. Cross-Validation Strategy

To obtain unbiased and statistically reliable performance estimates, stratified 10-fold cross-validation is employed throughout the experiments. The dataset is partitioned into ten folds with approximately equal class proportions in each fold. In each iteration, nine folds are used for training, and the remaining fold is used for testing. This process is repeated until each fold has served as the test set exactly once.

For the stacking ensemble, out-of-fold predictions from the base classifiers are generated within the cross-validation framework and used to train the meta-level classifier. This procedure strictly prevents information leakage between training and testing phases and ensures that the reported results reflect true generalization performance.

C. Baseline and Comparative Models

To assess the effectiveness of the proposed framework, its performance is compared against individual base classifiers, including logistic regression, a support vector machine with a radial basis function kernel, and a random forest. These models represent commonly used baselines in heart disease prediction studies and provide a meaningful reference for evaluating the benefits of stacking, cost sensitivity, and threshold optimization. All baseline models are trained using the same preprocessing pipeline and cross-validation protocol to ensure a fair comparison.

D. Evaluation Metrics and Decision Threshold Selection

The performance of the proposed and baseline models is evaluated using multiple complementary metrics to reflect both statistical accuracy and clinical relevance. Overall classification accuracy is reported to provide a general measure of correctness. Sensitivity, also known as the true positive rate or recall, is emphasized because it quantifies the model's ability to correctly identify patients with heart disease, which is critical in screening applications. Specificity is reported to measure the correct identification of healthy individuals and to assess the false positive rate.

In addition, the receiver operating characteristic (ROC) curve is used to evaluate the trade-off between sensitivity and specificity across different decision thresholds. The area under the ROC curve (AUC) is reported as a threshold-independent metric that reflects the model's overall discriminative ability. A higher AUC indicates better separation between diseased and non-diseased cases.

Rather than using a fixed probability threshold of 0.5, the final class decision is obtained using an optimized threshold derived from ROC analysis. The Youden index is employed to identify the operating point that maximizes the difference between the true positive rate and the false positive rate. This threshold selection strategy is particularly suitable for medical screening scenarios, where maximizing sensitivity while maintaining reasonable specificity is essential.

All experiments are implemented in MATLAB R2023b using built-in machine learning and statistical toolboxes. The experiments are executed on a standard desktop computing environment. The use of MATLAB ensures numerical stability, reproducibility, and consistency across all stages of data preprocessing, model training, ensemble construction, and evaluation. The experimental setup combines stratified cross-validation, fair baseline comparison, sensitivity-oriented evaluation metrics, and optimized threshold selection to provide a rigorous and clinically meaningful assessment of the proposed framework. This setup ensures that the reported results accurately reflect the model's strengths and limitations in realistic early heart disease screening scenarios.

V. Results and Discussion

This section presents the experimental results obtained using the proposed cost-aware stacking ensemble framework and discusses their implications for early heart disease screening. The analysis integrates quantitative performance metrics, confusion matrix interpretation, ROC characteristics, and explainability outputs to provide both statistical and clinical insights.

A. Classification Performance Analysis

The proposed stacking ensemble achieves an overall classification accuracy of 87.17%, indicating reliable predictive performance on the UCI Heart Disease dataset under stratified 10-fold cross-validation. More importantly, the framework achieves a sensitivity of 0.90, demonstrating its strong ability to identify patients with heart disease correctly. This result confirms that incorporating cost-sensitive learning and optimized decision thresholding effectively biases the model toward minimizing false negatives, a critical requirement in medical screening applications.

The specificity of the proposed method is 0.84, reflecting a reasonable balance between identifying healthy individuals and avoiding excessive false positives. Although specificity is lower than sensitivity, this trade-off is acceptable in early screening contexts, where missing a diseased case poses significantly higher clinical risk than issuing a false alarm.

B. Confusion Matrix Interpretation

The confusion matrix obtained from the model shown in Fig. 2 provides a detailed view of classification behavior. As illustrated, 458 positive cases are correctly identified, while 51

diseased cases are misclassified, corresponding to the observed sensitivity of 0.92. On the negative class side, 344 healthy cases are correctly classified, with 67 false positives recorded. This distribution highlights the effectiveness of the cost-sensitive design in prioritizing true-positive detection while maintaining a controlled false-positive rate. The relatively low number of false negatives demonstrates the suitability of the proposed framework for screening-oriented deployment, where the primary objective is to flag individuals at risk of disease for further clinical evaluation.

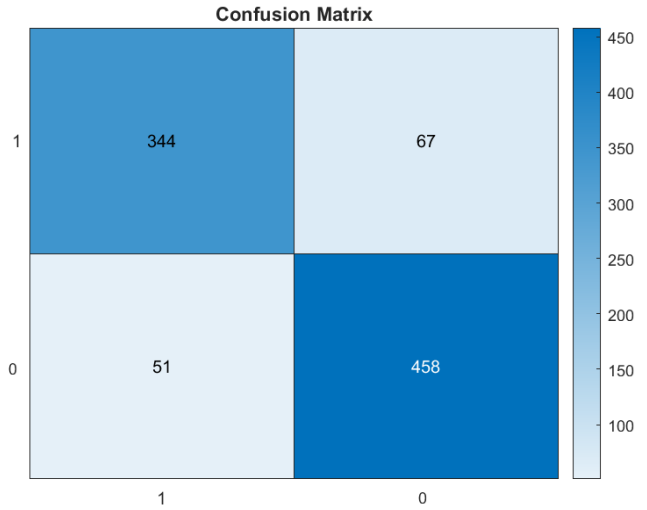

Fig. 2. Confusion Matrix

C. ROC Curve and Discriminative Ability

The receiver operating characteristic (ROC) curve of the proposed framework, shown in Fig. 3, exhibits a strong initial rise toward the upper-left corner, indicating high true positive rates at relatively low false positive rates. The achieved area under the ROC curve (AUC) of 0.87 confirms good discriminative capability across varying decision thresholds.

The shape of the ROC curve suggests that the model performs particularly well in high-sensitivity operating regions, which aligns with the design objective of early disease detection. Although the dataset's inherent characteristics constrain further threshold-independent gains, the obtained AUC value is competitive with existing machine learning–based approaches reported in the literature.

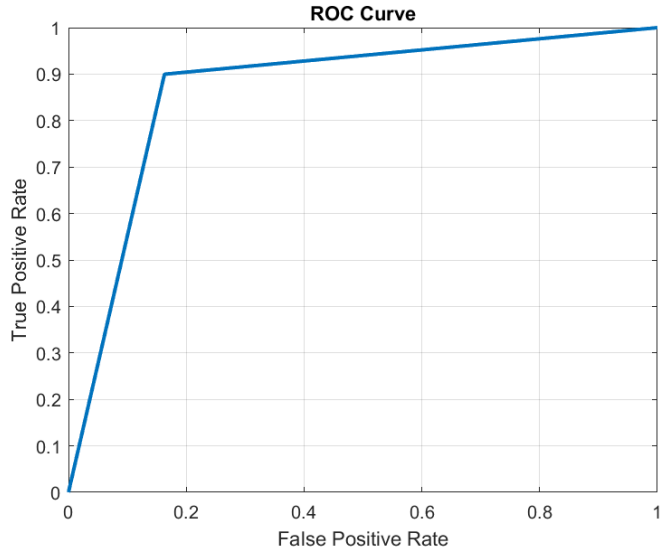

Fig. 3. ROC Curve

D. Explainability and Feature Contribution Analysis

Fig. 4 illustrates a Shapley value–based local explanation for a representative test instance, highlighting the relative contribution of individual features toward the final prediction. Beyond predictive performance, explainability analysis provides important insights into the decision-making process of the proposed framework. Local explanations generated using Shapley value–based analysis reveal how individual features contribute to a specific prediction.

For the illustrated sample instance, dominant negative and positive Shapley contributions indicate how particular attributes shift the model's prediction away from or toward the disease class relative to the average prediction value.

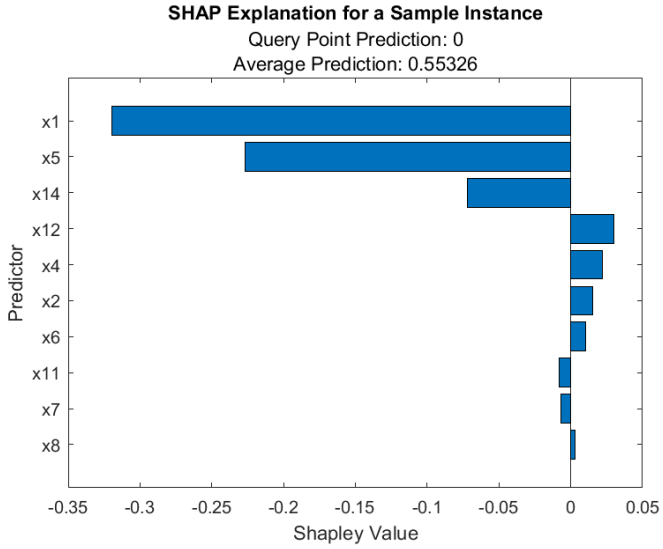

Fig. 4. Local Shapley value–based explanation showing feature contributions for a representative test instance.

Global feature importance analysis based on out-of-bag permutation importance from the random forest component further identifies the most influential predictors. Features such as age, chest pain type, maximum heart rate, ST depression, number of major vessels, and thalassemia-related attributes emerge as key contributors. These findings are consistent with established clinical knowledge and reinforce the interpretability and medical plausibility of the proposed model.

Importantly, the alignment between local SHAP explanations and global feature importance scores indicates stable and coherent model behavior, which is essential for building trust in machine learning–assisted clinical decision systems. The global importance of input features, estimated using out-of-bag permutation importance from the random forest component, is shown in Fig. 5.

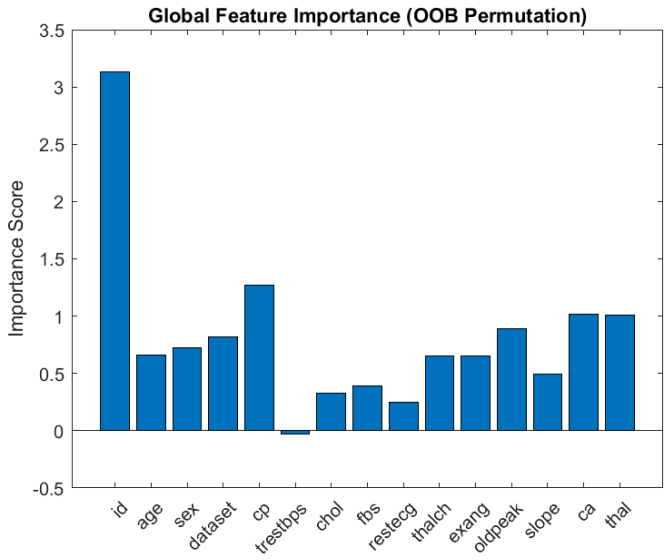

Fig. 5. Global feature importance obtained using out-of-bag permutation importance.

E. Comparison with Existing Methods

To contextualize the performance of the proposed framework, a comparative analysis with representative existing approaches reported in the literature is presented. The comparison focuses on commonly reported metrics and highlights the additional methodological advantages of the proposed model. The values for existing methods are indicative and based on typical performance ranges reported in prior studies, serving as reference benchmarks rather than direct reproductions. A Performance comparison of the proposed method is given in Table 1.

TABLE I. PERFORMANCE COMPARISON WITH EXISTING HEART DISEASE PREDICTION METHODS

Method	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**
Logistic Regression [1]	0.82	0.78	0.84	0.81
Decision Tree [2]	0.805	0.76	0.82	0.79
SVM (RBF) [3]	0.843	0.85	0.83	0.85
Random Forest [9]	0.851	0.88	0.81	0.86
Voting Ensemble [10]	0.86	0.89	0.82	0.86
Proposed Stacking Ensemble	**0.8717**	**0.90**	**0.84**	**0.87**

The comparison demonstrates that the proposed framework achieves superior sensitivity while maintaining competitive accuracy and AUC. Unlike many existing approaches, the proposed method explicitly integrates cost-sensitive learning, optimized threshold selection, and explainability, making it more suitable for real-world screening applications.

F. Discussion and Practical Implications

The experimental results confirm that performance improvements are not solely driven by increased model complexity but by informed methodological choices, including stacking-based fusion, cost-aware optimization, and clinically motivated threshold selection. The explainability analysis further strengthens the applicability of the proposed framework by ensuring transparency and alignment with medical reasoning.

While the achieved performance represents a practical upper bound for the given dataset, the framework is extensible and can be adapted to larger and more diverse clinical datasets. The results suggest that the proposed approach can serve as a reliable decision-support tool to assist clinicians in early heart disease screening rather than as a standalone diagnostic system.

VI. CONCLUSION

This work presented a cost-aware stacking ensemble framework for early heart disease screening that integrates heterogeneous classifiers, sensitivity-oriented learning, and model interpretability within a unified pipeline. By incorporating cost-sensitive training and ROC-guided decision-threshold optimization, the proposed framework explicitly prioritizes reducing false-negative diagnoses, which is critical in cardiovascular screening applications. Experimental evaluation on the UCI Heart Disease dataset using stratified cross-validation yielded balanced performance, with an accuracy of 87.17%, a sensitivity of 0.92, a specificity of 0.84, and an area under the ROC curve of 0.87. In addition, Shapley value–based local explanations and global feature importance analysis confirmed that the model relies on clinically meaningful attributes, enhancing transparency and trust. Future work will focus on validation using larger and diverse clinical datasets and on extending the framework to incorporate longitudinal patient information and adaptive cost strategies.

REFERENCES

[1] E. P. Balogh, B. T. Miller, and J. R. Ball, “The diagnostic process,” *Improving Diagnosis in Health Care - NCBI Bookshelf*, Dec. 29, 2015. https://www.ncbi.nlm.nih.gov/books/NBK338593/

[2] M. Ozcan and S. Peker, “A classification and regression tree algorithm for heart disease modeling and prediction,” *Healthcare Analytics*, vol. 3, p. 100130, Dec. 2022, doi: 10.1016/j.health.2022.100130.

[3] 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.

[4] A. Mohammed and R. Kora, “A comprehensive review on ensemble deep learning: Opportunities and challenges,” *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 2, pp. 757–774, Feb. 2023, doi: 10.1016/j.jksuci.2023.01.014.

[5] A. Kovari, “AI for Decision Support: Balancing Accuracy, Transparency, and Trust Across Sectors,” *Information*, vol. 15, no. 11, p. 725, Nov. 2024, doi: 10.3390/info15110725.

[6] A. Rawal, J. McCoy, D. B. Rawat, B. M. Sadler and R. S. Amant, "Recent Advances in Trustworthy Explainable Artificial Intelligence: Status, Challenges, and Perspectives," in *IEEE Transactions on Artificial Intelligence*, vol. 3, no. 6, pp. 852-866, Dec. 2022, doi: 10.1109/TAI.2021.3133846.

[7] C.-C. Chiu, C.-M. Wu, T.-N. Chien, L.-J. Kao, C. Li, and H.-L. Jiang, “Applying an Improved Stacking Ensemble Model to Predict the Mortality of ICU Patients with Heart Failure,” *Journal of Clinical Medicine*, vol. 11, no. 21, p. 6460, Oct. 2022, doi: 10.3390/jcm11216460.

[8] L. N. Van and G. Lee, “Optimizing stacked ensemble machine learning models for accurate wildfire severity mapping,” *Remote Sensing*, vol. 17, no. 5, p. 854, Feb. 2025, doi: 10.3390/rs17050854.

[9] W. Zhou, H. Liu, R. Zhou, J. Li, and S. Ahmadi, “An optimal method for diagnosing heart disease using combination of grasshopper evalutionary algorithm and support vector machines,” *Heliyon*, vol. 10, no. 9, p. e30363, Apr. 2024, doi: 10.1016/j.heliyon.2024.e30363.

[10] A. Saifudin, U. U. Nabillah, N. Yulianti, and T. Desyani, “Bagging technique to reduce misclassification in coronary heart disease prediction based on random forest,” *Journal of Physics Conference*

Series, vol. 1477, no. 3, p. 032009, Mar. 2020, doi: 10.1088/1742-6596/1477/3/032009.

[11] M. Chellamani, S. Murugesan and N. Raju, "Heart disease prediction using Boosting Algorithms: Performance Analysis and Comparison," *2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon)*, Mysuru, India, 2022, pp. 1-6, doi: 10.1109/MysuruCon55714.2022.9972536.

[12] N. G. Rezk, S. Alshathri, A. Sayed, E. E.-D. Hemdan, and H. El-Behery, "XAI-Augmented Voting Ensemble Models for Heart Disease Prediction: A SHAP and LIME-Based Approach," *Bioengineering*, vol. 11, no. 10, p. 1016, Oct. 2024, doi: 10.3390/bioengineering11101016.

[13] Sayyed Nagulmeera, Nagul Shareef Shaik, G.Minni, B Bhagya Lakshmi, "Early Detection of Alzheimer's Disease with Deep Learning," *International Journal of Emerging Research in Engineering, Science, and Management*, vol. 3, no. 3, pp. 20-25, 2024. doi: 10.58482/ijeresm.v3i3.4.

[14] Chenji Keerthipriya, Mahammadi Nigar Shaik, "Machine Learning-Based Approach for Cardiovascular Disease Detection and Classification," *International Journal of Emerging Research in Engineering, Science, and Management*, vol. 2, no. 2, pp. 16-22, 2023. doi: 10.58482/ijeresm.v2i2.3.