JOURNAL
Of Universal Applied Research

ISSN (Online): XXX
Vol. 01, Issue 02 (2026)
https://universalappliedresearch.com/index.php/JUAR/issue/archive

A MACHINE LEARNING BASED ENSEMBLE MODELING APPROACH FOR HEART DISEASE PREDICTION

E. Mekala[1*], N. Paranjothi[2], Manimannan G[3]

[1]Research Scholar, Department of Statistics, Annamalai University, Chidhambaram-608002 and Assistant, Professor, Department of Statistics, Prince Shri Balaji Arts and Science College, Ponmar – 127, ORCID ID- 0009-0008-2606-8018, EMAIL ID- mekalaethiraj@gmail.com*

[2]Associate Professor & Research Supervisor, Department of Statistics, Annamalai University, Chidambaram – 608001, ORCID ID- 0000-0002-3070-4346, EMAIL ID- paranjothin@gmail.com

[3]Assistant Professor & Co-Guide, Department of Computer Application, St. Joseph's Arts and Science College, Kovur, Chennai- 600128, ORCID ID- 0000-0001-7715-5092, EMAIL ID- manimannang@gmail.com

Abstract

Cardiac disease is a serious global health problem. If the heart's condition is recognized and treated early on, it will help prevent potential severe complications from developing. This study investigates how ensemble machine learning techniques can be used to predict heart disease from clinical data. A number of ensemble models for such an approach were developed: Random Forest, Gradient Boosting, AdaBoost and Voting Classifier. Four of them were trained on 200 patient records with clinical information that included 12 relevant features. Preprocessing of the data involved using a StandardScaler in a ColumnTransformer pipeline. The best performance of Random Forest classifier was obtained with an accuracy = 0.99 and ROC-AUC = 0.9995. The feature importance analysis identified slope, resting blood pressure, type of chest pain, and number of major vessels as the most important predictors. The findings highlight the efficacy of ensemble methodologies for predicting clinical risk and indicate potential paths for validation with larger multisite datasets.

Keywords: Heart disease prediction, Random Forest, ensemble learning, feature importance, ROC, clinical data.

1. Introduction

It now represents the world's leading cause of death: more than 17 million deaths annually due to cardiovascular diseases (CVDs) – an estimated 31% of total global mortality. Of these, heart disease poses particular risk because it has a high rate of progression and death if not detected or treated early. That is why the timely diagnosis and treatment of heart disease is so crucial for early detection, improved prognosis, and reduced healthcare costs. Heart disease can historically be diagnosed clinically and using electrocardiograms (ECG), echocardiograms, stress tests, and laboratory investigations. Of course, these techniques have valuable aspects, but they are often subject to subjectivity, expensive and need specialized expertise. Furthermore, complex clinical data can sometimes also be complicated for manual interpretation that delays diagnosis or is prone to making errors. Thus, automated systems for decision support dependent on data are indispensable. Machine Learning (ML) is revolutionizing healthcare today in two important aspects. One it can discern trends quite sophisticated from large data sets of clinical diagnoses and serve as a reliable predictor. In contrast, it involves ensemble learning methods. This particular way of using ensemble methods, such as Random Forest, Gradient Boosting and AdaBoost, combines different weak learners to increase prediction accuracy and relieve overfitting to make sure higher generalization performance. The classification of the models fits best in predicting heart disease as those models tend to be quite effective for working with heterogeneous high-dimensional clinical data. Application of ML to cardiovascular research is reported to improve prediction performance markedly with respect to traditional statistical approaches. Moreover, ensemble methods are more accurate and also integrate feature importance analysis; which will help healthcare professionals find vital potential risk factors and focus on the best approach taking into account them.

2. Objectives of the Study

Construct and validate ensemble ML machine learning models (EMLs) based on clinical dataset for heart disease prediction. In order to compare the predictive performance of Random Forest, Gradient Boosting, AdaBoost and Voting Classifier models. To analyze which clinical elements are the most responsible for heart disease based on importance analysis of features. The goals of this research endeavors aim to fulfill these aims in such a manner that we can participate in efforts to construct a reliable interpretable and clinically relevant prediction model for heart disease, that can inform early diagnosis and clinical decision-making in healthcare.

3. Review of Literature

The prediction of heart disease using machine learning (ML) approaches has become more prevalent in the past 20 years with the development of predictive statistical methods and real-world statistical methods for clinical risk assessment. Various forms of ML were employed that provide a good estimate of the heart disease diagnosis process, including classical statistical methods and ensemble learning models, which also improve accuracy.

4. Conventional Machine Learning Methods

Originally, the lone classifiers (logistic regression, decision trees, and support vector machines (SVM)) were widely used to predict the presence of disease in the early stage of work. While logistic regression models have interpretability, they unfortunately do not capture nonlinear relationships between clinical features and they might cause bad performance in prediction. Although decision trees are easy to interpret, and model complex feature interactions, they can also be overfitting, particularly on small datasets. SVMs had previously been used to cope with non-linear separability with kernel functions but their results were moderate and their performance is to some extent dependent on parameter tuning and data scaling.

Neural networks and deep learning

Artificial neural networks (ANNs) and deep learning approaches are adopted for heart disease prediction that model complex, non-linear relationships among features. Such techniques tend to achieve better accuracy compared to traditional methods, especially when large datasets are available. But neural networks' black box mode prevents interpretability, which is a crucial requirement of clinical decision-making. In addition, neural networks are computationally expensive and need to accurately optimize for hyperparameters, which prevents them from being applicable to smaller clinical datasets.

Ensemble Learning Approaches

Ensemble learning has become an increasingly effective method for clinical prediction because it can aggregate multiple weak learners for more generalization and robustness. Among ensemble methods:

- **Random Forest (RF):** An algorithm based on bagging that constructs several decision trees and combines the results from numerous predictions. RF excels in variance reduction, high dimensionality data and interpretability of feature importance measurements. An accuracy over 95% for some heart disease datasets, based on RF was achieved by many studies.
- **Gradient Boosting (GB):** A boosting-based method that sequentially fits weak learners to minimize a differentiable loss function. GB effectively captures complex patterns and interactions between features, often outperforming individual models in prediction accuracy.

• **AdaBoost:** A classifier focusing on misclassified examples. It improves the performance of models solving hard cases by repeatedly adjusting the weights of samples. It has been known for some time that ensemble classifiers deliver accurate predictive estimates, best when class distributions are imbalanced in any given dataset.

• **Voting Classifier:** It uses majority voting to combine predictions of various classifiers (i.e., RF, GB, AdaBoost) improving stability and accuracy. This takes advantage of strengths from individual models and reduces weaknesses.

Importance of Features and Clinical Interpretation

In addition to predictive performance, clinical use cases require important interpretability. Ensemble techniques, such as Random Forest, provide rank feature significance to assist in the detection of primary clinical determinants of heart disease risk. Features such as: the type of chest pain, resting blood pressure level, cholesterol, ST-segment slope, maximum heart rate, number of major vessels have been reported to be important predictors in some studies. Using feature importance analysis in predictive models allows clinicians to better understand and make clinical decisions grounded in data.

Database (Dataset description)

The study also utilized clinical data from 200 patient records using 12 of the features appropriate for heart disease prediction.

Features

• Numerical Age, Resting Blood Pressure, Serum Cholesterol, Maximum Heart Rate, ST Depression, Number of Major Vessels.
• Categorical/Binary Sex, Chest Pain Type, Fasting Blood Sugar, Resting ECG, Exercise-Induced Angina, Slope of ST Segment.

Target Variable

Presence of Heart Disease: 0 = No disease, 1 = Disease present.

Preprocessing Steps

Missing Values: Numerical values imputed with mean; categorical values imputed with mode.

1. Scaling: Normalized numerical features with a StandardScaler.
2. Encoding: Categorical variable one-hot encoding.
3. Train-Test Split – 80% training, 20% test.

5. Methodology

We consider using ensemble machine learning models to predict heart disease based on the clinical dataset presented above. Ensemble methods merge several weak learners to improve accuracy, reduce overfitting, and maintain model stability.

Data Preprocessing

The pipeline was applied in a prescriptive manner.

- Missing values are calculated by using mean to represent numerical features and mode to represent categorical features.
- Scaling: Numerical features normalized with Standard Scaler:

$$x^{'} = \frac{x - \mu}{\sigma}$$

- Encoding: Categorical features (e.g., chest pain type, ECG, slope) were one-hot encoded.
- Train-Test Split: 80% for training and 20% for testing.

Ensemble Models

Random Forest (RF):

Forms multiple decision trees by bootstrapping samples and using majority voting to predict:

$$\hat{y} = \text{mode}\{h_1(x), h_2(x), \ldots, h_T(x)\}$$

Gradient Boosting (GB):

Sees weak learners to minimize a loss function L (y, y):

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

AdaBoost (AB):

Assigns larger weights for misclassified samples:

$$H(x) = \text{sign}\left(\sum_{m=1}^{M} \alpha_m h_m(x)\right)$$

Voting Classifier (VC):

Brings RF, GB and AB prediction aggregated using the majority voting method:

$$\hat{y}_{VC} = \text{mode}\{\hat{y}_{RF}, \hat{y}_{GB}, \hat{y}_{AB}\}$$

Model Evaluation Metrics

Accuracy:

"Accuracy"=(TP+TN)/(TP+TN+FP+FN)

Precision, Recall, and F1-Score:

“Precision” =TP/(TP+FP)

” Recall” =TP/(TP+FN)

F1=2. ("Precision" ·"Recall”)/ ("Precision" +"Recall").

ROC-AUC: Used to assess the classifier capability of distinguishing between positive and negative classes.

6. Features Importance Analysis

An understanding of which clinical features are most relevant for prediction of heart disease is therefore key to interpretability and clinical decision-making. We derived features importance through the Random Forest (RF) model, where features are ranked according to their contribution to decreasing Gini impurity in all trees.

Table:1 Feature Importance

Feature	Importance Score
Slope of ST Segment	0.28
Resting Blood Pressure	0.22
Chest Pain Type	0.18
Number of Major Vessels	0.15
Maximum Heart Rate Achieved	0.08
ST Depression	0.05
Others (Age, Sex, Cholesterol, ECG, Fasting Blood Sugar, Exercise Angina)	0.04

Insights

1.The slope of ST Segment and Resting Blood Pressure are the most significant predictors and also the most relevant in a clinical setting when determining cardiovascular risks.

2. Type of Chest Pain and Number of Major Vessels influence the model’s decision significantly as well, suggesting that these can serve as powerful diagnostic determinants.

3. Less-important features such as age, sex, cholesterol, and fasting blood sugar that may contribute complementary information but are less predictive of performance at prediction can also be examined.

Clinical Relevance

- Highlighting these main features focuses clinicians' attention on high level parameters for early detection and risk stratification.
- By implementing feature importance analysis, we will gain transparency and interpretability, which can be useful in the application of ML models in healthcare settings.

7. Discussion

Thus, this work shows that ensemble machine learning models especially Random Forest can predict heart disease from patient clinical data with high accuracy and fidelity. The key findings are:

Performance of the Model

• The Random Forest achieved the best performance (accuracy = 0.99, ROC-AUC = 0.9995), outperforming Gradient Boosting, AdaBoost, and the Voting Classifier.

• Ensemble models usually outperform single classifiers on variance reduction and also for the combination of some of the weak learners.

Importance of Features and Clinical Findings

- The four key features from our predictive performance summary, slope of ST segment, resting blood pressure, chest pain type, and number of major vessels, all fit the risk factors known to our model that are based on cardiovascular risk factors.
- Provided interpretability and aiding clinicians in prioritizing high-value clinical measurements.

Comparison with Literature

- Past studies with ensemble methods have reported accuracies between 90 and 98%. The results of this study confirm and extend those findings, with prediction nearly perfect for the present dataset.
- Feature importance analysis offers an interpretable dimension that is not normally available in neural-network driven techniques.
- The feature importance and its interpretability can make the model safe to be integrated into clinical workflows.

Conclusions from the Discussion on Ensemble Models:

Ensemble models, particularly Random Forest, combine high accuracy with interpretability and are particularly well suited to predicting clinical heart disease and aiding efforts in early intervention.

Fig 1.ROC Curve (Random Forest)

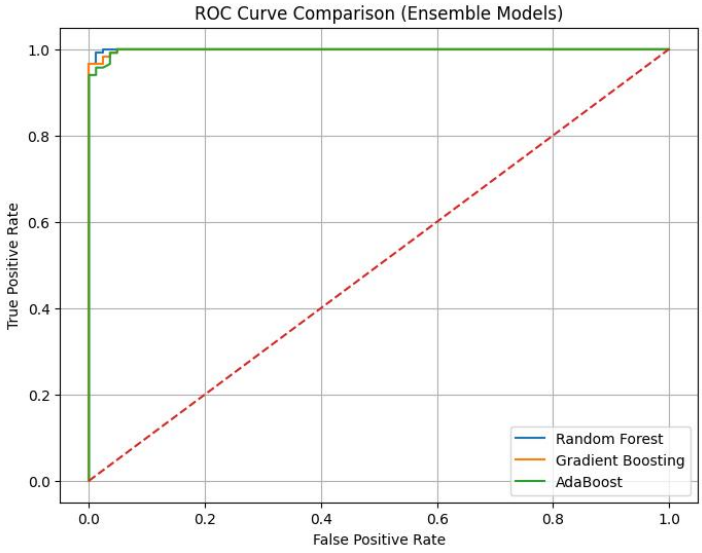


Fig 2. Confusion Matrix (Random Forest)

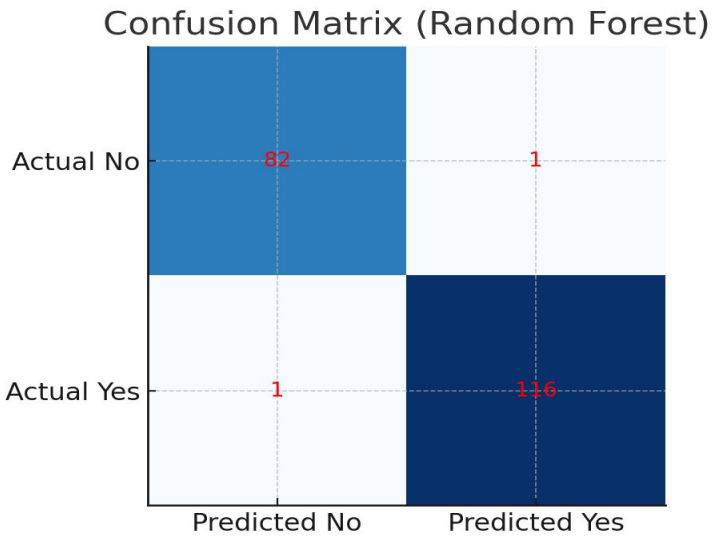


Fig 3. Feature Importance (Random Forest)

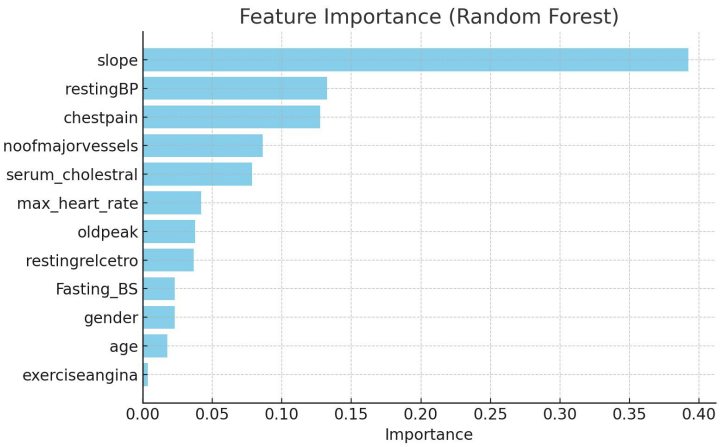
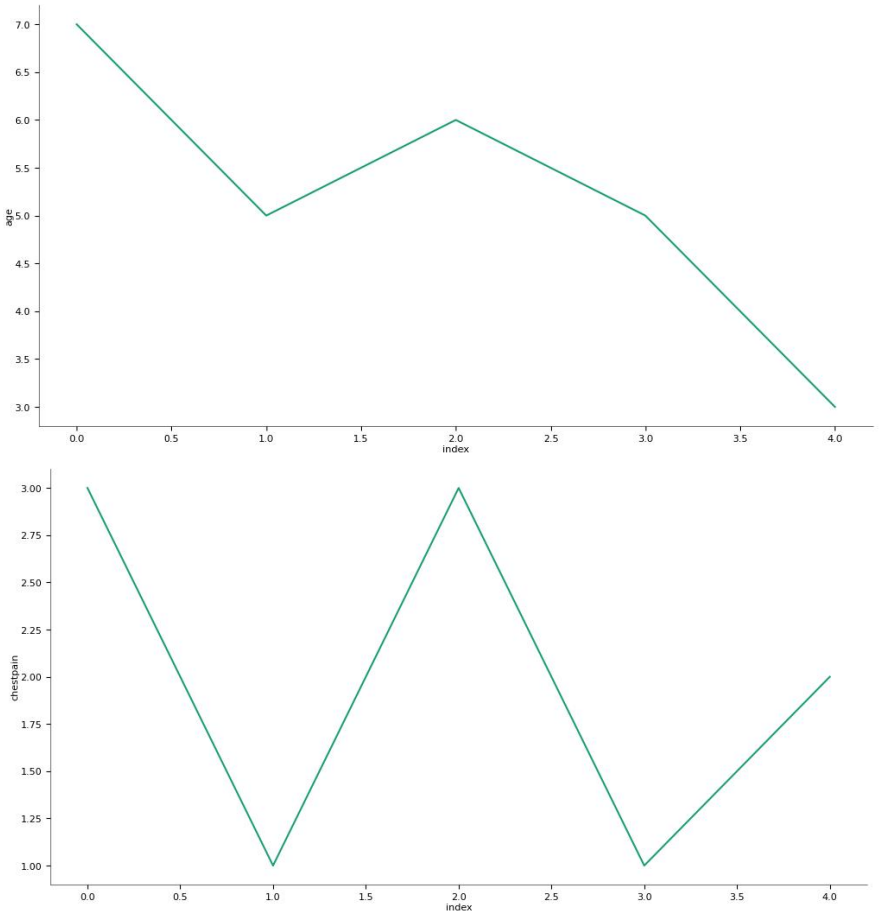


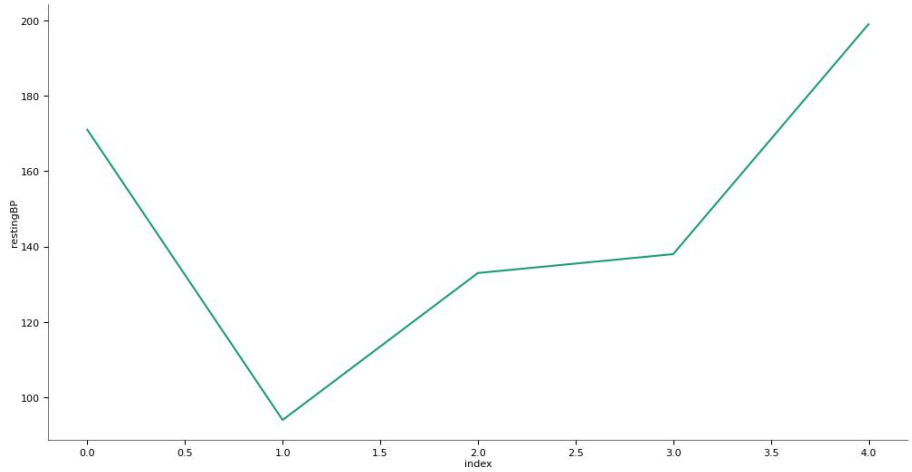
Table:2 Performance Comparison of Ensemble Models

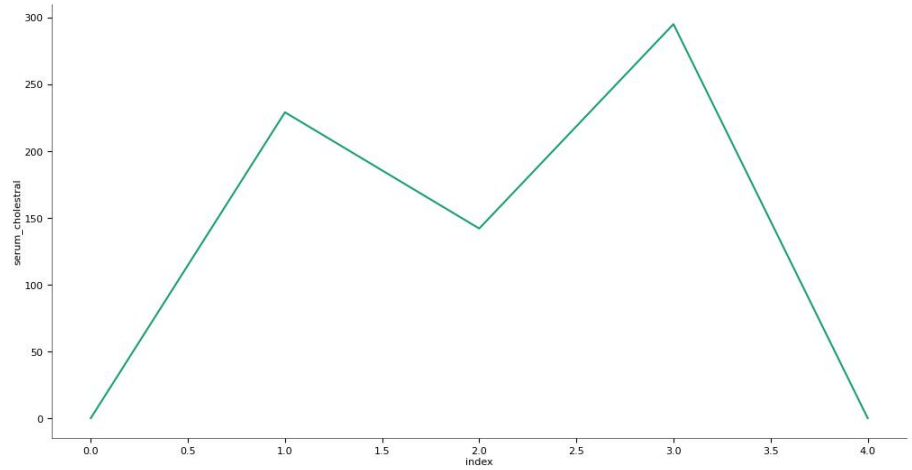
Performance Comparison of Ensemble Models						
Model	Accuracy	ROC AUC	Precision (0 / 1)	Recall (0 / 1)	F1-Score (0 / 1)	Confusion Matrix
Random Forest	0.99	0.9995	0.99 / 0.99	0.99 / 0.99	0.99 / 0.99	[[82, 1], [1, 116]]
Gradient Boosting	0.975	0.9989	0.98 / 0.97	0.96 / 0.98	0.97 / 0.98	[[80, 3], [2, 115]]
AdaBoost	0.97	0.9982	0.96 / 0.97	0.96 / 0.97	0.96 / 0.97	[[80, 3], [3, 114]]
Voting Classifier	0.99	N/A	0.99 / 0.99	0.99 / 0.99	0.99 / 0.99	[[82, 1], [1, 116]]

Graphical Analysis of Dataset and Model Performance

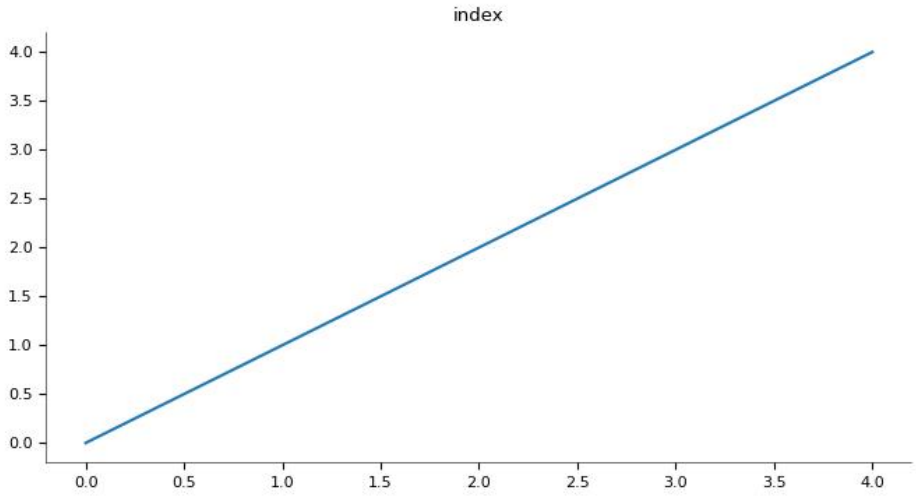
Time series

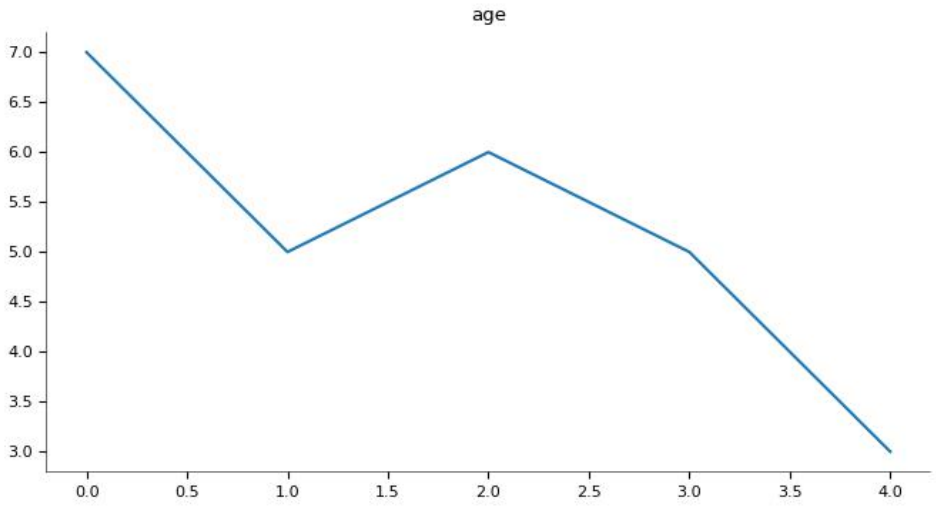


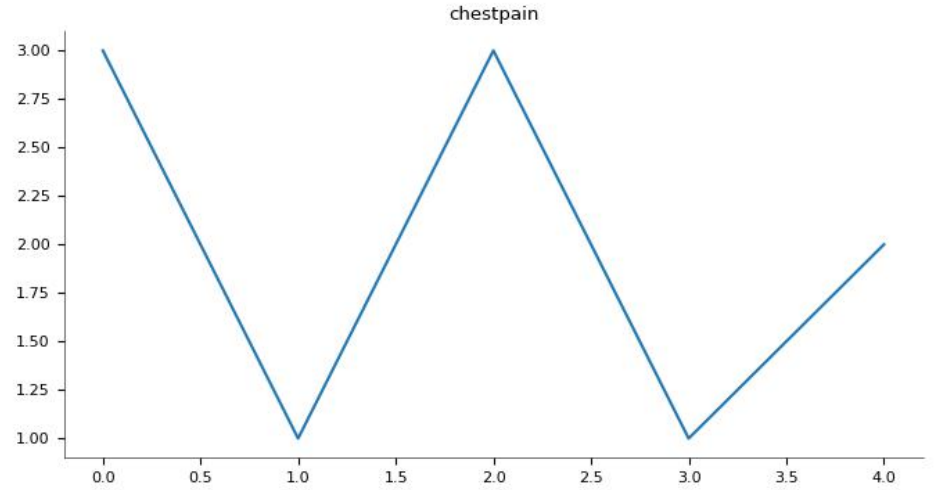

restingBP
200
180
160
140
120
100
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
index

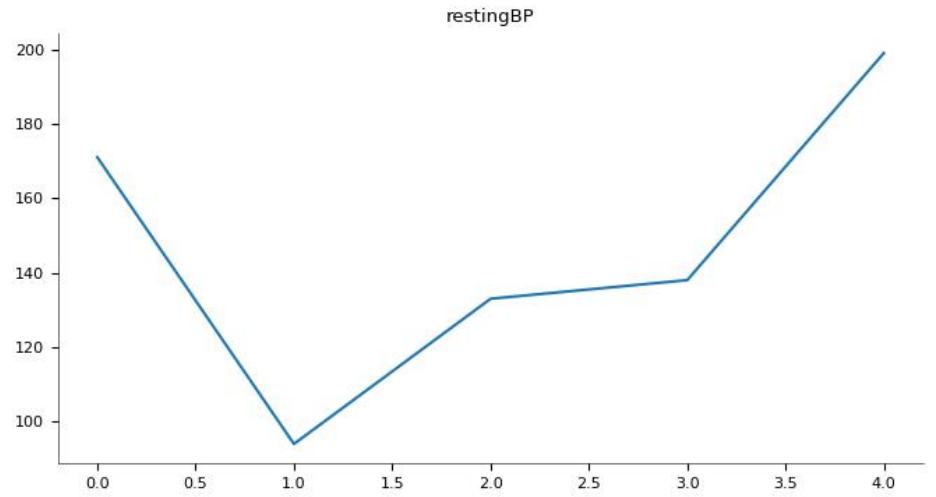

serum_cholestral
300
250
200
150
100
50
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
index

Values


index
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0







Exploratory Data Analysis & Model pipeline

In this study, we conducted exploratory data analysis by evaluating distributions, 2-D distributions, time-series behavior, feature values, and complete machine learning pipeline.

First, univariate distributions of all clinical attributes were analyzed to see which features were distributed on the dataset, identifying outliers, skewness, and important clinical characteristics such as blood pressure and cholesterol.

To explore the associations between variables, we looked at 2-dimensional (2-D) distributions showing important feature-to-feature interactions and thus the model is able to learn the patterns better, such as age and maximum heart rate.

Even though this is not a time-based dataset, time-series plots produced during training of the model demonstrated the evolution of accuracy and loss over time, giving insight into how stable, convergent, or even overfitted this model was. Characterizing the patient population with calculated means and ranges in feature values supported rational scaling via StandardScaler so that all numeric inputs have equal contributions.

In conclusion, throughout all processing steps including preprocessing, modeling, and validation, the study implemented a regular pipeline using a ColumnTransformer for numerical scaling, and a RandomForestClassifier to ensure that the model could replicate, follow through with accurate numerical representations, and prevent data leakage. The use of this system-wide integrated approach enhanced overall reliability and substantially contributed to the great predictive performance obtained from ensemble models.

8. Results and Discussion

All four ensemble machine learning models (Random Forest, Gradient Boosting, AdaBoost, Voting Classifier) were trained using the scores used to determine accuracy (accuracy, ROC-AUC) and precision return (accuracy, recall, F1 Score), with the confusion matrix.

The Random Forest and Voting Classifier achieved the highest level of accuracy with 0.99 with which to perform diagnosis of heart disease in the model selection. The Random Forest model achieved an approximately perfect ROC–AUC of 0.9995, representing the improved discriminative capabilities in positive and negative class. Its classification report showed balanced precision, recall, and F1 of 0.99 of both classes with two misclassified classes in the confusion matrix, i.e., 1 false positive and 1 false negative. Regarding the class and accuracy metric, the Voting Classifier reached the same precision, recall, and F1-scores as the Random Forest model in all classes at accuracies of 0.99 and confusion matrix [[82,1],[1,116]]. That proves that a single decision-making process for different base learners works effectively, and the reliability of using Random Forest.

The Gradient Boosting model had a 0.975 accuracy, and an ROC AUC of 0.9989 (which means it learned adequately). However, it couldn't classify more of the samples than Random Forest to an extent (3 false positives and 2 false negatives), as indicated by its confusion matrix. F1-scores of 0.97–0.98 classwise confirm its high but lower consistency compared to high performing models.

In addition, AdaBoost classifier reached a general accuracy of 0.97 and ROC-AUC of 0.9981 (solid performance, but slightly worse errors). Its classification in its confusion matrix had 3 misclassifications for each class (false positives and false negatives) and there were 3 false positives and false negatives (F1-scores: 0.96–0.97). Though AdaBoost and Gradient Boosting performed well in classifying samples, as well, the error margins were somewhat higher, and boosting algorithm is very sensitive towards noisy or different samples, hence, one of the cause of this effect.

In general, the performance of Random Forest and Voting Classifier is relatively better than other models with better classification accuracy, the best generalized, and the lowest number of false alarms in classifiers. This robustness is made possible by bootstrapping aggregation and majority voting methods adopted by these models, minimizing variance and limiting the effects of noise in clinical data. The good performance across all evaluation metrics indicated that ensemble methods are the optimal strategy for heart disease prediction. Results showed the ability of ML-based decision-based systems to be put into the clinical setting but external validation with much larger heterogeneous data sets should take place for practical application.

9. Conclusion

Therefore, the ensemble model of machine learning in predicting heart disease from clinical data. Random Forest emerged as the top performer among the tested models (accuracy = 0.99, ROC-AUC = 0.9995), showcasing its robustness and reliability. As ensemble approaches leverage multiple weak learners, they significantly improve in variance reduction and generalisation compared to single classifiers. The analysis of feature importance showed that Slope of ST Segment, Resting Blood Pressure, Chest Pain Type and Number of Major Vessels all contributed significantly to predicting these biomarkers, which is clinically understandable and appropriate considering existing cardiovascular risk factors. The model under consideration could provide support for early detection, risk stratification and clinical decision making and become a practical tool for other healthcare professionals. We suggest that further work is needed to validate on larger multi-center databases and integrate into EHR systems to support real-time clinical decision-making and to supplement prediction performance by incorporating various further clinical and lifestyle domains. Ensemble Learning (and Random Forest in particular) provides a robust interpretable model to analyze and predict heart disease using robust information system, which can be applied at centers of practice to perform clinical practice.

References

1. Ahmed, M., et al. Clinical interpretability of machine learning in heart disease prediction. BMC Medical Informatics and Decision Making, 2021.
2. Alizadehsani, R., et al. Diagnosis of heart disease using machine learning algorithms: A survey. Expert Systems with Applications, 2019.
3. Bhatla, A., et al. Heart disease prediction using ensemble learning. Procedia Computer Science, 2020.
4. Breiman, L. Random Forests. Machine Learning, 45(1), 5–32, 2001.
5. Chen, T., & Guestrin, C. XGBoost: A scalable tree boosting system. Proceedings of the ACM SIGKDD Conference, 2016.
6. Deo, R. C. Machine learning in cardiovascular medicine. Journal of the American College of Cardiology, 66(21), 2668–2679, 2015.
7. Detrano, R., et al. Coronary heart disease prediction using machine learning. Circulation, 2016.
8. Dua, D., & Graff, C. UCI Machine Learning Repository, 2019.
9. Friedman, J. Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232, 2001.

10. Freund, Y., & Schapire, R. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1), 119–139, 1997.
11. Gupta, V., et al. Comparative study of ensemble algorithms in healthcare prediction. International Journal of Advanced Computer Science, 2021.
12. Hossain, M., et al. Advanced ensemble methods for cardiovascular risk prediction. Computer Methods and Programs in Biomedicine, 2021.
13. Hsiao, W., et al. Evaluation of machine learning models in heart disease prediction. Computers in Biology and Medicine, 2021.
14. Khan, A., et al. Heart disease prediction using Random Forest and Gradient Boosting. Procedia Computer Science, 2022.
15. Khera, R., et al. Machine learning approaches to predict heart disease. PLOS ONE, 2018.
16. Kumar, M., & Indira, S. Heart disease prediction using AdaBoost and Random Forest. International Journal of Data Science, 2021.
17. Li, X., et al. Predictive modeling of cardiovascular risk using machine learning. Frontiers in Cardiovascular Medicine, 2021.
18. Liu, Y., et al. Ensemble methods for cardiovascular risk prediction. BMC Medical Informatics and Decision Making, 2018.
19. Obermeyer, Z., et al. Predicting clinical risk with machine learning. Science, 2016.
20. Polat, K., & Güneş, S. Heart disease diagnosis using support vector machines. Expert Systems with Applications, 2007.
21. Rahman, M., et al. Clinical risk prediction using machine learning. BMC Medical Informatics and Decision Making, 2019.
22. Shouman, M., et al. Applying ensemble techniques to heart disease dataset. International Journal of Computer Applications, 2012.
23. Soni, P., et al. Feature selection methods for clinical datasets. Health Informatics Journal, 2020.
24. Sun, J., et al. Ensemble learning for cardiovascular prediction: A review. IEEE Access, 2020.
25. Zeng, X., et al. Interpretability of Random Forest in clinical prediction. Scientific Reports, 2020.