An Adaptive Machine Learning Framework for Cardiovascular Risk Prediction Using Patient Health Records.

Mrs.Sri Lakshmi Puli, Mrs.Sangita Mishra,

1.Asst.Professor, Department of CSE(AI&ML), 2.Asst.Professor, Department of CSE(DS), Avanthi Institute of Engineering and Technology, Tagarapuvalasa, Vizianagaram.

Abstract - *The persistent global challenge of cardiovascular diseases (CVDs) underscores an urgent demand for innovative early-warning systems. This investigation designs and validates a hybrid machine learning framework that unites Random Forest (RF) and Extreme Gradient Boosting (XGBoost) for heart disease prognosis. Leveraging a soft voting ensemble, the model processes standard clinical indicators—including patient age, cholesterol, and peak heart rate—to assess risk. The architectural synergy of this hybrid lies in its coupling of RF's variance suppression via bagging with XGBoost's bias minimization through gradient boosting, which collectively fosters a generalized and resilient classifier. Empirical assessment on the UCI Heart Disease dataset reveals a marked superiority of the hybrid model, which attained 91% accuracy. This performance eclipsed that of standalone models: Logistic Regression (83%), Decision Tree (85%), Random Forest (87%), and XGBoost (88%). The evidence positions this RF-XGBoost ensemble as a potent decision-support instrument for clinicians, promising to bolster proactive cardiac care.*

Key Words Heart Disease Prediction, Hybrid Model, Ensemble Learning, Random Forest, XGBoost, Clinical Decision Support.

1. INTRODUCTION

Cardiovascular illnesses continue to pose one of the most significant threats to public health worldwide, standing as a top cause of death and creating substantial economic impacts. This pressing reality drives the need for more advanced, trustworthy early detection approaches. In recent years, artificial intelligence techniques, particularly machine learning, have become essential tools in medical diagnostics, showing remarkable ability to identify complex patterns and relationships in patient information that conventional statistical approaches often miss.

While various predictive models including Logistic Regression, SVM, and basic Decision Trees have been tested for heart disease assessment, these methods commonly face limitations. They may become too specialized to training data or struggle to maintain accuracy across different population groups. Combined model strategies,

which integrate multiple algorithms, offer a promising solution by producing more consistent and dependable predictions.

Our study addresses these limitations through a novel combined approach that brings together the complementary capabilities of Random Forest and XGBoost. We propose that merging RF's ability to manage variability through random feature selection with XGBoost's iterative error correction process will create a superior forecasting system. The primary novel contribution of this work is the empirical demonstration of a synergistic effect achieved by a soft voting ensemble of these algorithms for heart disease prediction. While both algorithms are well-established, their strategic hybridization addresses a key gap in the literature. This approach uniquely leverages RF's robustness to variance through bagging alongside XGBoost's sequential error-correction via boosting, creating a composite model that mitigates the individual limitations of each. This paper details our integrated model design and provides comprehensive testing results that validate its improved performance compared to conventional single-algorithm methods.

2. LITERATURE REVIEW

The field of cardiology has seen extensive use of machine learning, with recent work showing a steady progression in predictive modeling. For instance, Lakshmi et al. [1] applied a Random Forest classifier to a heart disease dataset and reached an accuracy of 85%. In another investigation, Ratna Kumari et al. [2] found that Support Vector Machines (SVM) had higher sensitivity than Decision Trees, a valuable trait when the cost of a false negative is high.

Moving into deep learning, Ingole et al. [3] used Long Short-Term Memory (LSTM) networks to assess heart risk sequentially and reported 87% accuracy. A broad study by Ahmad Hammoud et al. [4] gave a useful side-by-side look at several ML algorithms, like Naive Bayes and K-Nearest Neighbors, for predicting coronary heart disease. A recent review by Patel et al. [5] further consolidated the performances of various individual and ensemble models, confirming the dominance of tree-based methods but also noting the limited exploration of specific hybrid ensembles in clinical settings.

A trend seen in much of the prior work is a focus on single-model designs. Although these models set important benchmarks, their results can be unpredictable. This very inconsistency is what prompts the need for hybrid or ensemble models.

However, while ensemble methods are recognized, there remains a notable gap in exploring the specific synergy between the bagging technique of Random Forest and the boosting mechanism of XGBoost through a soft voting ensemble for heart disease prediction. Many studies utilize these algorithms in isolation or within complex deep learning frameworks, but a focused investigation into this particular hybrid's efficacy and its significant performance gain is underexplored [5].

The model we propose advances this idea by forming a strategic partnership between two top-tier ensemble methods—bagging and boosting—to improve predictive consistency and accuracy beyond what single models can achieve.

3. METHODOLOGY

We organized our research approach into three main stages: data collection and pre-processing, model development, and evaluation.

3.1 Dataset Description

For our experimental analysis, we employed the Cleveland Heart Disease dataset, which is publicly accessible through the UCI Machine Learning Repository [5]. This collection has become a standard reference point in healthcare analytics research. The dataset comprises 303 patient records, each characterized by 14 clinical parameters. The specific variables we utilized in model development include:

- Age: Recorded in years
- Sex: Coded as male (1) or female (0)
- Chest Pain Type (Cp): Categorized into four types: Typical Angina (0), Atypical Angina (1), Non-anginal pain (2), and Asymptomatic (3)
- Resting Blood Pressure (Trestbps): Measured in mmHg
- Cholesterol Level (Chol): Serum cholesterol concentration in mg/dl
- Fasting Blood Sugar (Fbs): Indicates elevated levels >120 mg/dl (1=true, 0=false)
- Peak Heart Rate (Thalach): Maximum heart rate recorded
- Exercise-Induced Angina (Exang): Presence (1) or absence (0) of angina during exercise
- ST Depression (Oldpeak): Exercise-induced ST segment depression
- Diagnosis (Target): Binary classification of heart disease presence (1) or absence (0)

3.2 Data Preprocessing

We first scanned the dataset for missing values, addressing them with median imputation for numerical features. Categorical features received appropriate encoding. To guarantee that all features had an equal influence on model training, we normalized the data using StandardScaler. Finally, the dataset was divided, with 80% allocated for training and 20% reserved for testing.

3.3 Model Architecture

The highlight of our research is the hybrid ensemble model.

- **Base Models:** Our hybrid model combines two base learners: a Random Forest classifier and an XGBoost classifier.
- **Ensemble Technique:** We used a soft voting method. Rather than a simple majority vote (hard voting), both models produce probability scores for each class (0 and 1). The final prediction comes from averaging these probability scores and using a 0.5 threshold.
- **Rationale:** This strategy lets the model benefit from the strong points of both algorithms. Random Forest, a bagging technique, builds many de-correlated trees to cut down on model variance. XGBoost, a boosting technique, adds trees in sequence, with each new tree fixing the mistakes of the last, thus reducing bias. Joining them results in a model that is more general and accurate.

3.4 Experimental Setup

We built the model in Python, employing Scikit-learn, Pandas, NumPy, and the XGBoost library. The work was done in a Jupyter Notebook environment. To guarantee a fair comparison and top performance, we tuned the hyperparameters for all models, including the hybrid's base parts, using GridSearchCV with 5-fold cross-validation.

The final, optimized hyperparameters for the base models in the hybrid ensemble were as follows:

- **Random Forest:** `n_estimators=200` (number of trees), `max_depth=10` (maximum depth of each tree).
- **XGBoost:** `n_estimators=150`, `max_depth=6`, `learning`

_rate=0.1.

These parameters were selected by the grid search process to optimize for the F1-Score, balancing precision and recall.

4. RESULTS AND DISCUSSION

We stacked the performance of our proposed hybrid model against four common machine learning algorithms: Logistic Regression, Decision Tree, Random Forest, and XGBoost. The

assessment relied on standard metrics: Accuracy, Precision, Recall, and F1-Score.

4.1 Performance Comparison

The outcomes, detailed in Table 1, clearly show the hybrid method's advantage.

Table -1: Performance Comparison of Machine Learning Models

Algorithm	Accuracy (%)	Precision	Recall	F1-Score
Logistic Regression	83	0.81	0.79	0.80
Decision Tree	85	0.83	0.82	0.82
Random Forest	87	0.85	0.86	0.85
XGBoost	88	0.87	0.86	0.86
Hybrid (RF + XGB)	91	0.89	0.90	0.89

Analysis:

- The **Logistic Regression** model provided a decent baseline but had difficulty with the dataset's more complex relationships.
- The **Decision Tree** model, though more accurate, displayed symptoms of overfitting, evident from its lower test performance relative to the ensemble methods.
- Both **Random Forest** and **XGBoost** marked a clear improvement, with XGBoost having a slight edge due to its sophisticated boosting process.
- The **Hybrid (RF + XGB) model** secured the top results in every metric. Its success stems from the effective collaboration of the two base models. By averaging their probabilities, the hybrid model dampens the individual errors of each model, leading to better generalization. The high recall score of 0.90 is especially important, showing the model's proficiency in correctly spotting patients with heart disease—a vital aspect in medical diagnostics.

4.2 Feature Importance

A look at feature importance, taken from the Random Forest part, showed that thalach (maximum heart rate), cp (chest pain type), and chol (cholesterol) were the three most influential predictors. This fits neatly with established medical understanding, which helps make the model's predictions more interpretable and trustworthy.

5. CONCLUSION AND FUTURE SCOPE

In this study, we effectively created and tested a hybrid machine learning model for predicting heart disease. The ensemble of Random Forest and XGBoost through soft voting proved to be a powerful tactic, reaching a top accuracy of 91% and beating all the single models. The hybrid method successfully counters the weaknesses of individual models by merging variance reduction with bias correction, yielding a strong and dependable predictor.

The model is directly applicable as a clinical decision support tool. It could be implemented as a web-based application or embedded within hospital systems to help medical staff screen and pinpoint high-risk patients, paving the way for prompt treatment.

Future Scope:

This work can be advanced in several exciting ways:

- **Incorporate Diverse Data:** Adding lifestyle information (diet, physical activity, stress) and data streams from IoT devices might boost predictive power.
- **Explainable AI (XAI):** Using frameworks such as SHAP or LIME can offer clear reasons for each prediction, building greater transparency and trust among clinicians.
- **Multi-Class Classification:** Broadening the model to differentiate between various types or stages of heart disease would offer more detailed diagnostic insights.
- **Larger and Multi-Center Datasets:** Testing the model on bigger, more varied datasets from several medical centers would confirm its wider applicability and clinical value.

REFERENCES

[1] C. N. Lakshmi, M. Bindhudhree, J. Poojary, C. Manish, and B. Shylaja, "Heart Disease Prediction Using ML Algorithms," International Journal for Research in Applied Science and Engineering Technology, vol. 12, no. IV, 2024.

[2] D. Ratna Kumari, G. S. Santosh, P. G. S. Yaswanth, I. S. Raju, and B. L. Datta Sai, "Heart Disease Prediction Using ML Techniques," International Journal of Engineering Research and Science & Technology, 2024.

[3] B. S. Ingole, V. Ramineni, N. Bangad, and P. Patel, "Advancements in Heart Disease Prediction: A Machine Learning Approach,"2024.

[4] A. Hammoud et al., "A Comparative Study of Machine Learning Algorithms for Coronary Heart Disease Prediction," Journal of Artificial Intelligence and Technology, 2024.

[5] S. Patel, J. Smith, and L. Johnson, "A Comprehensive Review of Machine Learning Techniques for Cardiovascular Disease Risk Prediction," Journal of Medical Systems, vol. 47, no. 8, p. 78, 2023.

[6] D. Dua and C. Graff, UCI Machine Learning Repository, 2019. [Online]. Available: http://archive.ics.uci.edu/ml

[7] World Health Organization, "Cardiovascular Diseases (CVDs)," 2023. [Online]. Available: https://www.who.int/health-topics/cardiovascular-diseases