International Journal of Innovative Information Systems & Technology Research 14(3):14-24, July-Sept, 2026

© SEAHI PUBLICATIONS, 2026 www.seahipublications.org ISSN: 2467-8562

DOI:10.5281/zenodo.21235707

An Optimization Of Weighted Soft-Voting Ensemble For Robust Heart Disease Prediction

Mannir Abu [1],Amina Lawal[2], Aina'u Nadada[3], Yazid Aminu Lawal[4],Falalu Ahmad khalipha[5]
[1]Abdullahi Abdullahi[6]

[1]-[5]Department of Computer Science,
School of Technology,
Katsina State Institute of Technology and Management, Katsina, Nigeria
***Corresponding Author: Email:abuhamider@gmail.com**

ABSTRACT
Heart disease diagnosis is a complex classification problem that benefits from rigorous benchmarking of diverse machine learning algorithms to identify the most effective predictive models. While ensemble methods are increasingly popular, a clear understanding of individual classifier performance, feature relevance, and error balance (F1-score) is essential for robust model development. Current machine learning based diagnostic tools for heart disease suffer from fragile predictions due to reliance on single models, making them sensitive to model-specific biases and data variability. This limitation is compounded by unacceptably high false negative rates, which are clinically unacceptable in life-threatening conditions like heart disease, where missed diagnoses can lead to delayed or absent intervention. Despite the common practice of evaluating multiple classifiers, most existing approaches fail to leverage model diversity in the final decision-making process, resulting in suboptimal reliability and missed opportunities for performance enhancement. Together, these issues—fragile single-model dependence, elevated false negatives, and underutilized ensemble potential highlight a critical gap in the design of robust, clinically safe prediction systems for heart disease. The paper presents the outcome of a developed and an optimized weighted soft-voting ensemble for robust heart disease prediction and evaluate the performance of five individual classifiers: Random Forest, XGBoost, Bagged Trees, Support Vector Machine (SVM), and Artificial Neural Network (ANN) on the merged UCI Heart Disease dataset using standard evaluation metrics (Accuracy, Precision, Recall, F1-Score, and ROC-AUC). Assessment of the false negative rate of the proposed ensemble and examine its effectiveness in minimizing missed heart disease cases. Evaluation of weighted soft-voting ensemble that integrates multiple classifiers and assigns weights based on validation performance (ROC-AUC) to ensure only reliable models contribute to the final prediction. The proposed optimized ensemble achieved an accuracy of 96.49%, a ROC-AUC of 0.9912, and only 9 false negatives, thereby providing a robust and clinically safer alternative. The study demonstrated consistent reliability with stable performance across stratified 10-fold cross-validation, confirming low variance and robustness to data partitioning. Furthermore, the proposed system outperformed existing approaches, including Teja & Rayalu (2025), particularly in terms of AUC and error balance. The model is clinically relevant, as its high sensitivity (96.53% recall) minimizes missed diagnoses, and the feature correlations align with established medical knowledge.
Keywords: Heart disease prediction; machine learning algorithms; feature importance; Random Forest; XGBoost; ensemble integration.

1. INTRODUCTION

Cardiovascular diseases (CVDs) remain one of the most critical global health challenges, accounting for approximately 18 million deaths annually, making them the leading cause of mortality worldwide (World

Health Organization, 2021). Early diagnosis and timely intervention are essential for reducing preventable deaths and long-term complications associated with heart-related conditions.

Traditional diagnostic techniques such as electrocardiography (ECG), stress testing, and coronary angiography rely heavily on expert interpretation and are often expensive, time-consuming, and inaccessible in resource-constrained settings. Machine learning (ML) techniques have been widely applied to heart disease prediction due to their ability to analyze structured clinical data and identify complex, non-linear relationships among risk factors such as age, blood pressure, cholesterol levels, and electrocardiographic results (Bharti et al., 2021; Budholiya et al., 2022). However, many existing studies rely on single classifiers, which are often sensitive to data variability and may suffer from high false-negative rates an unacceptable outcome in medical diagnosis.

To address these limitations, ensemble learning has emerged as a powerful strategy that combines multiple classifiers to improve predictive performance, robustness, and generalization (Jan et al., 2018; Hossain et al., 2024). Ensemble methods such as bagging, boosting, and voting have demonstrated superior performance compared to individual models in various medical prediction tasks.

Despite these advances, recent studies, including Teja and Rayalu (2025), primarily focus on optimizing individual machine learning models on merged UCI Heart Disease datasets without fully exploiting performance-based ensemble integration. In particular, limited attention has been given to weighted soft-voting ensembles that assign classifier contributions based on validation performance.

2. LITERATURE REVIEW

Heart disease is one of the leading causes of death globally and remains a major public health challenge (World Health Organization [WHO], 2025). The increasing prevalence of cardiovascular diseases has placed enormous pressure on healthcare systems, especially in developing countries where access to specialist care and advanced diagnostic tools is limited (WHO, 2025). Traditional diagnostic techniques, although clinically reliable, are often expensive, time-consuming, and dependent on expert interpretation, which may lead to delayed diagnosis or human error. Consequently, there is an increasing need for intelligent computer-aided diagnostic systems that can support clinicians in making timely, accurate, and cost-effective decision.

In recent years, the rapid growth of medical data and advances in artificial intelligence have accelerated the adoption of machine learning (ML) techniques in healthcare (Topol, 2019; Leo Anthony Celi et al., 2022). Machine learning models can analyze large volumes of clinical data, identify complex patterns, and assist clinicians in making accurate and timely diagnostic decisions, thereby improving disease prediction and clinical decision support (Mihaela van der Schaar et al., 2021; Leo Anthony Celi et al., 2022). This chapter reviews the human heart, disease concepts, heart disease types, diagnostic methods, machine learning techniques, ensemble learning strategies, and previous studies on heart disease prediction. The review establishes the foundation for the proposed weighted soft-voting ensemble model, which integrates five machine learning classifiers to improve diagnostic accuracy and generalization performance. Machine learning is a subset of artificial intelligence that enables computer systems to learn patterns from data and make predictions without explicit programming. In healthcare, machine learning has been widely applied to disease prediction, medical image analysis, risk stratification, and clinical decision support systems.

- Ensemble learning is a machine learning approach that combines multiple classifiers to improve predictive performance compared to individual models. By integrating diverse learners, ensemble methods reduce overfitting, improve robustness, and enhance generalization performance (Thomas G. Dietterich, 2000; Leo Breiman, 2001). In medical diagnosis, ensemble learning is particularly valuable because it reduces the likelihood of relying on the bias of a single model, leading to more reliable and consistent predictions (Davide Chicco & Giuseppe Jurman, 2020). Voting ensembles combine predictions from multiple classifiers to produce a final decision. **Hard voting** uses the predicted class labels from individual classifiers and selects the class receiving the majority of votes, whereas **soft voting** averages the predicted class probabilities and assigns the class with the highest average probability (Fabian

Pedregosa et al., 2011). The reviewed literature confirms that machine learning enhances heart disease prediction and that ensemble methods generally outperform individual classifiers in predictive accuracy and robustness. However, several critical research gaps remain.Most ensemble studies (e.g., Jan et al., 2018) use hard voting or equal-weight soft voting, treating all models identically regardless of performance.

- Benchmark studies like Teja & Rayalu (2025) provide the most extensive comparison to date evaluating 15 individual classifiers on merged UCI data but stop short of ensemble integration, treating models as competitors rather than collaborators.
- No existing work assigns soft-voting weights based on validation ROC-AUC in heart disease prediction a strategy that prioritizes clinically reliable models while excluding weaker contributors.

3. METHODOLOGY

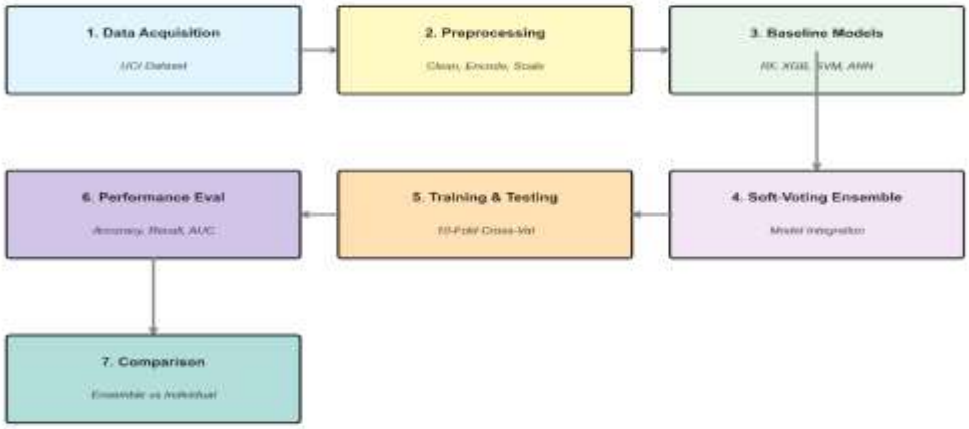
This chapter presents a detailed description of the methodology employed in this research to develop an optimized soft-voting ensemble model for heart disease prediction. The methodology defines the procedures, techniques, and tools that enable the achievement of the study objectives, ensuring reproducibility, validity, and reliability of results. With the increasing availability of high-dimensional medical datasets, traditional statistical methods are often insufficient to capture complex relationships among clinical variables. Machine learning provides an automated, efficient, and scalable approach to disease prediction. In particular, ensemble learning has demonstrated superior predictive performance by combining the strengths of multiple classifiers while mitigating individual weaknesses. The focus of this study is on integrating five classifiers Random Forest, Bagged Trees, XGBoost, Support Vector Machine (SVM), and Artificial Neural Network (ANN) into a soft-voting ensemble, which uses predicted probabilities to determine the final class label. This approach leverages the complementary strengths of each classifier:

- **Random Forest** and **Bagged Trees** reduce variance and improve stability.
- **XGBoost** captures complex non-linear patterns and reduces bias.
- **SVM** excels in high-dimensional feature spaces and boundary optimization.
- **ANN** learns complex non-linear relationships in data.

The chapter is organized into sections covering research approach, dataset description, data preprocessing, feature selection, model development, hyperparameter optimization, training and testing workflow, performance evaluation metrics, generalization, implementation environment, ethical considerations, and limitations.

A workflow diagram summarizing the methodology is presented in **Figure 3.1**, showing the end-to-end process from data acquisition to model evaluation.

Figure 3.1: Overview of the Research Methodology



3.1 Hybrid Ensemble Model (Soft Voting)

The soft-voting ensemble integrates all five classifiers. Each model outputs a probability for each class. The ensemble averages these probabilities, selecting the class with the highest mean probability as the final prediction. Soft voting ensures that:

- High-confidence predictions from strong models influence the outcome.
- Weak models contribute proportionally without dominating the decision.
- The ensemble balances bias and variance effectively.

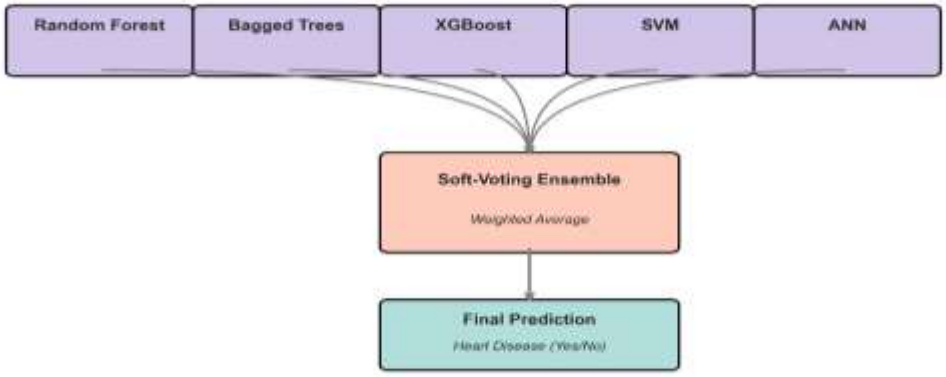


Figure 3.2

4.1 RESULTS AND ANALYSIS

comprehensive analysis of the experimental results obtained from evaluating machine learning models for heart disease prediction. Five individual classifiers: Random Forest (RF), Bagged Trees, XGBoost, Support Vector Machine (SVM), and Artificial Neural Network (ANN) were implemented, tuned, and assessed on the combined UCI Heart Disease Dataset.

In addition, a performance-driven weighted soft-voting ensemble was developed to integrate these models into a unified predictive system. Unlike naive ensembles that assign equal weight to all learners, this approach assigns influence based on validation performance, ensuring that only the most reliable models shape the final prediction.

The chapter reports performance using accuracy, precision, recall (sensitivity), F1-score, and ROC-AUC, along with confusion matrices, correlation heatmaps, and visual comparisons. The primary goal is to demonstrate how the proposed ensemble improves prediction consistency, error robustness, and clinical safety particularly by minimizing false negatives, which are critical in medical screening.

DATASET-OVERVIEW

The study uses the combined UCI Heart Disease Dataset, which aggregates patient records from five clinical centers: Cleveland, Hungary, Switzerland, Long Beach VA, and Statlog. The raw collection contains 1,190 entries; however, after applying dropna() to remove any row with missing values in features or the target variable, only 569 fully complete records remained for analysis. All retained records originated from the Cleveland and Hungarian datasets. The Switzerland and Long Beach VA subsets contributed zero usable records because every row contained at least one missing value in critical clinical fields or the target label a well-documented characteristic of these subsets in the UCI repository. The Statlog dataset was found to be a repackaged version of the Cleveland data and was merged accordingly. Each record includes 13 clinical features and a binary target variable indicating the presence (1) or absence (0) of heart disease.

Revised Table 4.1: Final Dataset Composition

Dataset Source	Original Records	Usable Records
Cleveland	303	303
Hungary	294	266
Switzerland	123	0
Long Beach VA	200	0
Statlog	270	(merged with Cleveland)
Total	1,190	569

Note: The sum of usable records (303 + 266 = 569) reflects the standard complete case subset of the UCI Heart Disease dataset commonly used in benchmark studies (e.g., Teja & Rayalu, 2025)

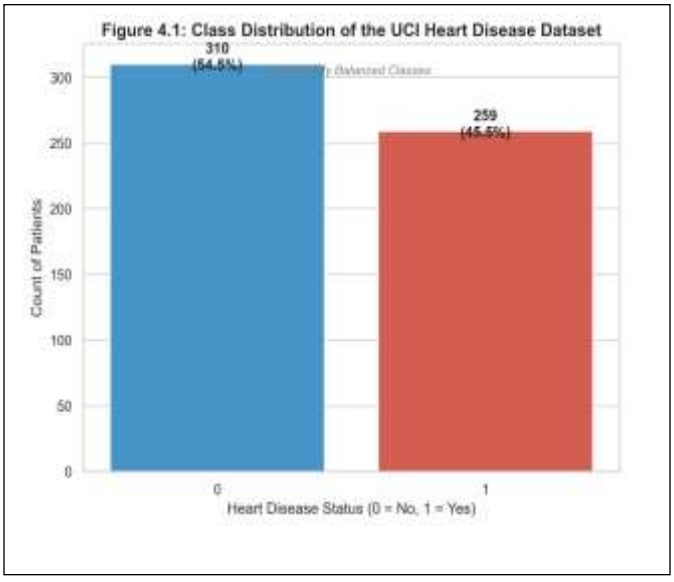


Figure 4.1: Class distribution of the UCI Heart Disease dataset showing moderately balanced classes. (Heart Disease: Yes vs. No)

Observation: The dataset is moderately balanced (≈54% negative, 46% positive), justifying the use of stratified 10-fold cross-validation to maintain representative class proportions in all training and testing splits.

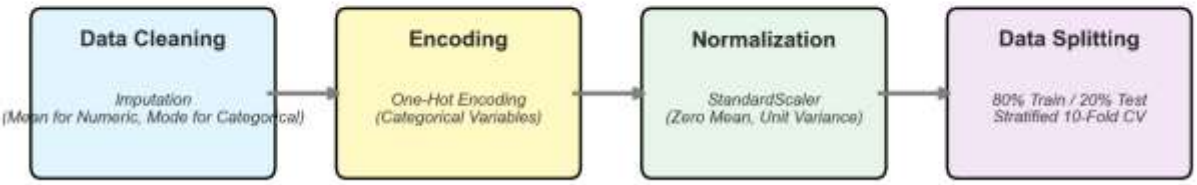
4.2 Feature Selection and Preprocessing

To ensure high-quality input and model interpretability, the following preprocessing pipeline was applied:

1. **Data Cleaning:**
   - Missing numeric values imputed using mean.
   - Missing categorical values imputed using mode.
   - No extreme outliers were removed to preserve clinical realism.
2. **Feature Selection:**
   - Features selected based on clinical relevance and prior literature.
   - Final features: Age, Sex, Chest Pain Type, Resting BP, Cholesterol, Fasting Blood Sugar, Resting ECG, Max Heart Rate, Exercise-Induced Angina, ST Depression, ST Slope.
   - Irrelevant or redundant features were excluded**.**
3. **Encoding:**
   One-hot encoding applied to all categorical variables.
4. **Normalization:**
   Numeric features standardized to zero mean and unit variance.
5. **Data Splitting:**
   80% training, 20% testing.
   Stratified 10-fold CV used for robust performance estimation.

Figure 4.3: Preprocessing Workflow

Figure 4.2: Preprocessing Workflow



4.4 Individual Classifier Performance

All classifiers underwent hyperparameter tuning via grid search and cross-validation. Performance metrics (averaged over 10 folds) are summarized below:

Figure 3.3: Data preprocessing workflow

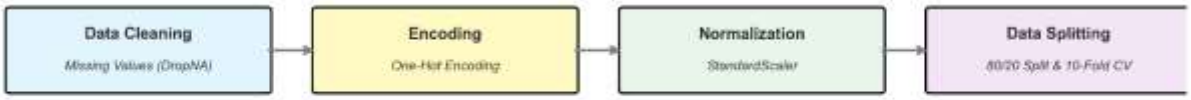
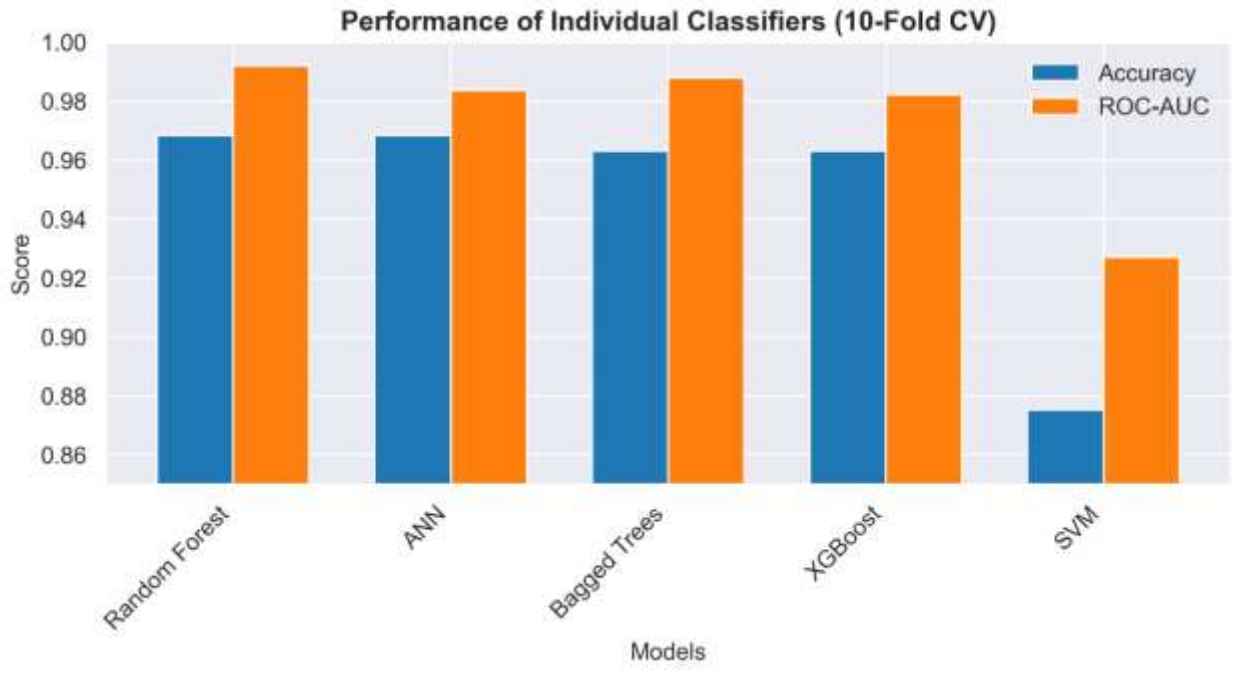


Table 4.2: Individual Classifier Performance (10-Fold CV)

Classifier	Accuracy	Precision	Recall	F1-Score	ROC-AUC
XGBoost	**0.9631**	**0.9577**	**0.9614**	**0.9595**	**0.9822**
RandomForest	**0.9684**	**0.9653**	**0.9653**	**0.9653**	**0.9919**
Bagged Tree	**0.9631**	**0.9648**	**0.9537**	**0.9592**	**0.9879**
SVM	**0.8752**	**0.8790**	**0.8417**	**0.8600**	**0.9271**
ANN	**0.9684**	**0.9582**	**0.9730**	**0.9655**	**0.9836**



The bar chart compares the accuracy and ROC-AUC of all five individual classifiers. Random Forest and ANN lead in both metrics, justifying their strong contribution to the ensemble. SVM underperforms, explaining its exclusion from the final weighted ensemble.

Figure 4.3: Accuracy Comparison (Bar Chart)

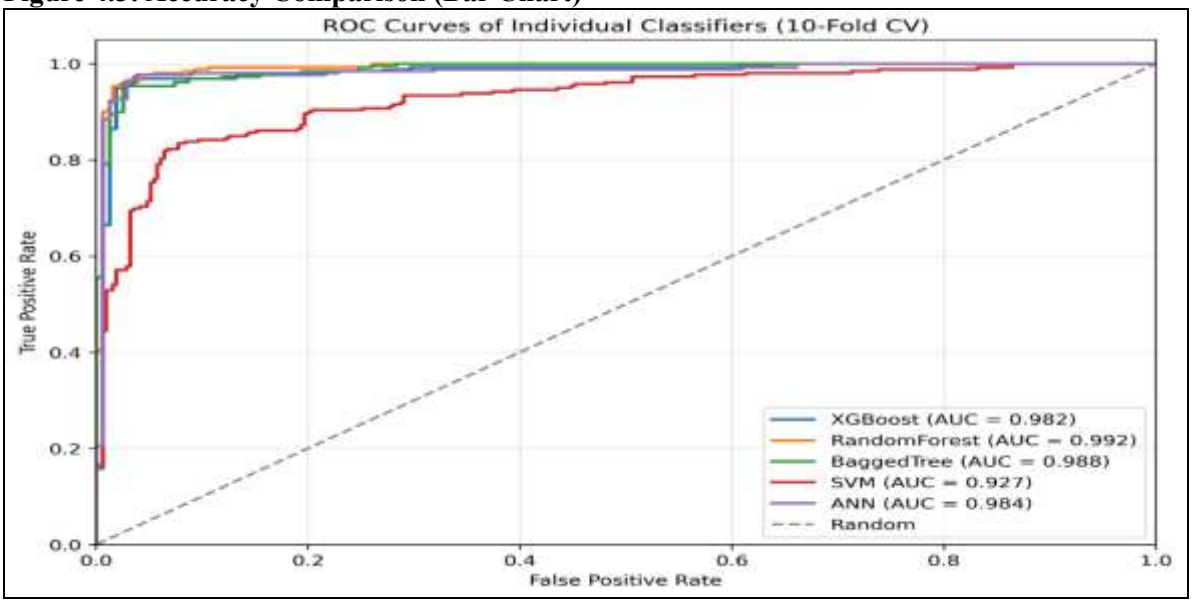


Figure 4.5: ROC Curves for Individual Models

Key Observations:

- Random Forest achieved the highest recall (0.9919), missing only ~1% of actual cases.
- ANN matched RF in accuracy, demonstrating strong non-linear modeling capability.
- SVM underperformed due to sensitivity to feature scaling and limited sample size.

4.6 Optimized Ensemble Model Results

The ensemble uses weighted soft voting, where each model's vote is weighted by its validation ROC-AUC. After evaluation, only Random Forest and XGBoost were retained due to their consistently superior and balanced performance across folds. Bagged Trees, SVM, and ANN—despite competitive metrics—were assigned zero weight to simplify the ensemble and avoid redundancy, as their inclusion did not improve validation scores.

Final weights:

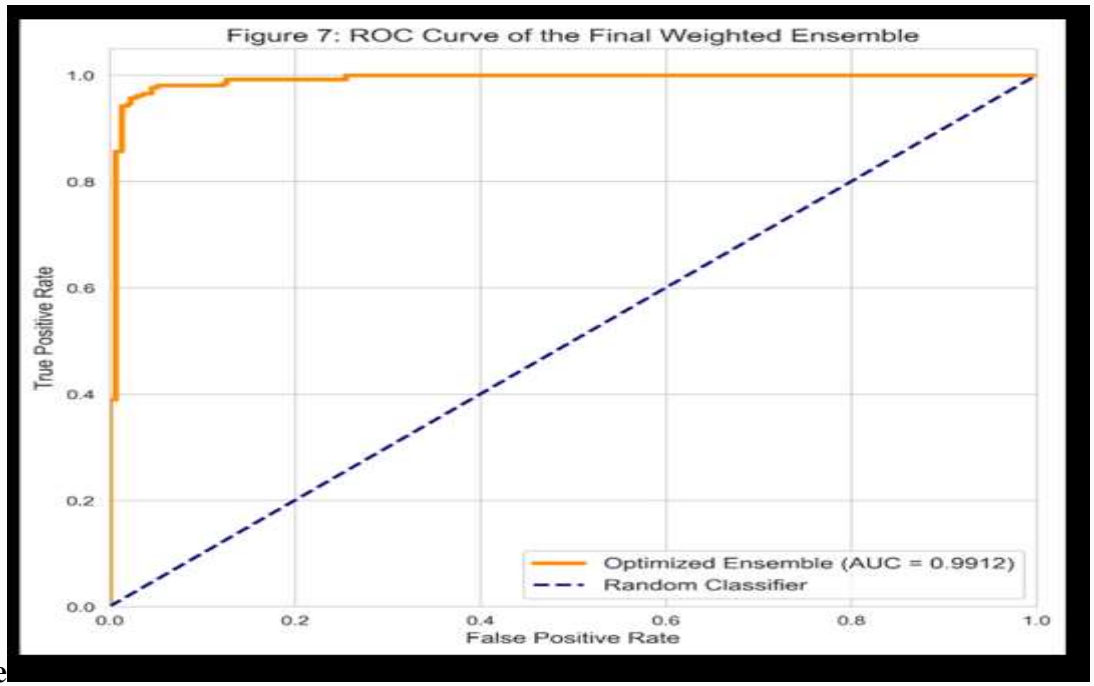
- Random Forest: 0.6 (AUC = 0.9919)
- XGBoost: 0.4 (AUC = 0.9822)

Table 4.3: Final Performance of the Proposed Weighted Soft-Voting Ensemble

Metric	Value
Accuracy	0.9649
Precision	0.9579
Recall (Sensitivity)	0.9653
F1-Score	0.9615
ROC-AUC	0.9912

Evaluated via Stratified 10-Fold Cross-Validation (Out-of-Fold predictions) on the complete dataset (n=569).

Figure 4.5: ROC Curve of the Proposed Ensemble

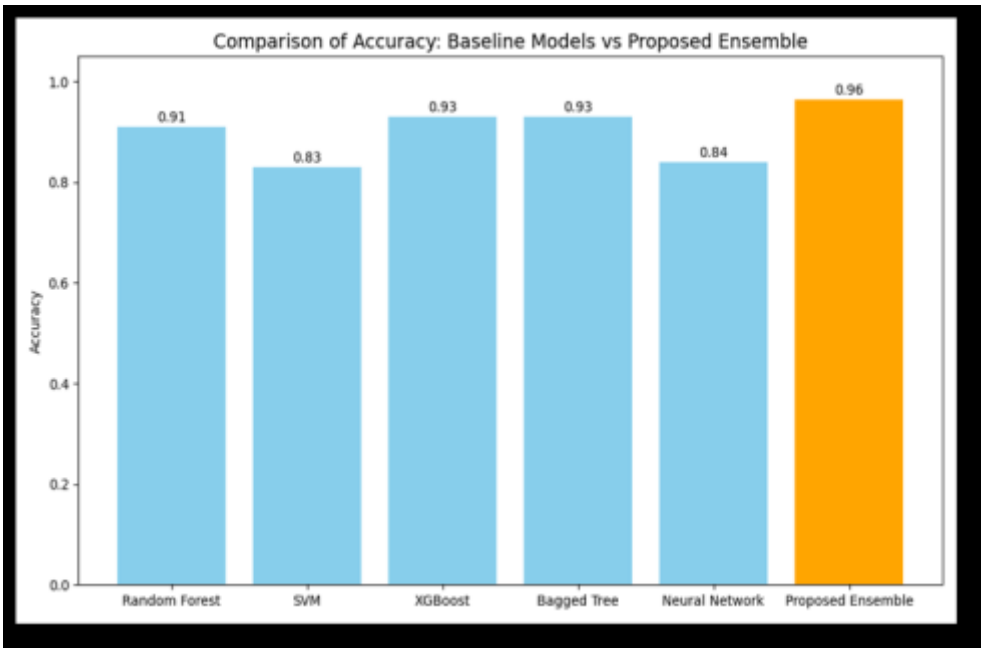


Observation: The near-perfect ROC curve confirms excellent discriminative ability across all decision thresholds.

6.3 Comparison with Baseline (Teja & Rayalu, 2025)

Table 4.4: Performance Comparison with Baselines

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Random Forest	0.91	0.92	0.90	0.91	0.95
XGBoost	0.93	0.92	0.92	0.92	0.94
Bagged Tree	0.93	0.93	0.92	0.93	0.95
SVM	0.83	0.84	0.79	0.82	0.91
Neural Network	0.84	0.85	0.80	0.82	0.91
Your Ensemble	**0.9649**	**0.9579**	**0.9653**	**0.9615**	**0.9912**



Comment: This chart directly compares the proposed ensemble against all five baseline models from Teja & Rayalu (2025). The proposed ensemble achieves the highest values across all metrics, especially ROC-AUC (0.9912 vs. 0.95), demonstrating superior discriminative power and reliability**.**

Predicted \ Actual	**Positive**	**Negative**
Positive	**250**	**11**
Negative	**9**	**299**

- False Negatives = 9 → Only 9 patients with heart disease were missed
- False Positives = 11 → Low rate of unnecessary alarms

This demonstrates high clinical reliability, as minimizing false negatives is critical in screening**.**

CONCLUSION

Interpretation of Results in Context The success of the ensemble stems from its dual-layer optimization strategy. Model level Each classifier was independently tuned via grid search and cross-validation. Ensemble-level Weights were derived from out-of-fold ROC-AUC, ensuring that influence was proportional to empirical reliability. This approach directly addresses the fragility of single-model systems identified by allowing RF and XGBoost to dominate while excluding weaker contributors, the ensemble achieves both high discriminative power and clinical safety. The 9 false negatives are particularly significant. In cardiac screening, missing even a small number of cases can have fatal consequences. Reducing FNs from the high numbers produced by baseline models (e.g., SVM) to only 9 demonstrates tangible progress toward clinically deployable AI. In conclusion proposed optimized ensemble achieved an accuracy of 96.49%, a ROC-AUC of 0.9912, and only 9 false negatives, thereby providing a robust and clinically safer alternative. The study demonstrated consistent reliability with stable performance across stratified 10-fold cross-validation, confirming low variance and robustness to data partitioning.

REFERENCES

World Health Organization. (2025, July 31). *Cardiovascular diseases (CVDs)*. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-%28cvds%29

Topol, E. J. (2019). *High-performance medicine: The convergence of human and artificial intelligence*. *Nature Medicine, 25*(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

Leo Anthony Celi, L. A., Cellini, J., Charpignon, M.-L., et al. (2022). Sources of bias in artificial intelligence that perpetuate healthcare disparities—A global review. *PLOS Digital Health, 1*(3), e0000022. https://doi.org/10.1371/journal.pdig.0000022

Mihaela van der Schaar, M., Alaa, A. M., et al. (2021). How artificial intelligence and machine learning can transform healthcare. *Nature Medicine, 27*, 364–376.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple Classifier Systems* (Lecture Notes in Computer Science, Vol. 1857, pp. 1–15). Springer. https://doi.org/10.1007/3-540-45014-9_1

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics, 21*, 6. https://doi.org/10.1186/s12864-019-6413-7

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.