Comparative Analysis of Logistic Regression and Random Forest Models for Cardiovascular Disease Prediction Using Clinical Data

Keliang Li

{xixuegui@tzc.edu.cn}

School of Economics and Management, Heilongjiang Bayi Agricultural University, Daqing, China

Abstract. Cardiovascular disease remains a significant global health burden, making it crucial to construct accurate risk prediction models. Analyzing a large clinical dataset, this study evaluates the performance of logistic regression and random forest models in predicting cardiovascular and cerebrovascular diseases. Data are divided into training and testing sets, and multiple indicators such as accuracy, precision, recall, F1 score, and area under the ROC curve (AUC) are used to measure the model's effectiveness. Results show that both models exhibited good predictive performance, with slightly better accuracy in random forest classification and outstanding discriminative ability in logistic regression (AUC of 0.78). Studying ensemble models can bring marginal performance improvements, but traditional statistical methods still have significant advantages and interpretability in clinical risk prediction. These conclusions provide clinically actionable guidance for clinicians and researchers to select appropriate machine learning models—such as preferring logistic regression for interpretable risk stratification or random forest for slightly higher accuracy—when screening for cardiovascular and cerebrovascular diseases in clinical practice.

Keywords: Cardiovascular disease; Machine learning; Logistic regression; Random forest; Clinical risk prediction.

1 Introduction

Cardiovascular disease continues to pose a significant threat to global public health, accounting for a high proportion of premature deaths and long-term disabilities. Timely identification of high-risk groups is particularly crucial for implementing effective prevention and intervention measures. Some traditional risk assessment methods often rely on linear algorithm techniques and pre-set clinical boundaries, which cannot fully capture complex connections between population, behavior, and physiological risk factors.

Recently, machine learning techniques have attracted widespread attention in predicting cardiovascular crises, and they have unique abilities to focus on modeling non-linearities, contextual correlations, and complex exchange interactions in high-throughput, multi-level patient data. A systematic literature review emphasizes continued expansion of machine learning models in the stratification of cardiovascular crises and implementation of precise preventive care, while also promoting dual needs of methodological innovation and real-world clinical execution [1]. Given that the heart blood crisis mode of machine learning has been proven to be significantly superior to traditional computing methods on autonomous queues, this highlights

ICIAAI 2026, May 22-24, Singapore, Singapore

DOI 10.4108/eai.22-5-2026.2365083

the key significance of the generalizability of the mode [2]. Similarly, compared to the model approach, the machine learning model that combines traditional and new crisis factors has shown better predictive performance [3].

Ensemble learning methods (e.g., random forest) exhibit substantial potential for cardiovascular disease prediction by capturing complex feature associations and reducing overfitting, often outperforming single classifiers [4]. However, their strong predictive performance is offset by limited interpretability, which hinders clinical acceptance—a critical limitation for real-world deployment.

In clinical practice, the transparency and interpretability of machine learning tools are now ethical and operational imperatives. Recent research has focused on developing interpretable models (e.g., via attribute analysis) to clarify individual risk factors and enhance clinician trust [5], highlighting an ongoing trade-off between predictive performance and interpretability in clinical machine learning.

While prior studies have compared traditional and machine learning models for cardiovascular disease prediction, few have evaluated these approaches on a large-scale, publicly available dataset (70,000 samples from Kaggle) or integrated both performance metrics and clinical interpretability into a unified analysis. This study addresses this gap by systematically comparing logistic regression (valued for its clinical interpretability) and random forest (noted for its predictive accuracy) using a comprehensive set of evaluation metrics. Our focus on real-world clinical interpretability and application scenarios further distinguishes this work from existing literature.

2 Experimental Framework and Methodological Approach

2.1 Dataset Description

This study utilized the publicly available Cardiovascular Disease Dataset from Kaggle (https://www.kaggle.com/sulianova/cardiovascular-disease-dataset), which contains 70,000 de-identified patient records and 13 variables. The dataset includes demographic attributes such as age (originally recorded in days and converted to years for analysis), gender (1 = female, 2 = male), height (cm), and weight (kg); physiological measurements including systolic blood pressure (ap_hi, mmHg) and diastolic blood pressure (ap_lo, mmHg); biochemical indicators such as cholesterol level (1 = normal, 2 = elevated, 3 = severely elevated) and glucose level (1 = normal, 2 = elevated, 3 = severely elevated); as well as lifestyle-related factors including smoking status (0 = non-smoker, 1 = smoker), alcohol consumption (0 = non-drinker, 1 = drinker), and physical activity (0 = inactive, 1 = active). The binary outcome variable (cardio) represents the presence (1) or absence (0) of cardiovascular disease according to the original dataset labeling scheme. A supplementary table (Table S1) provides a comprehensive statistical summary of all variables to support experimental reproducibility.

2.2 Data Preprocessing and Experimental Design

The dataset was processed using Python and the scikit-learn library. All variables in the original dataset were numerical, and no missing values were detected, eliminating the need for data imputation or interpolation. Categorical variables (gender, cholesterol, glucose, smoke, alcohol,

and active) were retained in their original ordinal numerical encoding (e.g., gender: 1 = female, 2 = male; cholesterol and glucose: 1 = normal, 2 = elevated, 3 = severely elevated; lifestyle factors: 0 = no, 1 = yes). One-hot encoding was not applied, as the ordinal structure of these variables is consistent with their clinical interpretation and preserves the interpretability of logistic regression coefficients. All input features were standardized using z-score normalization (mean = 0, standard deviation = 1) prior to model training to improve numerical stability. The dataset was randomly divided into training (70%) and testing (30%) sets using a fixed random seed (42) to ensure reproducibility. Although explicit stratified sampling was not applied, the original dataset exhibits an approximately balanced class distribution, which reduces potential sampling bias during model evaluation.

2.3 Machine Learning Models

Logistic regression and random forest classifiers were implemented using the scikit-learn library. Prior to model training, feature standardization was performed using the StandardScaler, which normalizes input variables to zero mean and unit variance based on the training data distribution.

For logistic regression, L2 regularization was applied with a regularization strength of C = 1.0. The L-BFGS solver was used for model optimization, and class weights were kept at the default setting (None). The maximum number of training iterations was set to 1000 to ensure convergence. No additional hyperparameter optimization was performed.

For the Random Forest model, an ensemble consisting of 500 decision trees (n_estimators = 500) was constructed. The maximum tree depth was set to None, allowing trees to grow until terminal leaf purity was reached. The Gini impurity criterion was adopted for node splitting, and the number of features considered at each split followed the default square root strategy (max_features = sqrt). A fixed random seed (random_state = 42) was applied to ensure experimental reproducibility. No further hyperparameter tuning was conducted.

3 Evaluation of Model Performance

3.1 Dataset Characteristics

Complete data preparation processing, divide samples into two subsets, use 49000 samples for model training, and retain 21000 samples for performance evaluation at a ratio of 70:30. In the test set, 10461 cases were labeled as non-cardiovascular disease cases, and 10539 cases were labeled as cardiovascular disease cases. Moreover, balance helps to reduce classification bias, which enables more accurate evaluation of model performance.

3.2 Evaluation Metrics

To evaluate model performance, multiple standard classification metrics were employed, including accuracy, precision, recall (sensitivity), F1-score, and the area under the receiver operating characteristic curve (AUC). Accuracy was used to assess overall predictive performance, while precision and recall quantified the reliability of positive predictions and the ability to identify true positive cases, respectively. The F1 score provided a balanced evaluation by combining precision and recall. In addition, AUC was used to measure the discriminative capability of the models across different decision thresholds, with higher values indicating

stronger classification performance.

3.3 Model Performance and Analysis

Table 1 summarizes the classification performance of the logistic regression and random forest models on the test dataset. Both models achieved stable predictive results across multiple evaluation metrics, indicating their suitability for cardiovascular disease risk prediction.

Table 1. Performance metrics of logistic regression and random forest on the test dataset.

Model	Accuracy	Precision (Class 1)	Recall (Class 1)	F1-score (Class 1)	AUC
Logistic Regression	0.7198	0.74	0.68	0.71	0.78
Random Forest	0.7286	0.74	0.70	0.72	0.79

As shown in Table 1, the random forest model achieved marginally higher accuracy and recall for the positive class compared with logistic regression, suggesting a slightly improved sensitivity in identifying cardiovascular disease cases. However, the absolute performance differences between the two models were relatively small. Without formal statistical significance testing (e.g., McNemar's test or cross-validation-based performance distribution analysis), these numerical improvements should be interpreted cautiously rather than being regarded as definitive superiority.

To formally evaluate whether these differences are statistically significant, a McNemar test was conducted on the test set predictions. The resulting test statistic was 1271.0 with a p-value of 0.00042, indicating that the difference between the two models is statistically significant.

In addition, the precision, recall, and F1-score values for the negative class were comparable between the two models, indicating balanced classification performance for both disease and non-disease cases. Inspection of the confusion matrices further showed that false negative and false positive errors remained within a similar range across models, which is clinically important because missed diagnoses and unnecessary alarms both carry potential risks.

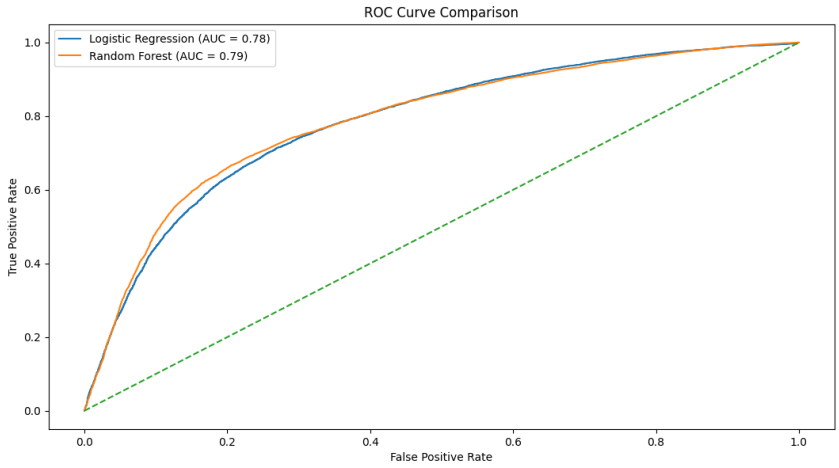

Figure 1. ROC curves of logistic regression and random forest models for cardiovascular disease prediction

Receiver operating characteristic (ROC) curve analysis was performed for both logistic regression and random forest models to further evaluate their discriminative capabilities. As illustrated in Figure 1, both models demonstrated strong separation between positive and negative cases across different classification thresholds. The random forest model exhibited a slightly higher AUC value, while logistic regression maintained competitive discriminative performance with an AUC of 0.78, highlighting the robustness of both approaches.

4 Discussion

The study demonstrates that both logistic regression and random forest models achieve reliable predictive performance for cardiovascular disease risk. Logistic regression attained an AUC of 0.78, indicating stable discriminative ability, while Random Forest achieved slightly higher accuracy and recall, reflecting marginal improvement in sensitivity for positive cases. These results suggest that, although ensemble methods can offer subtle performance gains, simpler statistical models remain competitive and interpretable for clinical applications [2, 7]. Examination of logistic regression coefficients indicates that several features have linear associations with the target variable, which may explain the limited advantage of more complex ensemble methods. Recent studies have also demonstrated that decomposing features into interpretable components can improve understanding of individual risk contributions, which aligns with the principles applied in this study [6].

Several limitations should be acknowledged. First, the analysis is based on a single public dataset, which may introduce selection bias and limit generalizability. Second, external validation was not performed, leaving the clinical applicability across different populations uncertain [8]. Third, the feature engineering employed was relatively simple, focusing primarily on raw clinical and lifestyle variables, potentially restricting predictive performance. Fourth,

comparisons with other state-of-the-art models, such as XGBoost or neural networks, were not conducted, leaving room for performance improvements [5].

Future studies should address these limitations by incorporating multiple and heterogeneous datasets to reduce selection bias, performing external validation to assess generalizability, and employing advanced feature engineering techniques, including longitudinal data and additional biomarkers. Furthermore, comparative analyses with more sophisticated machine learning models could provide insight into potential performance gains. The integration of explainable artificial intelligence methods is recommended to enhance model interpretability and clinician trust. Such efforts will support the development of robust, transparent, and clinically actionable predictive frameworks for cardiovascular risk assessment [9, 10].

5. Conclusion

This study conducted a comparative analysis of logistic regression and random forest models for cardiovascular and cerebrovascular risk assessment using a comprehensive clinical dataset. Both models demonstrated stable performance, with Random Forest showing slightly higher accuracy, while Logistic Regression achieved superior discriminative power (AUC) and interpretability. These results indicate that traditional statistical methods, such as logistic regression, remain highly valuable in clinical prediction tasks, especially where transparency and simplicity are prioritized.

The practical implications of these findings suggest that logistic regression can provide actionable insights for clinicians, supporting evidence-based decision-making and improving patient outcomes. At the same time, ensemble methods offer complementary advantages in sensitivity and classification stability.

Future research should aim to enhance model interpretability, evaluate generalizability across diverse patient populations, and incorporate additional clinical indicators and multimodal data. Improving robustness and validating performance in real-world clinical settings will be crucial to safely deploying machine learning–based cardiovascular risk prediction tools.

References

[1] Gautam, N., Mueller, J., Alqaisi, O., Gandhi, T., Malkawi, A., Tarun, T., Alturkmani, H. J., Zulqarnain, M. A., Pontone, G., & Al'Aref, S. J. (2023). Machine learning in cardiovascular risk prediction and precision preventive approaches. *Current Atherosclerosis Reports*, *25*, 1069–1081. https://doi.org/10.1007/s11883-023-01174-3

[2] Cai, Y., Cai, Y.-Q., Tang, L.-Y., Wang, Y.-H., Gong, M., Jing, T.-C., Li, H.-J., Li-Ling, J., Hu, W., Yin, Z., & Gong, D.-X. (2024). Artificial intelligence in the risk prediction models of cardiovascular disease and development of an independent validation screening tool: A systematic review. *BMC Medicine*, *22*, 56. https://doi.org/10.1186/s12916-024-03273-7

[3] Cheng, C.-H., Lee, B.-J., Nfor, O. N., Hsiao, C.-H., Huang, Y.-C., & Liaw, Y.-P. (2024). Using machine learning-based algorithms to construct cardiovascular risk prediction models for Taiwanese adults based on traditional and novel risk factors. *BMC Medical Informatics and Decision Making*, *24*, 199. https://doi.org/10.1186/s12911-024-02603-2

[4] Teja, M. D., & Rayalu, G. M. (2025). Optimizing heart disease diagnosis with advanced machine learning models: A comparison of predictive performance. *BMC Cardiovascular Disorders*, *25*, 212.

https://doi.org/10.1186/s12872-025-04627-6
[5] Ahiduzzaman, M., & Hasan, M. N. (2025). Interpretable machine learning for cardiovascular risk prediction: Insights from NHANES dietary and health data. *PLOS ONE*, *20*(11), e0335915. https://doi.org/10.1371/journal.pone.0335915
[6] Yu, L., Wu, J., Wu, X., Chen, C., Chen, Y., Fang, L., & Zheng, D. (2025). Interpretable machine learning model for cardiovascular disease risk prediction: A feature decomposition-based study. *BMC Public Health*, *25*, 3639. https://doi.org/10.1186/s12889-025-24921-4
[7] Rehman, M. U., Naseem, S., Butt, A. U. R., Mahmood, T., Khan, A. R., Khan, I., Khan, J., & Jung, Y. (2025). Predicting coronary heart disease with advanced machine learning classifiers for improved cardiovascular risk assessment. *Scientific Reports*, *15*, 13361. https://doi.org/10.1038/s41598-025-96437-1
[8] Zhang, Y., Li, S., Wu, W., Zhao, Y., Han, J., Tong, C., Luo, N., & Zhang, K. (2024). Machine-learning-based models to predict cardiovascular risk using oculomics and clinic variables in KNHANES. *BioData Mining*, *17*, 12. https://doi.org/10.1186/s13040-024-00363-3
[9] Sadr, H., Salari, A., Ashoobi, M. T., & Nazari, M. (2024). Cardiovascular disease diagnosis: A holistic approach using the integration of machine learning and deep learning models. *European Journal of Medical Research*, *29*, 455. https://doi.org/10.1186/s40001-024-02044-7
[10] Liu, T., Krentz, A. J., Lu, L., & Curcin, V. (2025). Machine learning based prediction models for cardiovascular disease risk using electronic health records data: Systematic review and meta-analysis. *European Heart Journal – Digital Health*, *6*(1), 7–22. https://doi.org/10.1093/ehjdh/ztae080