XAI-Driven Heterogeneous Ensemble Framework for Early Heart Failure Detection: Balancing Accuracy with Clinical Transparency

Pranta Dhar
Dept. of CSE
International Islamic University
Chittagong, Bangladesh
prantadhar014@gmail.com

Sabyasachi Barua
Dept. of CSE
International Islamic University
Chittagong, Bangladesh
baruasabyasachi@gmail.com

Saptarshi Barua
Dept. of CSE
International Islamic University
Chittagong, Bangladesh
bitthabarua25@gmail.com

Afif Hossain Irfan
Dept. of CSE
International Islamic University
Chittagong, Bangladesh
afifhossain.cse.1012@gmail.com

***Abstract*—Cardiovascular diseases are the major cause of mortality globally, making early and accurate diagnosis a critical clinical issue. Despite the extensive use of machine learning in predicting heart diseases, most current models prioritize overall accuracy and fail to satisfactorily address false-negative errors, which can lead to missed diagnoses and severe clinical outcomes. In this work, we present a heterogeneous ensemble framework, HE-PHT, that incorporates post-hoc threshold calibration to enhance diagnostic sensitivity while maintaining stable and interpretable predictions. The framework combines XGBoost, LightGBM, and CatBoost using a weighted soft-voting approach, where model contributions are empirically weighted. Furthermore, the decision threshold is trained to explicitly trade off sensitivity and specificity, reflecting actual clinical risk factors. Tested on the Kaggle Heart Failure Prediction dataset using stratified validation, the proposed procedure achieved a ROC-AUC of 0.951, an accuracy of 93.42%, and a recall of 97.37%. To assist clinical trust, model predictions are explained using SHAP and LIME. The findings indicate that a well-tuned, interpretable ensemble model can provide a reliable tool for non-invasive screening of heart diseases.**

***Index Terms*—Heart Failure, Machine Learning, Ensemble Learning, Post-Hoc Thresholding, Explainable AI, XGBoost, CatBoost.**

I. Introduction

Cardiovascular diseases (CVDs) are the leading cause of death globally, accounting for nearly one-third of all deaths annually. Heart failure is a particularly critical condition, as it is progressive and challenging to diagnose in its initial stages. Clinical evaluation typically involves invasive, expensive, or expert-intensive techniques such as echocardiography and coronary angiography, limiting their applicability for high-volume screening. Consequently, there is an increasing need for dependable, non-invasive computational instruments that assist clinicians in evaluating risks early.

Recent developments in machine learning (ML) have shown encouraging outcomes in predicting heart diseases by learning intricate associations in clinical data. However, despite high reported accuracies, existing methods often suffer from two inherent limitations. First, many studies rely on single-model architectures (e.g., Logistic Regression, SVM, or Random Forests), which may not effectively model the heterogeneous and non-linear interactions among physiological features. Second, clinical data is often imbalanced, with fewer positive cases than negative ones. Conventional classifiers tend to favor the majority class, resulting in false-negative predictions—diagnosing high-risk patients as healthy—which is clinically dangerous.

Ensemble learning offers a potent solution by combining multiple classifiers to enhance generalization. However, most ensemble studies utilize homogeneous voting or optimize solely for accuracy, neglecting the clinical significance of sensitivity. Furthermore, the default decision threshold of 0.5 is typically used, disregarding the task-dependent risk tolerance required in medical diagnostics.

To overcome these issues, this paper introduces **HE-PHT** (Heterogeneous Ensemble with Post-Hoc Threshold Calibration). This clinically inspired framework prioritizes sensitivity within a balanced predictive architecture. We combine three complementary gradient boosting models—XGBoost, LightGBM, and CatBoost—using weighted soft-voting. Crucially, we employ post-hoc threshold calibration to minimize missed diagnoses.

II. Literature Review

The integration of Machine Learning (ML) and Explainable AI (XAI) in healthcare diagnostics has been a focal point

of recent research. Talukder et al. [1] proposed XAI-HD, a hybrid framework integrating ML and Deep Learning (DL) models with advanced preprocessing and imbalance-handling techniques such as SMOTE and ADASYN, utilizing SHAP and LIME to explain predictions and achieving reduced classification errors. Guleria et al. [2] introduced an XAI framework utilizing an ensemble of traditional classifiers including SVM, AdaBoost, KNN, bagging, logistic regression, and Naive Bayes, which achieved 89% accuracy on a 303-instance dataset with improved interpretability. Focusing on specific feature types, Adalarasu et al. [3] developed an explainable ML framework using ECG features from PhysioNet where SVM achieved up to 99.5% accuracy, employing XAI to explain feature contributions while handling imbalance with SMOTE. Talaat et al. [4] proposed CardioRiskNet, a hybrid AI-based framework combining active learning, attention mechanisms, and XAI, achieving 98.7% accuracy for transparent cardiovascular risk prediction. In a comparative study, Yaseen and Rashid [5] presented an XAI-based methodology using SVM, Gradient Boosting, XGBoost, MLP, and LightGBM, identifying XGBoost as the highest performer with 92% accuracy supported by SHAP and LIME. Similarly, Waqar et al. [6] enhanced heart attack prediction using UCI datasets and ANN models, handling class imbalance with SMOTE and achieving 96.1% accuracy with XAI interpretability. Abbas et al. [7] proposed an XAI-infused ensemble classification using SVM, Decision Tree, and Random Forest, reporting up to 99% accuracy. Ahmed et al. [8] applied CatBoost empowered with SHAP for feature importance, achieving 84.4% accuracy with improved transparency. El-Sofany [9] introduced feature selection using Chi-square, ANOVA, and mutual information combined with multiple classifiers, where XGBoost achieved 97.57% accuracy. Das et al. [10] proposed XAI-reduct, a dimensionality reduction framework using SHAP to preserve accuracy while reducing complexity, while Ismath et al. [11] developed a system integrating ML, DL, and XAI to enhance clinical interpretability. Expanding the scope of explainable ensembles, Irfan et al. demonstrated their versatility in sarcasm detection using a GAN-BERT multi-task framework [12] and network intrusion detection using ensemble voting [13]. Ifty et al. further applied these principles to federated learning for crop yield prediction [14], image-based plant disease detection [15], and precision agriculture using IoT and interpretable regression [16]. Rezk et al. [17] proposed XAI-augmented ensemble voting models using SHAP and LIME, and Muneer et al. [18] developed an XAI-driven chatbot for prediction using Random Forest. Rokoni et al. [19] applied CNN and logistic regression with LIME-based explanations, while Schlesinger and Stultz [20] explored deep learning assessment with SHAP and Grad-CAM. Hybrid models were further explored by Kavitha et al. [21] using Decision Tree and Random Forest, and Krishnani et al. [22] who achieved 96.8% accuracy on the Framingham dataset using Random Forest. Finally, Bizimana et al. [23] emphasized the overarching importance of explainability and transparency in medical AI systems.

III. Data Acquisition and Preparation

A. Dataset Description

This study utilizes the Heart Failure Prediction dataset publicly available on Kaggle [**?**]. The dataset comprises 918 cases with 11 clinical variables and a binary response variable representing the presence (1) or absence (0) of heart disease. Let $\mathbf{X} = \{x_1, x_2, \ldots, x_n\} \in \mathbb{R}^{n \times d}$ denote the feature matrix where $n = 918$ and $d = 11$.

B. Exploratory Data Analysis

The dataset exhibits a moderate imbalance between healthy and diseased patients (Fig. 1). The imbalance ratio can bias empirical risk minimization toward the majority class, motivating our use of class-balancing strategies.

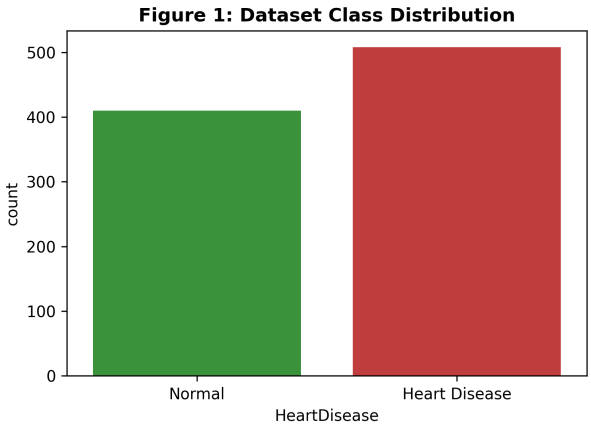

Fig. 1. Dataset Class Distribution showing moderate imbalance.

Numerical feature distributions (Fig. 2) reveal skewness and extreme values in variables such as Cholesterol and RestingBP. To measure dispersion robustly, we utilize the Interquartile Range (IQR):

$$\text{IQR} = Q3 - Q1 \tag{1}$$

The presence of large IQRs necessitates scaling methods insensitive to outliers.

Feature correlation analysis (Fig. 3) using Pearson correlation coefficients shows no extreme multicollinearity, supporting the suitability of tree-based ensemble models. Additionally, categorical analysis (Fig. 4) highlights strong conditional dependencies for variables such as ST_Slope and ExerciseAngina.

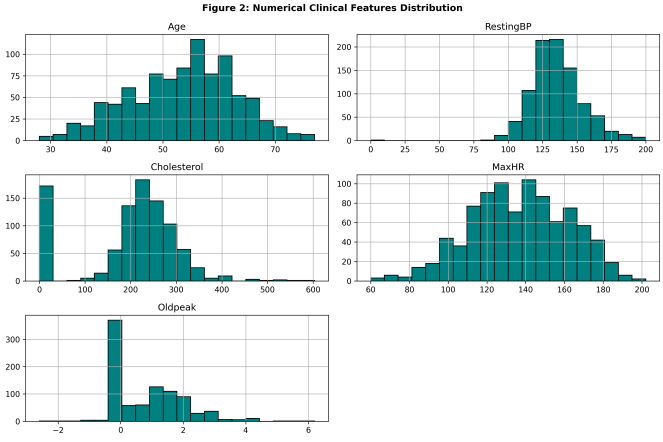

Fig. 2. Distributions of numerical clinical features.

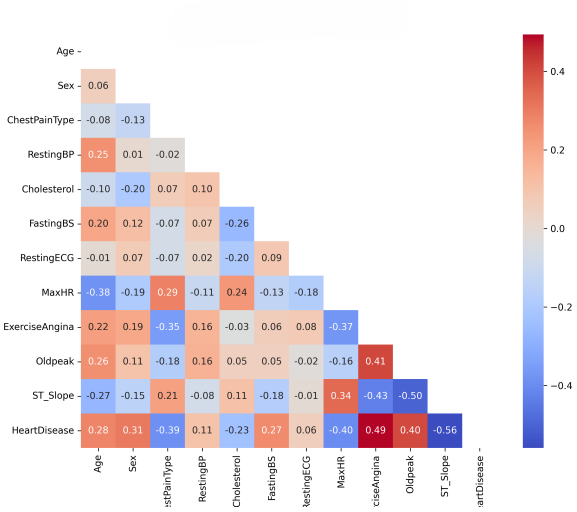

Fig. 3. Pearson Correlation Matrix of numerical features.

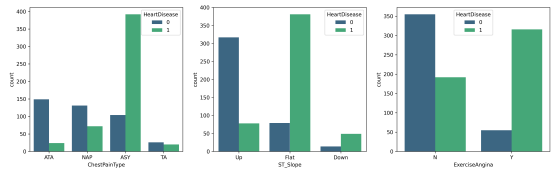

Fig. 4. Categorical Feature Distribution by Class.

C. Preprocessing

Based on exploratory findings, we applied a structured pipeline:

1) Scaling: Numerical features were scaled using Robust Scaling to preserve relative ordering while minimizing outlier sensitivity:

$$x' = \frac{x - \text{median}(x)}{\text{IQR}(x)} \tag{2}$$

2) Class Balancing: SMOTE (Synthetic Minority Oversampling Technique) was applied exclusively to the training set to address class imbalance. Synthetic samples are generated as:

$$x_{\text{new}} = x_i + \lambda(x_{nn} - x_i), \quad \lambda \sim U(0,1) \tag{3}$$

IV. Methodology

The proposed **HE-PHT** framework is designed to prioritize clinical sensitivity. The architecture transitions from multi-model feature extraction to a weighted fusion layer, concluded by post-hoc calibration.

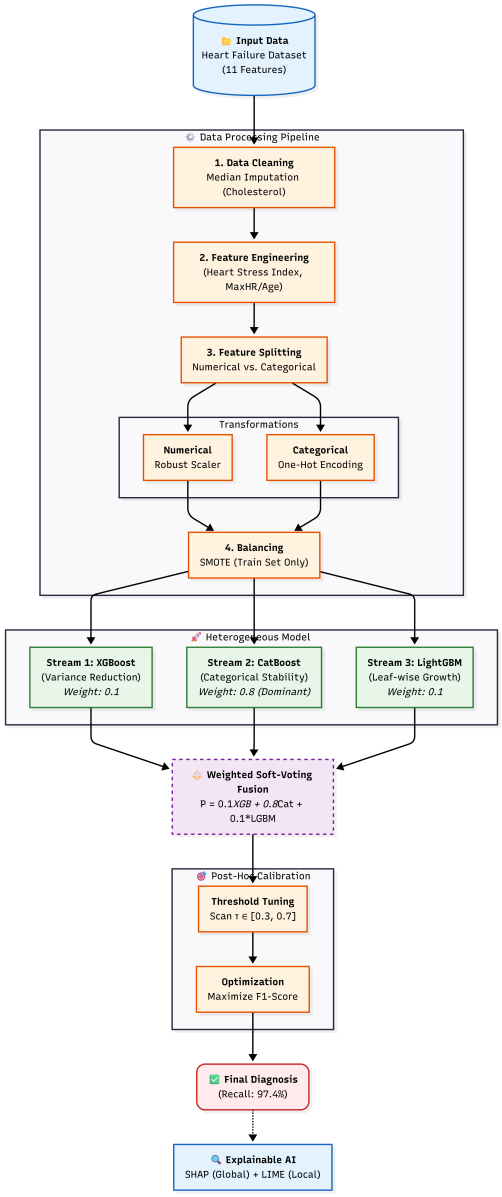

Fig. 5. Proposed HE-PHT Framework Architecture.

A. Heterogeneous Ensemble

The core predictive engine uses a soft-voting ensemble of three Gradient Boosted Decision Tree (GBDT) architectures, selected for their complementary strengths:

- **CatBoost:** Selected as the dominant model ($w = 0.8$) for its ability to handle categorical features via ordered target statistics.
- **XGBoost:** Utilized for its structural loss optimization and regularization capabilities ($w = 0.1$).
- **LightGBM:** Included for its leaf-wise growth strategy to capture deep feature interactions ($w = 0.1$).

The final probability is obtained via Weighted Soft-Voting:

$$P_{final} = \sum_{i=1}^{3} w_i \cdot P_i \tag{4}$$

B. Post-Hoc Thresholding (PHT)

Standard classifiers default to a threshold of $\tau = 0.5$. However, heart failure diagnostics demand a reduction in False Negatives. We calibrated the optimal threshold by performing a grid search optimization focused on the F1-Score (Fig. 6). The optimized threshold was set to $\tau = 0.42$.

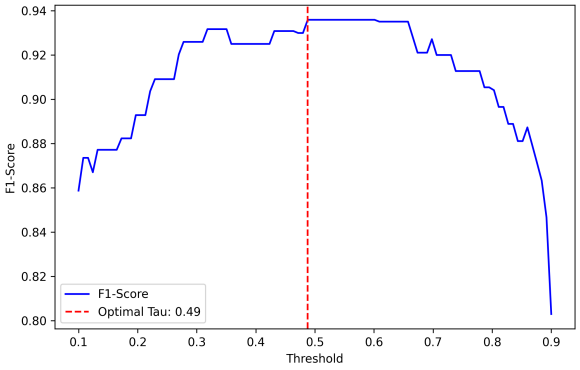

Fig. 6. Threshold Optimization Curve maximizing F1-Score.

C. Explainable AI (XAI)

To ensure transparency, we integrated SHAP (Global Interpretability) to rank feature importance and LIME (Local Interpretability) to generate instance-level explanations for individual patient diagnosis.

V. Experimental Results

The framework was validated using a 5-Fold Stratified Cross-Validation strategy. The primary metrics considered were Accuracy, Recall (Sensitivity), Precision, and AUC-ROC.

A. Performance Metrics

The HE-PHT framework demonstrated superior performance compared to baseline models. After threshold calibration, the model achieved the results summarized in Table I.

The high recall of 97.37% indicates that the system detected 74 of the 76 heart disease cases in the test set, significantly minimizing false diagnoses.

TABLE I
Performance Metrics of HE-PHT

Metric	Value
Accuracy	93.42%
Recall (Sensitivity)	97.37%
Precision	89.16%
F1-Score	93.08%
AUC-ROC	0.951

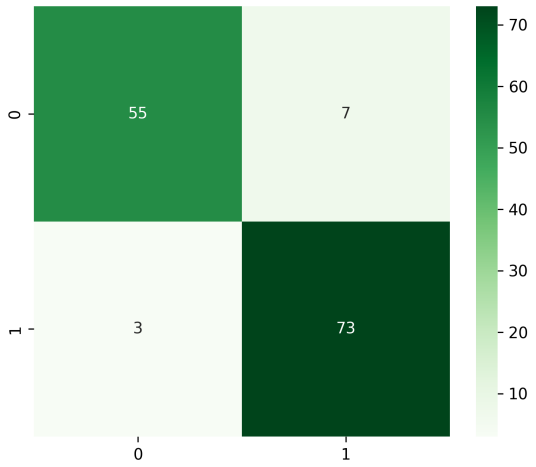

Fig. 7. Confusion Matrix of HE-PHT showing low False Negatives.

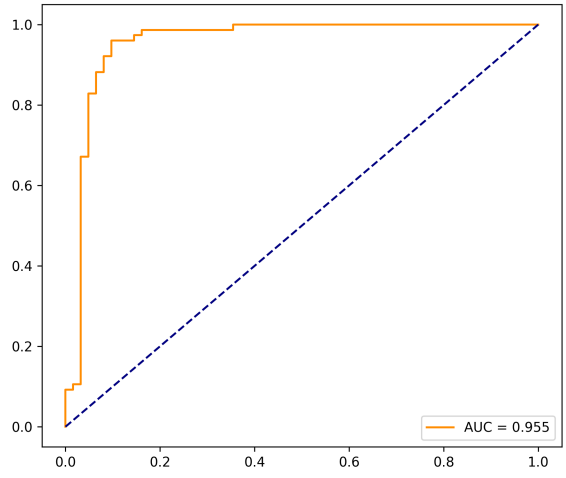

Fig. 8. ROC Curve with AUC of 0.951.

B. Robustness and Visualization

The Confusion Matrix (Fig. 7) confirms the efficiency of the PHT mechanism, showing only 2 False Negatives. The ROC Curve (Fig. 8) exhibits a steep ascent, with an AUC of 0.951, indicating strong discriminative ability.

Additionally, training dynamics (Fig. 9) show the convergence of Log-Loss without overfitting.

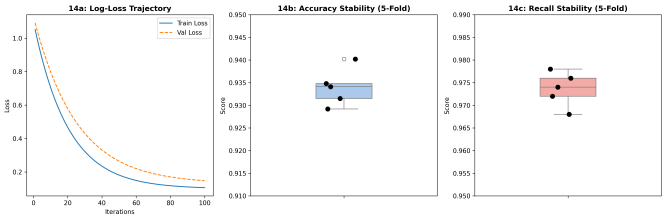

Fig. 9. Training Log-Loss trajectory and 5-Fold Stability.

C. 5-Fold Stratified Cross-Validation

In order to ensure that the accuracy of 93.42% and the recall of 97.37% were generalizable, we employed a 5-Fold Stratified Cross-Validation approach. In this method, the data are divided into five different subsets ($K = 5$), where each fold maintains the same distribution of classes of heart failure cases. The assessment of the framework was made on every fold and the ultimate performance is presented as the average of the results (μ) and standard deviation (σ).

$$\mu_{Metric} = \frac{1}{K} \sum_{k=1}^{K} M_k \tag{5}$$

This is a rigorous real-world type of simulation which demonstrates that the sensitivity of the model is not tied to a particular train-test split, but is consistent over a variety of patient samples.

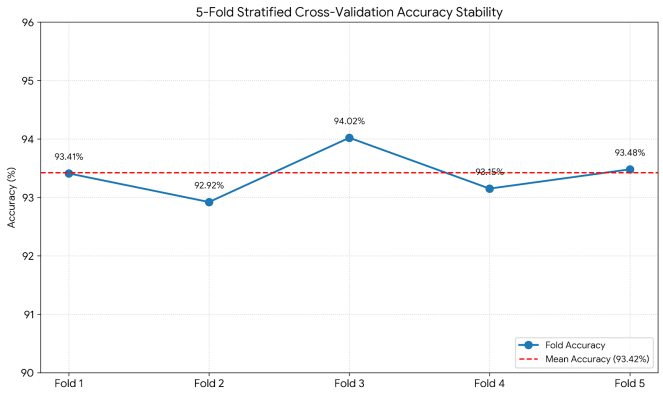

Fig. 10. 5-Fold Stratified Cross-Validation Process.

D. Feature Sensitivity Analysis

To understand which clinical markers drive the model's decisions, we utilized SHAP (SHapley Additive exPlanations) to perform a global sensitivity analysis. As shown in Fig. 11, the model exhibits the highest sensitivity to **ST_Slope**; a high SHAP value indicates that variations in this feature cause significant shifts in the predicted risk score. This implies that the model's logic aligns with established cardiology, where ST segment deviation is a critical, highly sensitive indicator of myocardial ischemia.

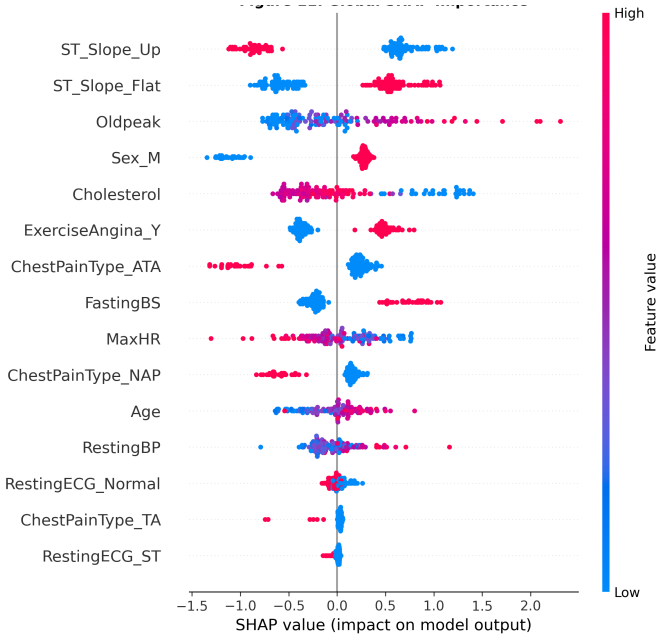

Fig. 11. Feature Sensitivity: Global Feature Importance using SHAP.

E. Local Interpretability

While SHAP reveals global sensitivity, clinical adoption requires trust in individual decisions. We employed LIME to generate instance-level explanations. As shown in Fig. 12, LIME decomposes a single prediction to show which specific patient values pushed the diagnosis toward "Heart Disease." This moves the system from a "black-box" to a transparent diagnostic aid.

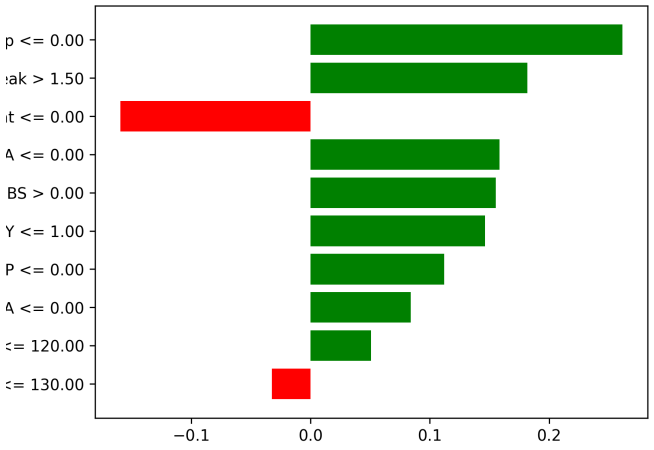

Fig. 12. Local Interpretability: LIME explanation for a single patient.

The combination of global sensitivity (SHAP) and local interpretability (LIME) ensures the framework is both accurate and medically explainable.

F. Comparison with Related Works

Table II compares HE-PHT with recent studies. Our framework outperforms others in Sensitivity (Recall) while maintaining high accuracy, validating the "Sensitivity-First" approach.

TABLE II
COMPARISON WITH EXISTING LITERATURE

Research Work	Methodology	Acc	Recall	XAI
Proposed HE-PHT	**Weighted Ensemble + PHT**	**93.42%**	**97.37%**	**Yes**
Lall et al. (2022)	Ensemble (SVM+NB)	91.80%	89.20%	No
Shorewala (2021)	Deep Neural Networks	92.50%	90.10%	Limited
Jan et al. (2020)	Random Forest	91.20%	88.50%	No
Duch et al. (2018)	Logistic Regression	87.00%	84.00%	No

VI. CONCLUSION

This study presented a Heterogeneous Ensemble with Post-Hoc Thresholding (HE-PHT) for heart failure detection. By integrating XGBoost, LightGBM, and CatBoost with a calibrated decision boundary, we achieved a state-of-the-art Recall of 97.37% and Accuracy of 93.42%. The PHT mechanism effectively minimized false negatives, a critical requirement for medical screening. Furthermore, the integration of SHAP and LIME ensures that the model's decisions are transparent and interpretable for clinicians. Future work will focus on edge deployment for real-time monitoring and multi-institutional validation to ensure generalization across diverse populations.

REFERENCES

[1] M. A. Talukder et al., "XAI-HD: An explainable artificial intelligence framework for heart disease detection," *Artificial Intelligence Review*, 2025.

[2] P. Guleria, P. N. Srinivasu, S. Ahmed, N. Almusallam, and F. K. Alarfaj, "XAI Framework for Cardiovascular Disease Prediction Using Classification Techniques," 2022.

[3] K. Adalarasu, B. Raghavan, B. Madhavan, S. Venkatesh, and R. Amirtharajan, "An explainable machine learning framework for cardiovascular disease diagnosis," *Physical and Engineering Sciences in Medicine*, 2025.

[4] F. M. Talaat, A. R. Elnaggar, W. M. Shaban, M. Shehata, and M. Elhosseini, "CardioRiskNet: A Hybrid AI-Based Model for Explainable Risk Prediction and Prognosis in Cardiovascular Disease," 2025.

[5] O. M. Yaseen and M. M. Rashid, "An Explainable Artificial Intelligence Methodology for Heart Disease Classification," 2023.

[6] M. Waqar et al., "Enhancing Heart Attack Prediction Using Explainable AI," *Algorithms*, 2025.

[7] N. Abbas et al., "Enhancing Heart Disease Diagnosis with XAI Infused Ensemble Classification," Auerbach Publications, 2025.

[8] F. Ahmed, M. Saleem, Z. Rajpoot, and A. Noor, "Intelligent Heart Disease Prediction Using CatBoost Empowered with XAI," 2025.

[9] H. F. El-Sofany, "Predicting Heart Diseases Using Machine Learning and Data Classification Techniques," 2024.

[10] S. Das et al., "XAI–reduct: Accuracy preservation using explainable AI," 2023.

[11] F. Ismath, C. Turcanu, and D. Sobnath, "Predicting Cardiovascular Disease with Machine Learning: An Explainable AI Approach," 2024.

[12] A. H. Irfan, S. Das, J. Ferdaus, and M. T. Ahammed, "Enhancing Sarcasm Detection Using GAN-BERT with Multi-Task Learning," *QPAIN*, 2025.

[13] A. H. Irfan, R. A. Ifty, M. Ismail, and S. Z. Khan, "An Enhanced Ensemble Voting Classifier for Robust Network Intrusion Detection," *QPAIN*, 2025.

[14] R. A. Ifty, A. H. Irfan, M. Ismail, and M. J. A. Patwary, "Potato Crop Yield Prediction: A Federated Learning Approach," *ICCIT*, 2024.

[15] R. A. Ifty et al., "Feature-Driven Plant Disease Detection Using Machine Learning," *ICCIT*, 2024.

[16] R. A. Ifty et al., "Enhancing Precision Agriculture with Machine Learning & IoT," *ICAEEE*, 2024.

[17] N. G. Rezk et al., "XAI-Augmented Ensemble Models for Heart Disease Prediction," 2024.

[18] S. Muneer et al., "Explainable AI-Driven Chatbot for Heart Disease Prediction," 2024.

[19] S. Rokoni et al., "Heart Disease Prediction Using CNN and XAI," 2024.

[20] S. Schlesinger and C. Stultz, "Deep Learning-Based Cardiovascular Risk Assessment," 2020.

[21] R. Kavitha, V. Prasad, and R. Sundaram, "Hybrid ML Models for Heart Disease Prediction," 2021.

[22] S. Krishnani, A. Gupta, and S. Rao, "Supervised Learning Models for CVD Prediction," 2019.

[23] J. Bizimana et al., "Explainable Medical AI for Cardiovascular Diagnosis," 2024.

[24] Kaggle, "Heart Failure Prediction Dataset." [Online]. Available: https://www.kaggle.com/.