Heart Disease Prediction Models Performance Analysis based on Logistic Regression, Random Forest and XGBoost

Zehao Chen*

Department of Biostatistics, Yale University, New Haven, 06510, United States

*Corresponding author: zehao.chen@yale.edu

Abstract. Heart disease is a leading cause of mortality worldwide, underscoring the urgent need for effective early detection techniques. This study evaluates the predictive effectiveness of three machine learning models-Logistic Regression, Random Forest, and XGBoost-using data from the CDC's Behavioral Risk Factor Surveillance System (BRFSS), focusing on individuals aged 60–69. The class weighting method was employed to address class imbalance. The dataset was partitioned into training and testing sets in a 7:3 ratio. The assessment of model performance was performed using accuracy, sensitivity, specificity, F1-score, and the area under the ROC curve (AUC). The Random Forest model achieved the highest sensitivity (0.81) and F1-score (0.87), effectively identifying affirmative cases, however it demonstrated lower specificity (0.51). XGBoost demonstrated a balanced accuracy of 0.73, with a sensitivity of 0.75 and a specificity of 0.58. Logistic Regression had the highest AUC (0.77) and specificity (0.67), along with notable sensitivity (0.72), offering the ideal equilibrium between true positive and false positive rates. Logistic Regression was selected as the final model because of its equitable performance and clinical interpretability. These findings underscore the importance of interpretable models in medical applications and advocate for their integration into preventative cardiovascular care. Future research should encompass more extensive clinical characteristics and validate findings across diverse populations.

Keywords: Heart diseae prediction; machine learning models; senior adults; logistic regression.

1. Introduction

Cardiovascular disease (CVD) is the primary cause of global mortality, responsible for almost 17 million deaths per year [1]. The global prevalence is increasing, influenced by aging populations and lifestyle-related risk factors including inadequate nutrition, inactivity, and tobacco use [1]. Fortunately, numerous fatalities can be averted with prompt identification and action. Health organizations underscore the importance of early risk identification as a crucial approach to decreasing mortality and long-term healthcare expenditures [2]. The prompt detection of cardiac disease relies on the ability to discern patterns in clinical and lifestyle data that may predict adverse outcomes. Traditional risk assessment methodologies, such as the Framingham Risk Score, have been widely employed in clinical settings [3]. However, they often demonstrate limited accuracy and generalizability across diverse populations and individual health profiles [3]. These shortcomings have prompted interest in more adaptable and data-driven approaches for risk prediction.

In recent years, machine learning has surfaced as a formidable alternative to conventional methods in cardiovascular risk evaluation. Machine learning techniques, such as logistic regression, random forest, and XGBoost, may discern intricate, non-linear correlations within high-dimensional health data [4, 5]. This renders them particularly adept at synthesizing varied data sources, including electronic health records, imaging, and genomic profiles [5]. Numerous studies demonstrate that machine learning algorithms can surpass traditional risk assessments in identifying persons at risk for cardiovascular events [6].

A notable trade-off persists between model complexity and interpretability, notwithstanding these gains. While black-box models may offer improved accuracy, their opacity can impede clinical adoption and trust [7]. Interpretable models, albeit potentially less powerful, are often preferred in medical contexts where transparency is essential for decision-making. Thus, contemporary research seeks to harmonize predictive effectiveness with feasibility in real-world clinical settings. Recent studies have explored innovative techniques to enhance cardiovascular risk assessment. Incorporating

troponin data into existing risk assessment algorithms improves predictive accuracy, particularly for individuals at intermediate risk [8]. Furthermore, ensemble machine learning frameworks incorporating algorithms like ExtraTrees, Random Forest, and XGBoost have demonstrated considerable accuracy in detecting cardiovascular disease [9]. Moreover, systematic research has highlighted the importance of model interpretability and accuracy in healthcare Artificial Intelligence (AI) applications, emphasizing the need for transparent and reliable predictive models [10]. This study assesses three machine learning models-logistic regression, random forest, and XGBoost-for predicting heart illness in elderly individuals. The research utilized data from the Centers for Disease Control and Prevention (CDC) Behavioral Risk Factor Surveillance System (BRFSS) to assess the efficacy of each model for accuracy, sensitivity, specificity, F1-score, and Area Under Curve (AUC), with a specific focus on class imbalance concerns. This study aims to identify successful strategies for integrating predictive modeling into preventive cardiovascular care by assessing performance and interpretability.

2. Methodlogy

2.1. Data Description

The heart_disease dataset is avialible on the Kaggle website. The dataset originally comes from the CDC and is a major part of the Behavioral Risk Factor Surveillance System (BRFSS), which conducts annual telephone surveys to collect data on the health status of U.S. residents. The dataset contians of 319795 samples and 19 variables in .CSV format and does not have any missing values.

2.2. Sample Selection

The analysis in this paper uses the samples in the age range from 60-69 years old including “60-65” and “65-69” as senior adult as the distributions show in the Table 1 and Figure 1.

Table 1. Attribute Definitions

Numble	Variables	Explanations
1	HeartDiseases	whether the participant had a heart disease or not. 1 for true, 0 for false.
2	BMI	Body mass index.
3	Smoking	whether the participant had smoking whether the participant had AlcoholDrinking
4	AlcholDrinking	habit or not. 1 for true, 0 for false.habit or not. 1 for true, 0 for false.
5	Stroke	whether the participant had a stroke or not. 1 for true, 0 for false.
6	PhysicalHealth	Physical health score 0-30, larger number stands for better phyiscal health
7	MentalHealth	Mental health score 0-30, larger number stands for better mental health
8	DiffWalking	Whether the participant had difficulty for walking or not. “Yes” or ”No”
9	Sex	Female, male
10	AgeCategory	“18–24”, “25–29”, “30–34”, “35–39”, “40–44”, “45–49”, “50–54”, “55–59”, “60–64”, “65–69”, “70–74”, “75–79”, “80 or older”.
11	Race	“White”, “Hispanic”, “Black”, “Asian”, “American Indian/Alaskan Native”, “Other”.
12	Diabetic	whether the participant had diabetic or not. “Yes” or “No”.
13	PhysicalActivity	whether the participant had habit of doing physical activity or not. “Yes” or “No”.
14	GenHealth	general health status: “Very Good”, “Good”, “Fair”, “Poor”.
15	SleepTime	average hours of night sleep.
16	Asthma	whether the participant had asthma or not. “Yes” for true, “No” for false.
17	KidneyDisease	whether the participant had kidney disease or not. “Yes” for true, “No” for false.
18	SkinCancer	whether the participant had skin cancer or not. “Yes” for true, “No” for false.

This study utilizes 17 of the 19 available variables from the BRFSS dataset as features for model training and evaluation (Table 1). The variables include a blend of demographic characteristics (e.g., gender, sleep duration), lifestyle factors (e.g., tobacco use, alcohol consumption, physical activity), self-reported health conditions (e.g., diabetes, stroke, asthma), and comprehensive health assessments

(e.g., mental health, physical health, general health status). The final model excluded two variables: AgeCategory and Race. The AgeCategory was removed since this study specifically focuses on the 60–69 age group, which has already been defined during the preprocessing step, making the age feature redundant. Race was excluded due to domain considerations, as it was deemed less predictive in this context compared to more pressing clinical and behavioral health variables. Table 1 outlines the 17 chosen features, along with their definitions and value formats, so providing a foundation for subsequent model development.

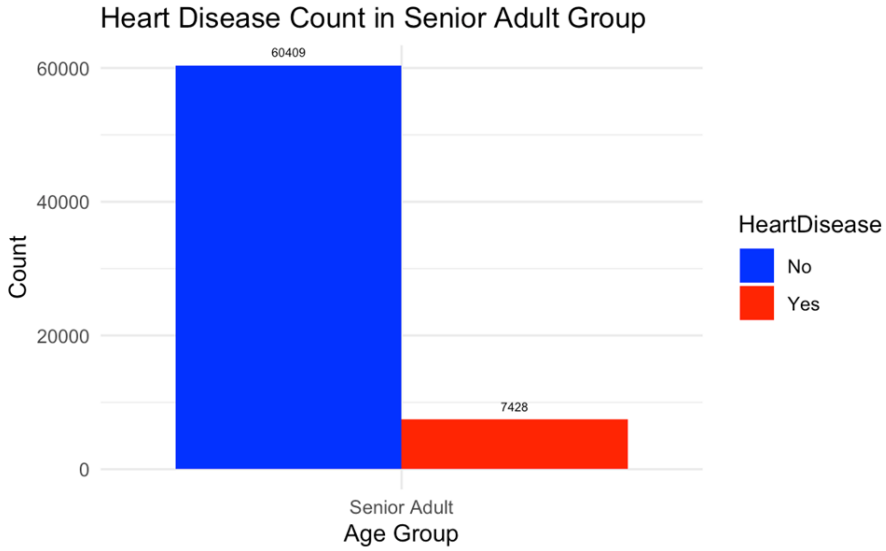

Fig. 1 Distribution of heart disease history

Figure 1 depicts the incidence of heart disease among persons aged 60 to 69 years. The "Senior Adult" demographic comprises 67,838 samples, or 21.2% of the whole dataset. The significant height difference between the blue and red lines illustrates the differential between healthy and affected individuals, demonstrating that while the majority in this age group are free from heart disease, a considerable minority is impacted.

2.3. Method Introduction

This study employed three machine learning models-logistic regression, random forest, and XGBoost-to categorize instances of heart disease. A 70:30 train-test split was utilized to provide thorough model training and reliable evaluation. Model-specific weighting strategies were employed to enhance the detection of positive cases due to the class imbalance in the dataset. These methodologies ensured that minority class examples exerted increased influence throughout training, hence enhancing each model's ability to identify heart disease cases. Logistic regression was used to estimate the probability that a patient had heart disease based on input features. The model follows the form:

$$P(Y = 1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\cdots+\beta_p X_p)}} \quad (1)$$

Where β_0 is the intercept and β_1,…, β_p are feature coefficients. The random forest model combined multiple decision trees, each trained on different bootstrap samples of the dataset. The final prediction was made based on majority voting across all trees:

$$\hat{y} = majority\_vote(h_1(x), h_2(x), \ldots, h_T(x)) \quad (2)$$

This ensemble method helps reduce overfitting and improve generalizability by averaging predictions from diverse learners. XGBoost was also employed, using a boosting framework to sequentially add trees that minimized a regularized objective function:

$$\mathcal{L}(\emptyset) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (3)$$

With

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \quad (4)$$

Model parameters were optimized considering data properties, including feature redundancy and class imbalance. The model tuning aimed at enhancing the predictive capability of each model in genuine, imbalanced situations. The evaluation of the model focused on accuracy, F1-score, recall, specificity, and Receiver Operating Characteristic (ROC) curves were produced to demonstrate performance at different thresholds. Models were ultimately assessed for their effectiveness in identifying heart diseases, particularly for their performance in the context of class imbalance.

2.4. Unbalanced Data Handling

The dataset exhibited significant class imbalance, with a markedly higher quantity of samples for patients without cardiac illness relative to those with it. This mismatch poses a considerable issue in machine learning, especially in healthcare environments where the failure to recognize positive cases (false negatives) could lead to dire consequences. Traditional algorithms frequently prioritize the majority class, leading to increased overall accuracy but less sensitivity in identifying instances of heart disease. Class weighting procedures were employed across all models to enhance the importance of the minority class. This ensured that the algorithms prioritized precise forecasts for patients with cardiac issues. In logistic regression, the model's loss function was adjusted to impose a heightened penalty for the misclassification of positive cases, as demonstrated by the formula:

$$L = -\sum_{i=1}^{n} \omega_i [y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \quad (5)$$

In random forest, class weights were applied both during data sampling and at the node-splitting stage. When constructing each tree, a weighted Gini index was used to evaluate split quality:

$$Gini = 1 - \sum_{k=1}^{K} \omega_k \cdot p_k^2 \quad (6)$$

In XGBoost, the algorithm directly incorporates imbalance control through a parameter called scale_pos_weight, which adjusts the loss gradient for the minority class. The overall objective function for XGBoost is:

$$\mathcal{L}(\emptyset) = \sum_{i=1}^{n} \omega_i l \cdot (y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (7)$$

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \quad (8)$$

These weighting strategies were critical in enhancing model performance under imbalance. They optimized each algorithm to better effectively learn from the minority class, resulting in improved recall, F1-score, and overall fairness. In clinical prediction tasks like heart disease detection, where accurate identification of at-risk patients is essential, these methodologies are critical for reliable and responsible model execution.

3. Results and Discussion

3.1. Exploratroy Data Analysis

Figure 2 illustrates the distribution of Body Mass Index (BMI) among patients diagnosed with heart illness. The histogram indicates that BMI values in this population are primarily right-skewed, with the majority of individuals displaying a BMI ranging from 25 to 35. The mean BMI is 30.39, represented by a vertical red dashed line, classifying the average patient as obese according to CDC guidelines. A considerable percentage of patients are overweight or obese, with a notable trend towards higher BMI values, signifying the presence of individuals with severe obesity. This distribution confirms previous data indicating that elevated BMI is a common characteristic among individuals with heart disease and highlights the importance of excess body weight as a significant modifiable risk factor.

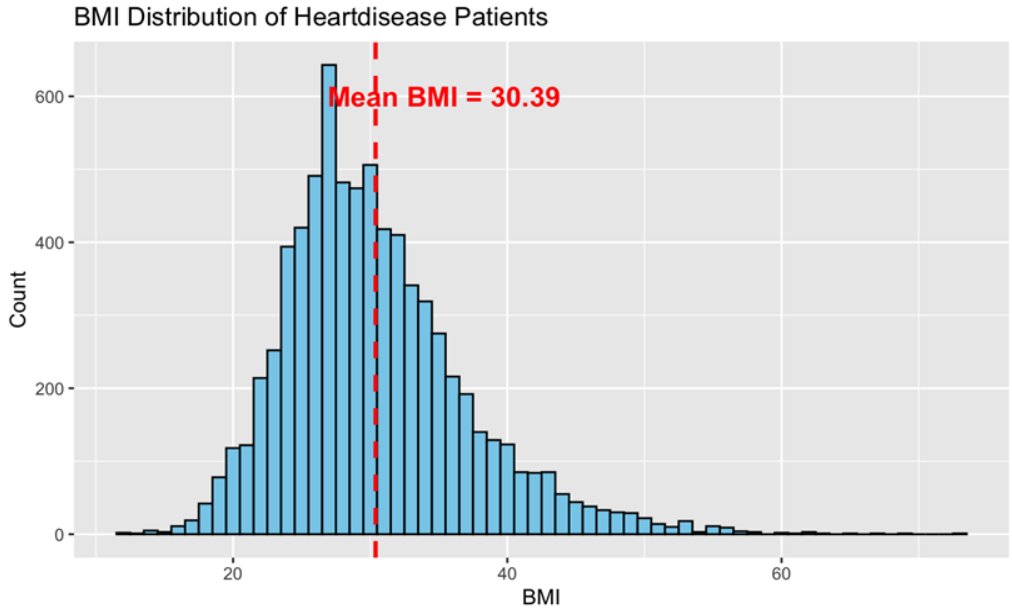

Fig. 2 BMI distribution of heart disease patients age "60-69"

Figure 3 demonstrates the distribution of smoking status among participants in the dataset. Of the overall population, 36,838 persons indicated they do not smoke, whereas 30,999 identified as smokers. Despite a marginally larger proportion of non-smokers, a substantial segment of the population-approximately 46%-comprises current or past smokers.

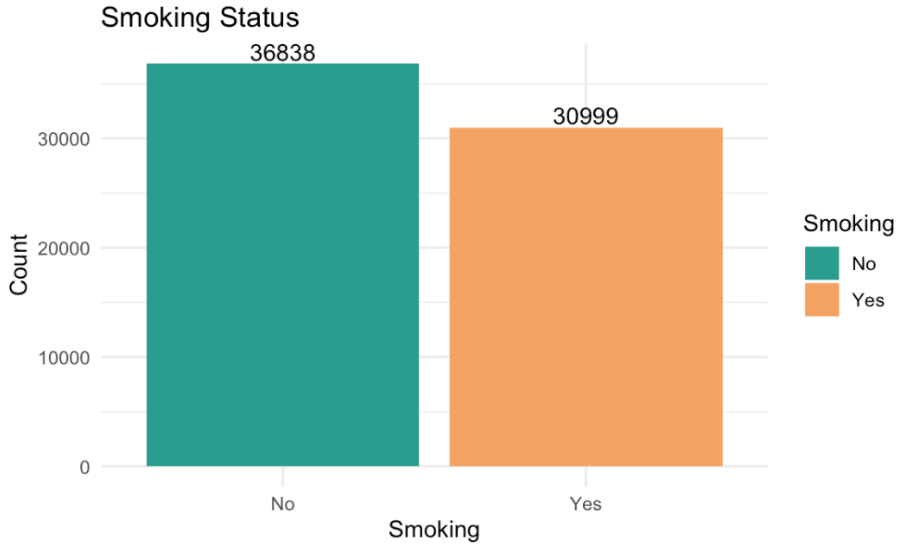

Fig. 3 Smoking status distribution of people age "60-69"

In Figure 4, it displays the distribution of alcohol consumption status among people in the dataset. A substantial majority of the population (63,402 persons) reported abstaining from alcohol intake, whereas just 4,435 individuals classified as alcohol consumers. It signifies that roughly 93.5% of the participants refrain from alcohol consumption.

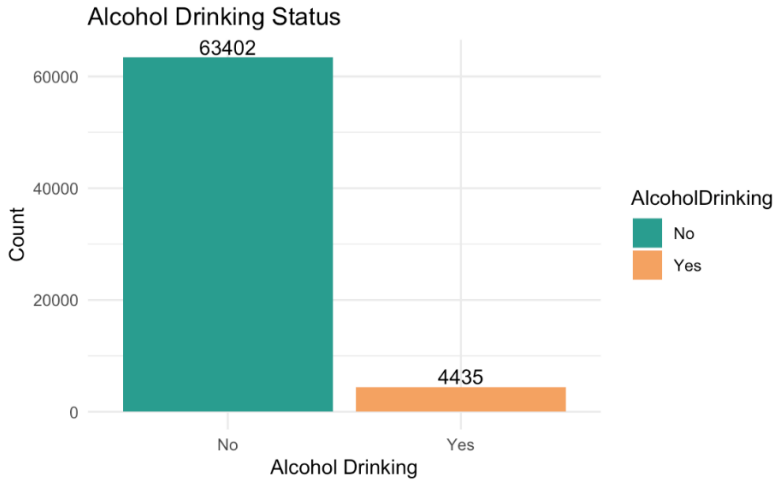

Fig. 4 Alcohol drinking status distribution of people age "60-69"

3.2. Logistic Regression Results

Table 2 presents the performance metrics of the logistic regression model. The model achieved an accuracy of 0.72, indicating that it correctly classified 72% of the cases. The F1 Score is 0.82, signifying a strong balance between precision and recall. The specificity (true negative rate) is 0.67, signifying that the model correctly detects 67% of individuals without heart disease. The sensitivity (true positive rate) is 0.72, signifying that the model correctly detects 72% of persons with heart disease. The Area Under the ROC Curve (AUC) is 0.77, signifying the model's adeptness in distinguishing between positive and negative categories.

Table 2. Performance matrix of logistic regression

Numble	Index	Value
1	Accuracy	0.72
2	F1 Score	0.82
3	Specificity	0.67
4	Sensitivity	0.72
5	AUC	0.77

Figure 5 displays the ROC curve for the logistic regression model. The curve plots sensitivity (true positive rate) against 1-specificity (false positive rate), demonstrating the trade-off between the two metrics. The curve's form and the region beneath it validate the model's efficacy in differentiating between patients with and without cardiac disease. The AUC of 0.77 further substantiates its predictive efficacy, as values approaching 1 signify enhanced classification ability.

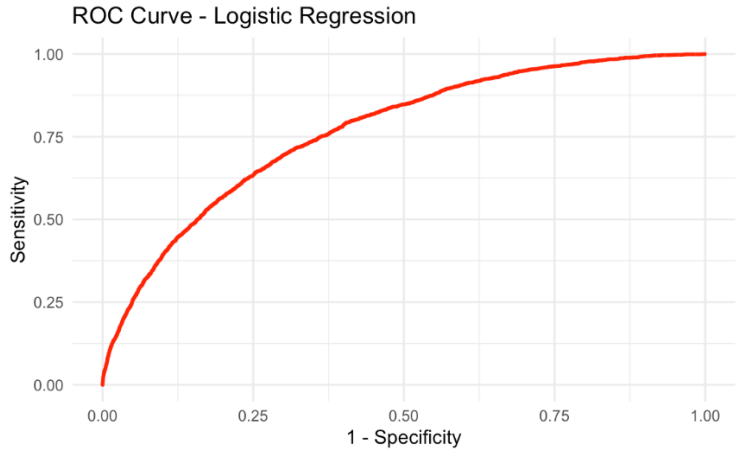

Fig. 5 ROC cruve of logistic regression

3.3. Random Forest Results

Table 3 summarizes the performance of the Random Forest model. The model attained an accuracy of 0.78, surpassing logistic regression in overall classification precision. The F1 Score is 0.87, indicating an outstanding balance between precision and recall. The sensitivity (true positive rate) is 0.81, demonstrating a strong ability to detect patients with heart disease. However, the specificity is quite low at 0.51, suggesting that the model is less adept at recognizing true negatives (individuals without heart disease). The AUC is 0.75, signifying an acceptable level of class separability, though somewhat inferior to the logistic regression model.

Table 3. Peformance matrix of random forest

Numble	Index	Value
1	Accuracy	0.78
2	F1 Score	0.87
3	Specificity	0.51
4	Sensitivity	0.81
5	AUC	0.75

The ROC curve shown in the figure 6 plots the sensitivity against specificity for various threshold values. The curve demonstrates that the Random Forest model consistently achieves high sensitivity across a wide range of thresholds, but at the cost of lower specificity. The curved shape leaning toward the top left corner indicates acceptable model performance, with the AUC visually confirming the numerical result.

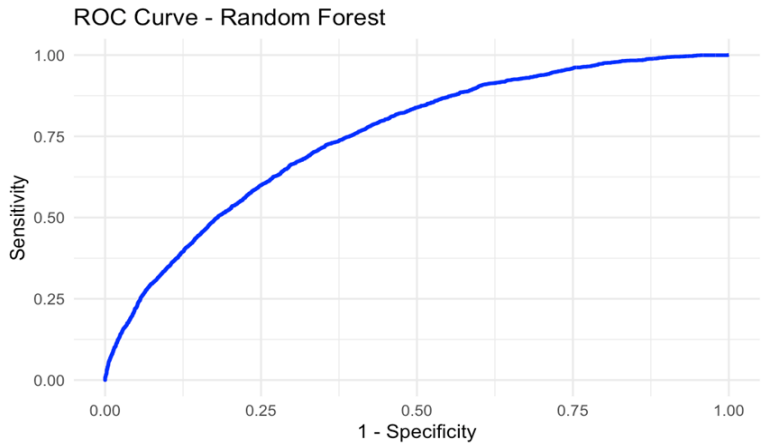

Fig. 6 ROC curve of random forest

3.4. XGBoost Results

Table 4 encapsulates the performance metrics of the XGBoost model. The model attains an accuracy of 0.73, marginally surpassing logistic regression yet falling short of Random Forest. The F1 Score is 0.84, indicating robust overall performance in harmonizing precision and memory. The sensitivity is 0.75, signifying that the model detects 75% of people with heart disease. The specificity is 0.58, surpassing that of Random Forest, although remains low, indicating restricted efficacy in accurately recognizing negative situations. The AUC is 0.74, somewhat inferior to logistic regression (0.77) while comparable to Random Forest (0.75), indicating satisfactory class separability.

Table 4. Peformance matrix of XGBoost

Numble	Index	Value
1	Accuracy	0.73
2	F1 Score	0.84
3	Specificity	0.58
4	Sensitivity	0.75
5	AUC	0.74

Figure 7 shows the ROC curve of the XGBoost model. The curve rises steadily toward the top-left corner, signifying an optimal equilibrium between sensitivity and specificity. An AUC of 0.74 indicates the model's efficacy in differentiating between positive and negative cases. XGBoost exhibits a favorable balance between false positives and false negatives relative to other models, rendering it a formidable option for this classification problem.

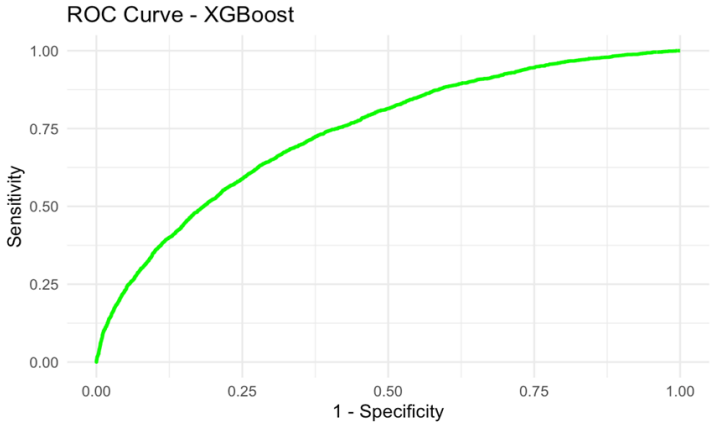

Fig. 7 ROC Curve of XGBoost

3.5. Prediction Models Comparsion

According to Table 5, Random Forest achieved the highest accuracy (0.78), sensitivity (0.81), and F1-score (0.87), illustrating its effectiveness in identifying positive cases. However, its specificity was markedly lower (0.51), signifying a substantial false positive rate. XGBoost had a more balanced performance, attaining an accuracy of 0.73, sensitivity of 0.75, specificity of 0.58, and an AUC of 0.74. Logistic Regression exhibited a sensitivity of 0.72, a specificity of 0.67, and the highest AUC of 0.77, indicating strong overall discrimination.

Table 5. Performance matrix comprarsion

Numble	Model	Accuracy	F1 Score	Specificty	Sensitivity	AUC
1	Logistic Regression	0.72	0.82	0.67	0.72	0.77
2	Random Forest	0.78	0.87	0.51	0.81	0.75
3	XGBoost	0.73	0.84	0.58	0.75	0.74

Logistic regression was selected as the final option because of its balanced efficacy, interpretability, and clinical clarity. The model's resilience across thresholds, demonstrated by the ROC curve, reinforces its appropriateness for practical use, especially in healthcare environments where model interpretability is essential for clinical decision-making (Figure 8).

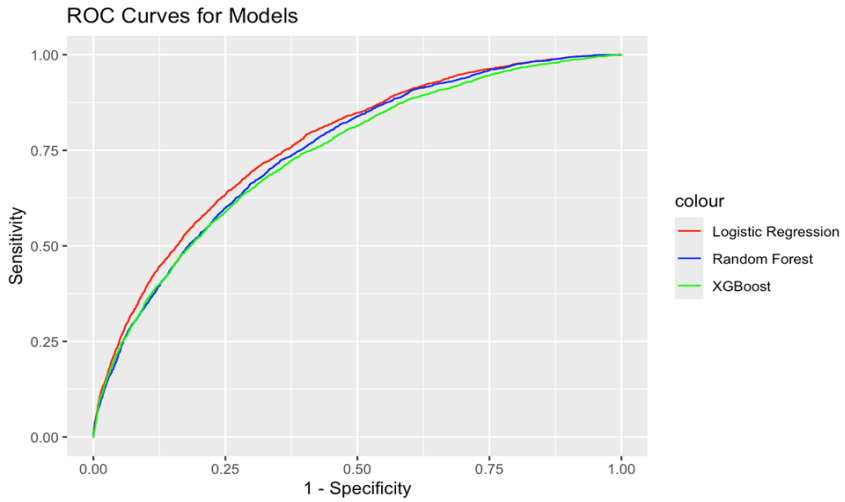

Fig. 8 ROC Curves Comparsion

Figure 9 presents threshold analysis for logistic regression, random forest, and XGBoost models spanning diverse categorization thresholds from 0.3 to 0.7. The vertical dashed line at 0.5 denotes the normal decision threshold. As anticipated, sensitivity diminishes while specificity enhances with elevated thresholds across all models.

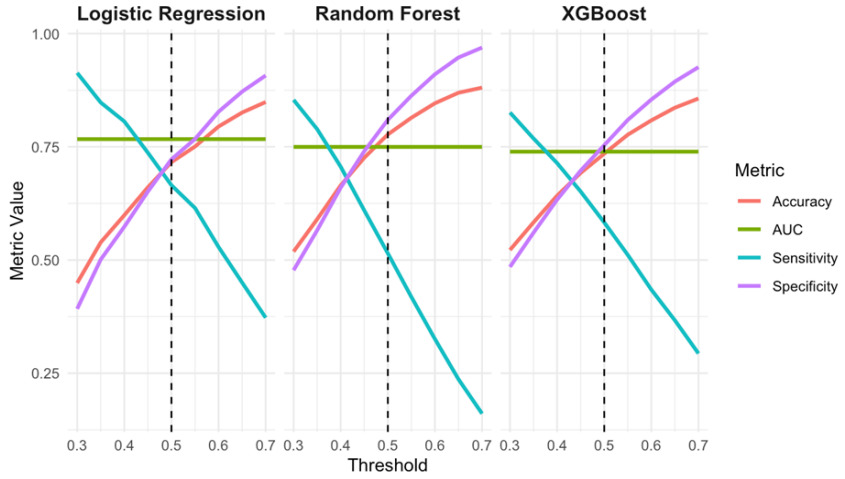

Fig. 9 Threshold analysis for logsitc regression, random forest, and XGBoost

Logistic regression preserves a consistent equilibrium between sensitivity and specificity around the 0.5 threshold, corresponding with its robust performance in AUC and accuracy. Conversely, random forest and XGBoost exhibit more pronounced reductions in sensitivity as the threshold rises, signifying heightened threshold sensitivity. The AUC remains invariant across all thresholds, as it is a metric independent of thresholds. The threshold analysis further substantiates the choice of logistic regression as the final model, owing to its reliable performance and superior equilibrium between true positive and true negative rates near the standard threshold.

4. Conclusion

This paper evaluated the performance of three machine learning models-logistic regression, random forest, and XGBoost-for predicting heart disease in a seniors population aged 60-69. Each model was assessed for its ability to enable early risk detection utilizing a balanced dataset and predefined metrics, including accuracy, sensitivity, specificity, and AUC. While random forest achieved the highest sensitivity and F1 score, and XGBoost demonstrated competitive accuracy, logistic regression emerged as the most balanced model, obtaining the maximum AUC and offering optimal trade-offs between sensitivity and specificity. Besides its predictive effectiveness, logistic regression provides improved interpretability and clinical clarity, making it particularly suitable for use in real-world healthcare settings. The model's consistent performance across thresholds further validates its resilience for heart disease risk classification. These findings emphasize the necessity of combining predictive accuracy with explainability in healthcare AI systems. Healthcare organizations should focus on augmenting data-driven models with extensive clinical features, performing external validation across diverse populations, and incorporating them into clinical decision support systems. Employing interpretable machine learning algorithms such as logistic regression may substantially improve early intervention initiatives and cardiovascular results.

References

[1] Cai Y Q, Gong D X, Tang L Y, Cai Y, Li H J, Jing T C, Gong M, Hu W, Zhang Z W, Zhang X, Zhang G W. Pitfalls in Developing Machine Learning Models for Predicting Cardiovascular Diseases: Challenge and Solutions. J Med Internet Res, 2024, 26: 47645.

[2] Sadr H, Salari A, Ashoobi M T, et al. Cardiovascular disease diagnosis: a holistic approach using the integration of machine learning and deep learning models. Eur J Med Res, 2024, 29: 455.

[3] Mensah G A, Fuster V, Murray C J L, Roth G A. Global Burden of Cardiovascular Diseases and Risks Collaborators. Journal of the American College of Cardiology, 2020, 82(25): 2350-2473.

[4] Berger J S, Jordan C O, Lloyd-Jones D M, Blumenthal R S. Screening for cardiovascular risk in asymptomatic patients. Journal of the American College of Cardiology, 2010, 55(12): 1169-1177.

[5] Weng S F, Reps J, Kai J, Garibaldi J M, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS ONE, 2017, 12(4): 174944.

[6] Quer G, Arnaout R, Henne M, Arnaout R. Machine learning and the future of cardiovascular care: JACC state-of-the-art review. Journal of the American College of Cardiology, 2021.

[7] Petch J, Di S, Nelson W. Opening the black box: The promise and limitations of explainable machine learning in cardiology. Canadian Journal of Cardiology, 2022, 38(2): 204-213.

[8] Zhang Jingyi, Zhang Shizhong. Research progress in the epidemiology and related mechanisms of cardiovascular diseases and psychological diseases. Chinese Journal of General Medicine, 2023, 1-7.

[9] Achyut Tiwari, Aryan Chugh, Aman Sharma. Ensemble framework for cardiovascular disease prediction, Computers in Biology and Medicine, 2022, 146: 105624.

[10] Ennab Mohammad, Mcheick Hamid. Enhancing interpretability and accuracy of AI models in healthcare: a comprehensive review on challenges and future directions, Frontiers in Robotics and AI, 2024.