Dean&Francis
ISSN 2959-6157

From Data to Decision: Explainable Risk Prediction for Cardiovascular Diseases Using Multicenter Patient Records

Jiaming Ou

Department of Statistics and Data Science, Beijing Normal-Hong Kong Baptist University, Zhuhai, 519087, China
s230005042@mail.uic.edu.cn

Abstract:

Cardiovascular disease (CVD) remains one of the leading global causes of mortality, highlighting the critical need for early prediction to reduce fatality rates. This study utilizes a publicly available CVD dataset to develop and compare three supervised learning models—Lasso-regularized logistic regression, random forest, and an ensemble model (Stacking)—for assessing individual disease risk. Through comprehensive preprocessing, including interaction terms and dummy variable encoding, this research enhanced model expressiveness and feature representation. The experimental results demonstrate robust predictive performance across all models, with the Stacking ensemble achieving the highest accuracy (90.00%), surpassing logistic regression (87.78%) and random forest (89.44%). Feature importance analysis further reveals ST depression induced by exercise (Oldpeak), Slope of peak exercise ST segment (ST_slope), and maximum heart rate achieved during exercise (MaxHR) as the most influential predictors. These findings not only validate machine learning's effectiveness in CVD risk assessment but also emphasize the value of feature engineering and model assembling in boosting predictive accuracy. The study provides a reliable framework for clinical decision support, potentially enabling earlier interventions and improved patient outcomes.

Keywords: Cardiovascular disease prediction; Lasso logistic regression; Random forest; Machine learning.

1. Introduction

Cardiovascular Diseases (CVDs) remain the leading cause of death globally, accounting for 31% of annual mortality worldwide (approximately 17.9 million deaths). Projections indicate a 73.4% increase in crude mortality rates by 2050 [1]. While traditional risk assessment tools such as the Framingham Risk Score and ASCVD model have laid the groundwork for preventive strategies, they exhibit notable limita-

ISSN 2959-6157

tions in capturing nonlinear relationships, incorporating novel biomarkers, and addressing multicenter data heterogeneity [2, 3]. For instance, these models often overlook psychosocial factors (e.g., depression) and synergistic effects among metabolic markers (e.g., triglyceride-glucose index) [4, 5].

Recent advances in Machine Learning (ML) have demonstrated remarkable performance in CVD risk prediction. For example, ML algorithms leveraging electronic health records achieved 91.7% accuracy in hypertension risk prediction, while deep learning models outperformed conventional methods in predicting 2-year post-myocardial infarction survival [6]. In a UK Biobank study of 229,000 participants, integrating clinical features with metabolomic data significantly improved the C-index for CVD mortality prediction to 0.822 [7,8]. Additionally, dynamic ECG analysis enabled real-time myocardial ischemia monitoring post-PCI, achieving 73.6% accuracy [5]. However, variability in diagnostic standards and population structures across centers poses challenges for model generalization and multicenter data integration [3]. A 15-year cohort study revealed that individuals with high total cholesterol variability faced a 20% independent increase in CVD risk, suggesting that fixed-threshold models may underestimate risk in high-variability subpopulations [9].

Progress has also been made in model interpretability. A study of 5.4 million fatty liver patients combined ML with logistic regression to predict carotid plaque formation, achieving an AUROC of 0.831 using just 5 key features (e.g., age, LDL-C) [7]. Similarly, SHAP value analysis in a Chinese diabetic cohort identified age and cystatin C as top predictors, with the model's Harrell's C-statistic (0.923) significantly surpassing Cox regression (0.890) [6]. Nevertheless, generalizability remains limited due to single-center sampling or lack of ethnic diversity (e.g., UK Biobank) [7, 8].

This study analyzes a multicenter CVD dataset (n=918) collected from four regions, incorporating 11 clinical features with interaction analysis to enhance predictive performance. The proposed framework aims to develop an interpretable and robust risk assessment model for early intervention in cardiovascular care.

2. Data Description

This study integrates five previously unmerged public datasets from Kaggle/UCI repositories to create one of the largest multicenter cardiovascular datasets in current research, encompassing 11 key clinical features for heart disease prediction [10]. The original data sources are shown in Table 1.

Table 1. Data Sources and Record Counts for the Multicenter Cardiovascular Dataset

Original Data Source	Number of Record
Cleveland	303
Hungary	294
Switzerland	123
VA Long Beach	200
Stalog (Heart)	270
Total (After Merging)	918

Note. The initial dataset included 1,190 records before removing 272 duplicates.

3. Variable Definitions

The variables in the research are essential for predicting cardiovascular disease. They include basic demographics, clinical measurements, and health outcomes. Each variable is defined to ensure clarity and reproducibility of the study's predictive model.

The study utilizes a harmonized cardiovascular disease dataset (n=918) sourced from the UCI Machine Learning Repository, comprising 11 key clinical features. These include demographic variables (Age, Sex), cardiovascular indicators (Chest Pain Type categorized as Typical angina (TA), Atypical angina (ATA), Non-anginal pain (NAP) or Asymptomatic (ASY); Resting Blood Pressure (RestingBP); Cholesterol levels, metabolic markers (FastingBS), electrocardiographic measurements (RestingECG results including Normal, ST abnormalities or LVH), exercise response parameters (Max HR achieved, Exercise Angina occurrence, Oldpeak ST depression), and diagnostic outcomes (Heart Disease status). The dataset also incorporates functional capacity assessments through ST Slope measurements during peak exercise (Up, Flat, Down). This comprehensive feature set enables multidimensional analysis of both traditional risk factors and

functional cardiovascular parameters for robust predictive modeling.

4. Data Preprocessing

The outliers of the data set are detected, and the box diagram of the numeric data in the data set is drawn, and the obvious outliers are observed through the image. After that, use the Z-score method to detect abnormal values. Those with Z-score greater than 3 or less than -3 are considered abnormal values. Clean up these abnormal values and delete them. Finally, the classification variable is converted into a dummy variable. The purpose is to convert the classification variable into a numerical form, so that it is easier to process these data later.

Table 2 provides the encoding and descriptions of the variables used in the cardiovascular disease prediction model. This encoding is essential for statistical analysis and machine learning algorithms to process categorical data effectively.

Table 2. Variable Encoding and Descriptions

Variable	Levels (Original Labels)	Integer Encoding
Sex	F (Female), M (Male)	1, 2
ChestPainType	ASY (Asymptomatic), ATA (Atypical Angina), NAP (Non-Anginal Pain), TA (Typical Angina)	1, 2, 3, 4
RestingECG	LVH (Left Ventricular Hypertrophy), Normal, ST (ST-T abnormality)	1, 2, 3
ExerciseAngina	N (No), Y (Yes)	1, 2
ST_Slope	Down, Flat, Up	1, 2, 3

Note. The integer encodings for each variable facilitate the application of machine learning techniques, allowing for the conversion of qualitative data into a numerical format that can be easily analyzed.

Table 3 presents a sample of the cardiovascular disease dataset after preprocessing steps, including the removal of missing values and outliers. Table 3 illustrates the distribution of key variables among the included records.

Table 3. Sample Data from the Cardiovascular Disease Dataset

Variable	Value					
ID	1	2	3	4	5	6
Age	46	50	37	46	54	39
Sex	2	1	2	1	2	2
ChestPainType	2	3	2	1	3	3
RestingBP	140	160	130	138	150	120
Cholesterol	289	0	283	214	195	339
FastingBS	0	0	0	0	0	0
RestingECG	2	2	3	2	2	2
MaxHR	172	156	98	108	122	170
ExerciseAngina	1	1	1	2	1	1
Oldpeak	0.0	0.0	0.0	1.5	0.0	0.0
ST_Slope	1	2	3	2	1	3
HeartDisease	0	0	0	1	1	0

Note. The sample data shown in Table 4 is representative of the overall dataset, demonstrating the range and variability of the variables included in the analysis.

5. Preliminary Data Analysis and Visu-

ISSN 2959-6157

al Presentation

5.1 Distribution Visualization

The age distribution of heart disease cases was visualized using a dual-color scheme to distinguish between affected and unaffected groups, as shown in Fig. 1.

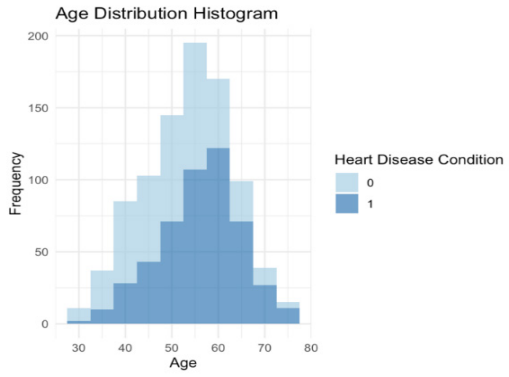

Fig. 1 Age distribution histogram by heart disease condition (Photo/Picture credit: Original).

The majority of samples fell within the 40-70 age range, with disease prevalence showing a progressive increase with age, particularly between 50-60 years. Histograms revealed a positive correlation between age and disease incidence, suggesting elevated risk in older populations.

Evaluation of chest pain types (ASY, ATA, NAP, TA) showed ASY presenting the highest disease association (>75%), followed by TA and NAP, while ATA showed the lowest (<20%), as clearly demonstrated in Fig. 2.

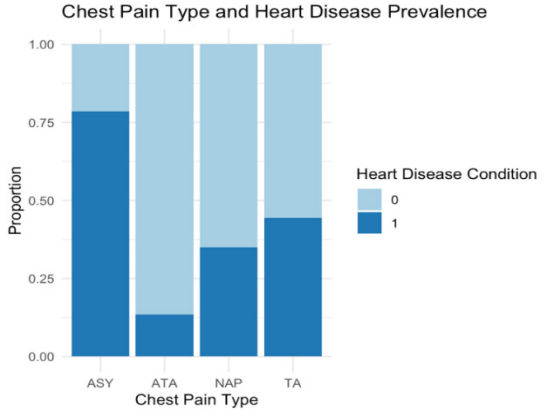

Fig. 2 Chest pain type and heart disease relationship (Photo/Picture credit: Original).

Violin plots (Fig. 3) illustrated cholesterol level distributions, indicating potentially elevated values in affected groups. Maximum heart rate analysis revealed depressed values among cardiac patients, possibly reflecting impaired cardiac function.

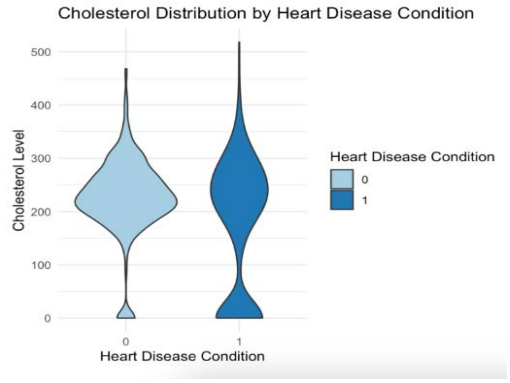

Fig. 3 Max heart rate distribution by heart disease condition (Photo/Picture credit: Original).

The analysis of maximum heart rate revealed distinct distribution patterns between groups, as visualized in Fig. 4. While affected individuals (represented by dark points) appeared clustered in both lower (100-150 bpm) and higher (150-200 bpm) ranges, unaffected subjects (light points) demonstrated a more uniform distribution across the spectrum.

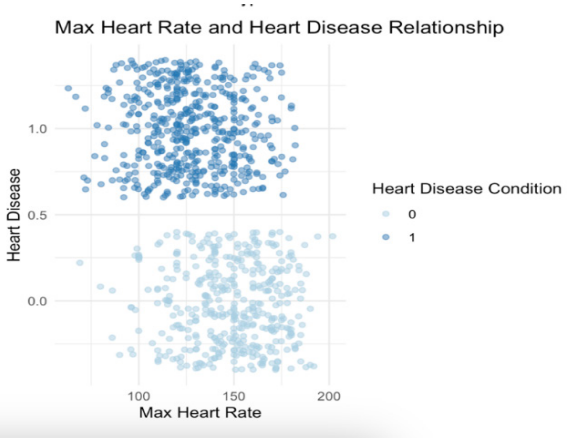

Fig. 4 Cholesterol distribution by heart disease condition (Photo/Picture credit: Original).

5.2 Correlation Analysis

The correlation analysis employed Spearman's rank correlation coefficients (ranging from -1 to 1) to quantify monotonic relationships between variables. Age showed a positive correlation with heart disease (0.294), while RestingBP demonstrated a weak positive association (0.111). In contrast, Cholesterol exhibited a slight negative correlation (-0.141). FastingBS displayed a moderate positive correlation (0.268), whereas Maximum heart rate (MaxHR) revealed a strong negative correlation (-0.410). Notably, Oldpeak showed the strongest positive correlation among all variables examined (0.425).

These relationships were further visualized in Fig. 5, which employs a red-blue gradient heatmap (red: positive; blue: negative) to illustrate correlation strengths. Diagonal elements appeared deep red representing perfect autocorrelation. Age, fasting glucose and Oldpeak showed positive disease associations, while MaxHR demonstrated negative correlation. RestingBP and cholesterol displayed minimal correlation strength.

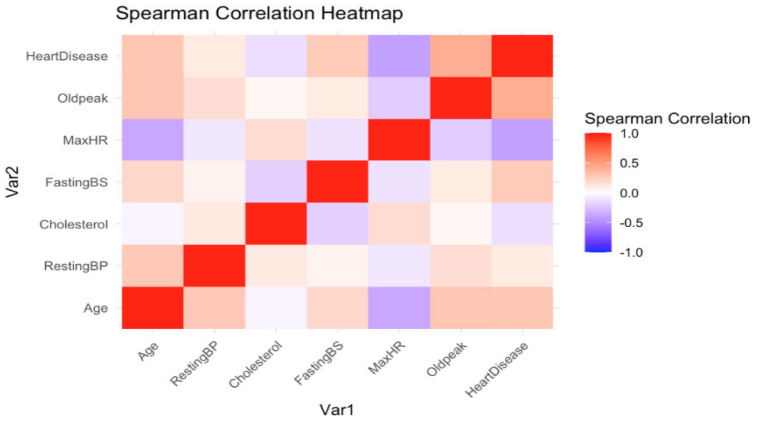

Fig. 5 Spearman correlation heatmap (Photo/Picture credit: Original).

6. Methodology

To evaluate the predictive capacity of various clinical factors for heart disease, it developed two primary classification models: logistic regression and random forest. The logistic regression model was selected for its interpretability and feature selection capabilities, while the random forest approach was employed to capture complex nonlinear relationships and assess feature importance. Below, it details the model specifications and performance evaluation outcomes.

7. Model Construction and Optimization

7.1 Logistic Regression: A Foundational Tool for Feature Selection

Given its interpretability and capacity for significance analysis, logistic regression served as the primary predictive modeling approach. The dataset was partitioned into training and testing sets at an 8:2 ratio. During training, it implemented Lasso (L1) regularization to enable feature selection and model simplification. With over 900 initial variables, Lasso regression effectively compressed irrelevant or redundant features by shrinking their coefficients to zero, thereby enhancing the model‘s generalizability and stability.

The evaluation results demonstrated 88.98% test-set accuracy (95% CI: 83.36%-93.08%). The model significantly outperformed random classification (No Information Rate: 56.11%, p<2e-16), with a Kappa coefficient of 0.7744 indicating strong label-prediction agreement. McNemar‘s test (p=1) confirmed the absence of statistically significant classification bias. These performance metrics are detailed in Table 4.

Table 4. Key performance metrics of logistic regression model

Matric	Value
Sensitivity	87.34%
Specificity	90.10%
PPV/NPV	87.34%/90.10%
Balanced Accuracy	88.72%

While demonstrating robust performance suitable for clinical screening, the 38.33% detection rate suggests room for improvement.

7.2 Random Forest: Capturing Nonlinear Relationships

To better model complex nonlinear patterns, it developed

ISSN 2959-6157

a random forest classifier (500 trees) for cardiovascular risk prediction. All variables except the outcome were included as features, with post-training importance analysis revealing each predictor's relative contribution, as detailed in Table 5.

Table 5. Key Performance Metrics of Random Forest Model

Matric	Value
Accuracy	88.89% (consistent CI)
Kappa	0.7725
Sensitivity	83.54%
PPV/NPV	90.41%/87.75%

McNemar's test (p=0.2636) indicated non-significant error differences versus logistic regression. While showing comparable overall performance, the random forest demonstrated particular strengths in handling high-dimensional data and complex variable interactions, as illustrated in Fig. 6.

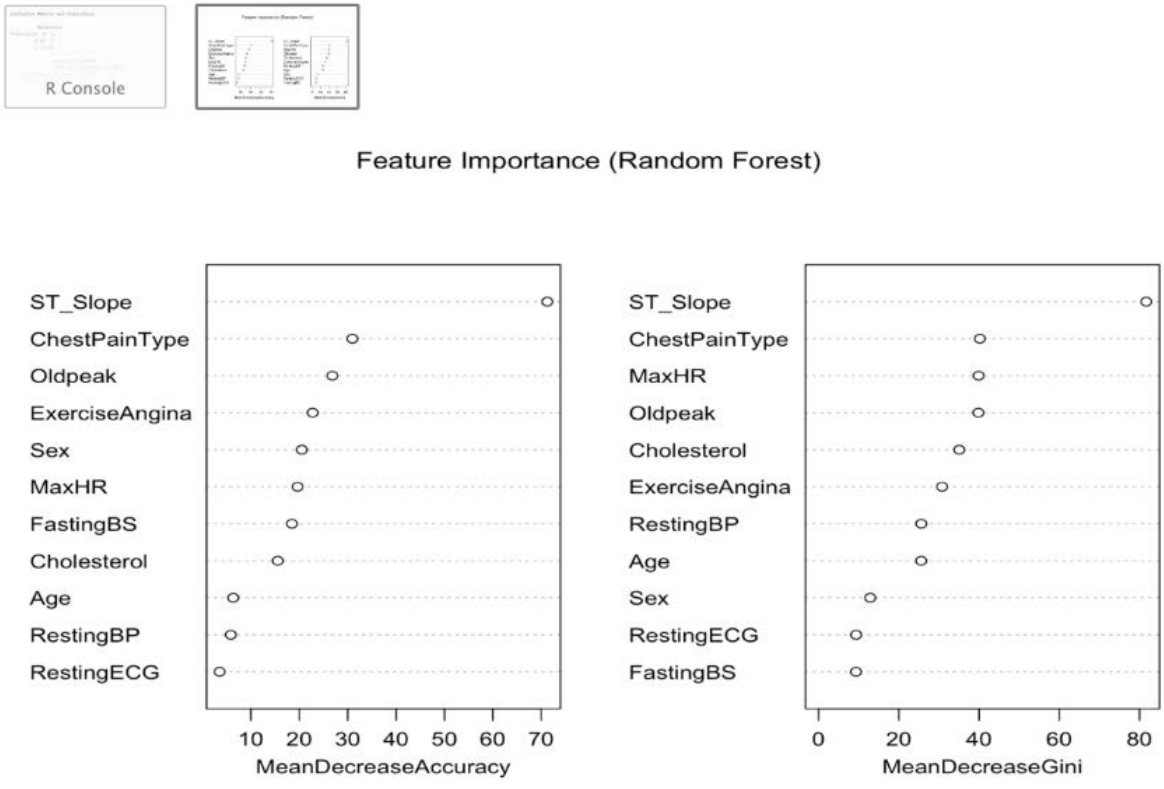

Fig. 6 Random Forest feature importance (Photo/Picture credit: Original).

7.3 Ensemble Strategy and Refinement

Both models demonstrated comparable accuracy, with logistic regression offering superior interpretability for initial screening, while random forest provided enhanced positive predictive value (PPV) for case verification. The optimization framework incorporated several key improvements. For logistic regression refinement, it added interaction terms to better address linearity assumptions and fine-tuned the Lasso regularization strength to optimize feature selection. In random forest tuning, it adjusted class weights to handle imbalanced data and optimized both the maximum features per tree and minimum leaf samples to improve generalization. Additionally, it implemented model stacking by combining prediction probabilities from both models as meta-features, which were then used to train a meta-learner for final classification. This comprehensive approach enhanced the overall predictive performance while leveraging the strengths of each model.

For clinical implementation, it recommends a two-stage

approach: high-sensitivity logistic regression screening followed by random forest confirmation of high-risk cases to reduce false positives. This hybrid strategy balances interpretability with predictive power while addressing each model's limitations.

8. Results

The logistic regression model, with its optimal λ parameter selected through cross-validation, identified several significant predictors through non-zero coefficients. Key variables included age, cholesterol levels, maximum heart rate (MaxHR), Oldpeak measurement, along with important interaction terms such as age of the patient with serum cholesterol and the slope of the peak exercise ST segment with maximum heart rate, all demonstrating substantial predictive value for cardiac risk assessment.

For the random forest model, after implementing class weights to address sample imbalance, variable importance analysis revealed „Oldpeak", „ST_Slope", „MaxHR", and „ChestPainType" as the most influential features contributing to model decisions.

The ensemble approach, which combined predictions from both logistic regression and random forest models as meta-features, achieved superior performance compared to individual models (as shown in Table 6). This enhanced performance indicates complementary learning capabilities - where logistic regression effectively captures linear relationships while random forest excels at identifying complex nonlinear patterns in the data distribution.

Table 6. Key Performance Metrics of Random Forest Model

Model	Accuracy	Kappa	Sensitivity	Specificity
Logistic Regression	87.78%	0.7512	84.81%	90.10%
Random Forest	89.44%	0.7842	84.81%	93.07%
Stacking Model	90.00%	0.7958	86.08%	93.07%

This comparative analysis demonstrates how different modeling approaches can provide unique insights into cardiovascular risk prediction, with the ensemble method leveraging the strengths of both constituent models for optimal performance. The consistency of certain key variables (Oldpeak, MaxHR) appearing as important predictors across different methodologies reinforces their clinical relevance in cardiac risk assessment.

9. Discussion

The comparative analysis of the three models reveals key insights into cardiovascular disease (CVD) risk prediction. The logistic regression model demonstrates robust performance, confirming that even after variable selection and interaction term enhancement, linear models retain strong predictive capability. The random forest model, however, achieves superior specificity by leveraging nonlinear relationships and automated feature selection, while the ensemble model combines the strengths of both approaches, attaining the highest overall accuracy (90.00%). Notably, all models consistently identify Oldpeak, ST_Slope, and MaxHR as the most influential predictors, reinforcing their clinical significance in CVD risk assessment. The inclusion of interaction terms (age of the patient with serum cholesterol and the slope of the peak exercise ST segment with maximum heart rate) further enhanced logistic regression performance, highlighting the importance of considering variable interdependencies.

Despite these strengths, several limitations must be acknowledged. First, the models were trained on a single public dataset (n=918), and while internally validated, multicenter studies with larger cohorts are needed to ensure generalizability. Second, the cross-sectional nature of the data restricts dynamic risk assessment, as longitudinal tracking of disease progression was unavailable. Third, while random forest and ensemble models offer higher accuracy, their „black-box" nature may hinder clinical adoption compared to the interpretability of logistic regression. Finally, the study did not incorporate cost-sensitive learning, which could better account for the severe consequences of false-negative diagnoses in clinical practice.

Future research should focus on external validation using real-world electronic health records (EHR), integration of temporal modeling (e.g., LSTM networks) for dynamic risk prediction, and incorporation of personalized features (e.g., genetic markers, lifestyle factors) to enhance individualized assessments, as conceptually illustrated in Fig. 7. Additionally, ensemble diversification with algorithms like XGBoost could further optimize predictive performance while maintaining clinical interpretability. These advancements will be critical in translating machine learning models into practical, reliable tools for early CVD detection and precision intervention.

ISSN 2959-6157

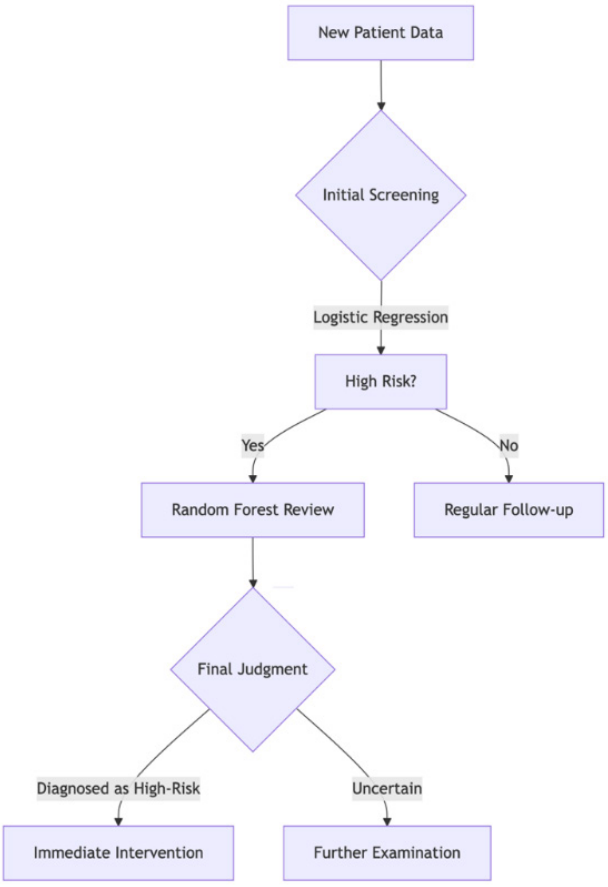

Fig. 7 Clinical Decision (Photo/Picture credit: Original).

10. Conclusion

This study developed predictive models for cardiovascular disease using Lasso-regularized logistic regression, random forest, and an ensemble approach. All models demonstrated strong predictive accuracy, with the ensemble model achieving optimal performance (90% accuracy). Key clinical variables, including Oldpeak, maximum heart rate, and ST segment slope, were consistently identified as significant predictors, warranting particular clinical attention.

The findings highlight two critical insights: that ensemble modeling strategies can effectively enhance predictive performance and that incorporating interaction terms and rigorous feature engineering substantially improves traditional model efficacy. Future research directions should focus on expanding sample sizes and incorporating temporal/longitudinal data to develop more robust personalized prediction models.

References

[1] Roth G. A., Mensah G. A., Johnson C. O., Addolorato G., Ammirati E., Baddour L. M. Global burden of cardiovascular diseases and risk factors, 1990–2019: Update from the GBD 2019 study. Journal of the American College of Cardiology, 2020, 76(25): 2982–3021.

[2] Kelly B. B., Fuster V. Promoting cardiovascular health in the developing world. National Academies Press, 2010.

[3] OECD. Is cardiovascular disease slowing improvements in life expectancy? OECD Publishing, 2020.

[4] Tousoulis D., Stefanadis C. Biomarkers in cardiovascular diseases. Taylor & Francis, 2013.

[5] Liu R., Sun Q., Pang J., et al. Dynamic learning-enabled ECG for evaluating PCI efficacy in acute coronary syndrome. Chinese Journal of Emergency Medicine, 2022, 31(7): 922–929.

[6] Pulkkinen M., Varimo T. J., Hakonen E. T., Hero M. T., Miettinen P. J., Tuomaala A. K. During an 18-month course of automated insulin delivery treatment, children aged 2 to 6 years achieve and maintain a higher time in tight range. Diabetes Obesity and Metabolism, 2024, 26(6): 2431–2438.

[7] Walsh J., Cave J., Griffiths F. Combining topic modeling, sentiment analysis, and corpus linguistics to analyze unstructured web-based patient experience data: Case study of Modafinil experiences. Journal of Medical Internet Research, 2024, 26: e54321.

[8] Singh C. Interpretable machine learning in real-world contexts. University of California, Berkeley, 2022.

[9] Wilemon K. A., MacDougall D. E., McGowan M. P., Howard W., Myers K. D. The vast majority of high- and very-high risk hypercholesterolemia patients never reach below LDL-C thresholds in the 2018 ACC/AHA guidelines. Journal of Clinical Lipidology, 2023, 17(4): e4–e5.

[10] Fedesoriano. Heart failure prediction dataset. Kaggle, 2021.