Proceedings of ICMMGH 2026 Symposium: Cell and Molecular Biology: Understanding Immune Cell Functions
DOI: 10.54254/2753-8818/2026.PJ30865

Heart Disease Risk Prediction via Feature Engineering and Stacking Ensembles

Conghao Wang

Faculty of Life Sciences and Medicine, King's College London, London, United Kingdom
k23166965@kcl.ac.uk

Abstract. Cardiovascular diseases (CVD) cause many deaths in the world. Heart disease is one important type of CVD. Classical tools such as the Framingham Risk Score use only a few variables and a simple linear form, so they cannot show complicated links between many risk factors. Electronic medical records are now very common, and machine-learning methods give another way to build risk prediction models from this data. This study uses the UCI Heart Disease (Cleveland) dataset to build a heart disease risk prediction method that puts feature engineering together with a stacking ensemble. It start from 14 clinical variables and design five enhanced features and three new composite features from blood pressure, cholesterol, ST depression, chest pain type, and exercise-induced angina. Logistic regression (LG), random forest (RF), XGBoost, support vector machine (SVM), and k-nearest neighbours (KNN) are used as base models in a two-layer stacking structure. It judges model performance with AUC, accuracy, precision, recall, and F1-score. The results show that the engineered features give small but clear improvements for most base models, and the stacking model has the best performance with a test AUC of 0.8771. Under a small sample size, the method keeps good accuracy and stability, and may be a useful tool for early screening of high-risk groups.

Keywords: cardiovascular disease prediction, heart disease, feature engineering, stacking ensemble learning.

1. Introduction

Cardiovascular diseases (CVD) cause many deaths in the world [1,2]. Heart disease, which includes problems like coronary heart disease and heart attacks, is one important type of CVD and takes up a large share of global deaths [1,2]. Usual risk factors include high blood pressure, high cholesterol, smoking, excess body weight, and diabetes. In many low- and middle-income areas, these factors often appear together and are not well controlled. So it is hard to judge a person's real risk level only by simple rules [2].

Traditional linear models, for example the Framingham Risk Score, have been used for cardiovascular risk assessment for many years [3]. However, they assume linear relations between predictors and outcomes and they use only a small number of variables. They cannot model non-linear effects or interactions between several risk factors at the same time. In the last decade, many studies have used machine-learning methods to build cardiovascular risk prediction models from

DOI: 10.54254/2753-8818/2026.PJ30865

structured clinical data [4-6]. Tree-based models like RF and XGBoost often reach higher AUC than classic risk scores when they predict cardiovascular events, so these data-driven methods usually work better [4-6].

There are still some important gaps. Many studies use only raw clinical variables and do not design composite features that match pathophysiological mechanisms [7]. This limits how well the models can find hidden patterns. Some studies only test a single model and do not make full use of ensemble learning to get more robust results [4-6]. Stacking ensembles can join several base models and may reduce variance and bias, but careful tests of stacking structures for heart disease prediction are still quite few [8-10]. Also, the interpretability of these models is often not fully discussed, and this makes it harder to use them in real clinical work.

To deal with these problems, this study uses the UCI Heart Disease (Cleveland) dataset and builds a risk prediction method that puts feature engineering together with a stacking ensemble [7]. At the data level, it designs five enhanced features and three new composite features from key clinical variables, such as blood pressure, cholesterol, ST depression, chest pain type, and exercise-induced angina, based on known cardiovascular risk factors [2,6,11-14]. At the model level, five common classifiers are used as base learners and are joined by a two-layer stacking framework [8-10]. The aim is to see whether feature engineering and ensemble learning can give stable improvements under a limited sample size and provide interpretable decision support for possible clinical use [4-6,15,16].

2. Methods

2.1. Dataset and preprocessing

The paper uses the Heart Disease (Cleveland) dataset from the UCI Machine Learning Repository [7]. The dataset contains 920 records and 14 clinical variables. The main variables cover age and sex, chest pain type, resting blood pressure, blood cholesterol, fasting blood sugar, resting ECG result, maximum heart rate, whether angina appears during exercise, ST-segment depression (oldpeak), the slope of the ST segment, the count of major vessels on fluoroscopy and the type of thalassemia, and a diagnosis label (num) [7]. The original diagnosis takes values from 0 to 4. Here, value 0 is treated as “no heart disease”, and values 1–4 are combined into 1, so the task becomes binary classification [7].

In the preprocessing stage, missing values and outliers are checked. For a small number of missing continuous variables, the k-nearest neighbors (KNN) imputer ($k = 5$) is used so that as many samples as possible are kept [4-6]. Categorical features like sex, type of chest pain, resting ECG findings, the presence of exercise-induced angina, and the ST-segment slope and thalassemia type are converted into dummy (0/1) indicators using one-hot encoding. For continuous variables, extreme outliers are winsorised: values far below the typical range are raised to a lower bound defined from the first quartile and inter-quartile range, and values far above the typical range are reduced to an upper bound defined from the third quartile and inter-quartile range. After this step, all numeric features are standardized to have a mean of zero and unit variance using the StandardScaler [4-6,11].

Then the data are randomly divided into a training set and a test set, with about 70% for training and 30% for testing. Stratified sampling is used so that the share of positive and negative cases is similar in both sets [4-6]. The distribution of the target variable (0 = healthy, 1 = diseased) is shown in Figure 1. The labels are relatively balanced, which is suitable for binary classification [4-6].

DOI: 10.54254/2753-8818/2026.PJ30865

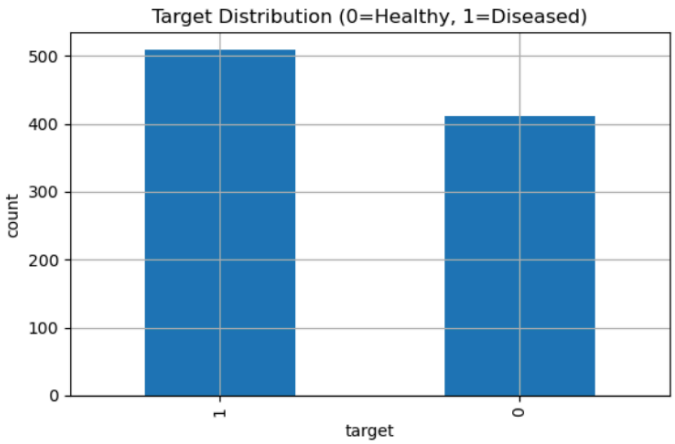

Figure 1. Distribution of the target variable (0 = healthy, 1 = diseased) (photo credit: original)

2.2. Feature engineering

Based on previous studies on feature engineering in medical prediction, this work designs five enhanced features and three new composite features on top of the original variables [11,12,14]. The aim is to capture non-linear effects and interactions that reflect clinical knowledge about cardiovascular risk [2,6,11,14].

For enhanced features, interaction terms and simple transformations are used. For example, the Age-Chol ratio is defined as cholesterol divided by (age + 1) to reflect the burden of blood lipids at different ages [11,12,14]. Rate-pressure product is defined as heart rate × blood pressure to describe the workload of the heart. ST-depression slope combines ST depression with the slope of the ST segment to show possible ischemia-related ECG changes. An interaction between chest pain type and exercise-induced angina is added to distinguish typical and atypical symptoms. Cholesterol is also transformed by a log function to reduce the effect of skewness [11,12,14].

For the new composite features, a log-risk score is built from blood pressure, cholesterol, and ST depression to measure the combined effect of several main risk factors. Chest pain type, maximum heart rate, and exercise-induced angina are then combined to construct the RBG Log, which reflects the risk of ischemia under exercise. A composite feature called RISC Combo (Risk-Integrated Score Combination) integrates chest pain type, maximum heart rate, and ST depression to describe multi-factor risk [14]. Finally, Rank High is designed to mark samples that show high-risk patterns in several features at the same time, such as high blood pressure, high cholesterol, and obvious ST depression [2,6,14].

Table 1 gives the definitions and medical meanings of all eight new features. These features are consistent with common clinical knowledge about hypertension, high cholesterol, and myocardial ischemia risk [2,6,11,12,14].

DOI: 10.54254/2753-8818/2026.PJ30865

Table 1. Definitions and medical meanings of the new features

Feature name	Calculation method	Medical meaning
Age-Chol ratio	$chol \div (age + 1)$	Reflects the burden of lipid load at different ages.
Rate-pressure product	$trestbps \times thalch$	Describes the workload of the heart under resting conditions.
St-depression slope	$oldpeak \times slope\ index$	The electrocardiogram interaction effect of ST segment depression and slope
Cp-exang interact	$cp \times (1 - exang)$	The interaction characteristics between chest pain types and exercise-induced states
Chol log	$\log(1 + chol)$	$\log(1 + chol)$
RBG Log	$\log(1 + trestbps \times chol \div (1 + oldpeak))$	Composite indicator of exercise-related ischemia based on symptoms and stress response.
RISC Combo	$cp \times (thalch \div (1 + age)) - exang \times oldpeak$	A combined risk score that uses chest pain type, maximum heart rate, whether symptoms appear during exercise and ST-segment depression
Rank High	Risk stratification based on multi-feature weighted scoring	Marks patients who show multiple high-risk patterns at the same time.

2.3. Models and experimental design

To balance interpretability and non-linear fitting ability, this study uses LG, RF, XGBoost, SVM and k-nearest neighbours KNN as base learners [4-6,8,9]. LG acts as a simple linear baseline and its coefficients are easy to explain. RF and XGBoost are tree-based ensemble models with strong non-linear modelling power and they often work well on tabular clinical data [4-6,8,9]. SVM can learn complex decision boundaries in a medium-sized dataset. KNN is a simple distance-based model that gives a different view of the data [4-6].

The ensemble uses a two-layer stacking structure [8-10]. In the first layer, the five base learners are trained on the training set. It uses stratified five-fold cross-validation and records the out-of-fold prediction probabilities. These probabilities, together with prediction scores on the test set, are used as new input features for the second layer. In the second layer, an XGBoost model is chosen as the meta-learner. It learns how to join the outputs of the base models in probability space and gives the final prediction [9,10].

DOI: 10.54254/2753-8818/2026.PJ30865

On the training set, it runs stratified five-fold cross-validation to test each model under two feature settings: using only the original "base" features and using the extended "enhanced" feature set [4-6]. For every algorithm, key hyperparameters are tuned by grid search within set ranges. I measure performance with AUC, accuracy, precision, recall and F1-score. It also draws ROC curves and precision–recall (PR) curves to show how sensitivity and the number of false alarms change under different thresholds [4-6,13].

3. Result

3.1. Overall model performance

With only the original features, the AUC of single models in five-fold cross-validation is around 0.87. After adding the engineered features, the average AUC of the five models increases slightly, and precision and F1-score also rise. This shows that the designed features bring small but generally positive gains for most algorithms [11,12,14].

On the independent test set, the stacking model achieves an AUC of 0.8771. Its accuracy is 0.8087 and the F1-score is 0.8285, which are very similar to the values of the strongest single model (enhanced SVM). At the same time, stacking has the highest AUC and recall among all models. Overall, it provides the best balance between sensitivity and overall discrimination (Table 2).

In Figure 2, the ROC curves of four key models on the test data are drawn. The curve for the stacking model is above the other curves for most threshold values. Figure 3 shows the corresponding PR curves. The stacking model has higher precision over a wide range of recall, especially in the high-recall region, which is important for screening tasks [4-6,13].

Table 2. Performance of different models on the base and enhanced feature sets

Setting	Model	ACC	PRE	PRC	F1	AUC
Base	KNN	0.8065	0.8188	0.8350	0.8264	0.8581
Enhanced	KNN	0.8033	0.8150	0.8330	0.8235	0.8604
Base	LogisticRegression	0.7870	0.8174	0.7918	0.8042	0.8677
Enhanced	LogisticRegression	0.8054	0.8286	0.8173	0.8227	0.8719
Base	RandomForest	0.7891	0.7999	0.8253	0.8121	0.8559
Enhanced	RandomForest	0.7946	0.8029	0.8331	0.8171	0.8580
Base	SVM	0.8076	0.8331	0.8153	0.8240	0.8727
Enhanced	SVM	0.8163	0.8425	0.8212	0.8314	0.8716
Base	Stacking	0.7946	0.8212	0.8035	0.8119	0.8719
Enhanced	Stacking	0.8087	0.8207	0.8369	0.8285	0.8771
Base	XGBoost	0.7707	0.7810	0.8134	0.7967	0.8347
Enhanced	XGBoost	0.7783	0.7933	0.8095	0.8012	0.8411

DOI: 10.54254/2753-8818/2026.PJ30865

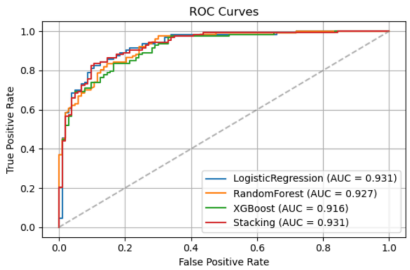

Figure 2. ROC curves of different models on the test set (photo credit: original)

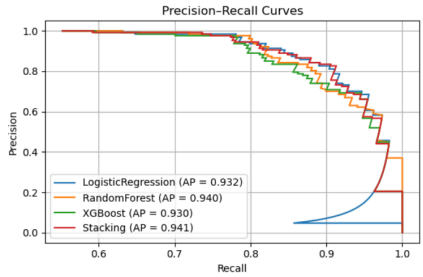

Figure 3. Precision–recall curves of different models on the test set (photo credit: original)

3.2. Feature importance and interpretability

To analyze the contribution of different features, feature importance scores are calculated on the enhanced feature set using the permutation importance method based on a RF model [8,11,12]. The results, shown in Figure 4, indicate that Chol log, Age-Chol ratio, ST depression (oldpeak), RBG Log and RISC Combo are the top important features. Rank High also has a high importance score and plays a key role in risk stratification [11,12,14].

These findings are consistent with clinical knowledge that the combination of high blood pressure, high cholesterol and ST segment changes is linked to higher cardiovascular risk [2,6,14]. The importance ranking suggests that the new engineered features carry useful information that is not fully captured by the original variables. It also shows that the model pays special attention to core clinical indicators when classifying high-risk individuals, which supports interpretability [15,16].

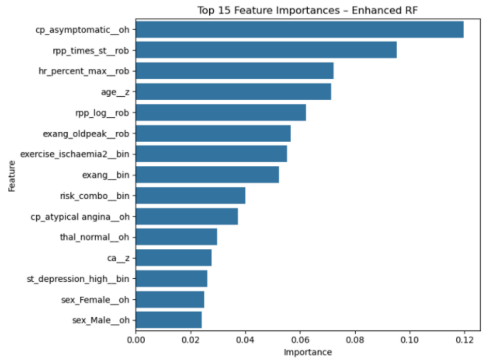

Figure 4. Top fifteen important features based on permutation importance in the RF model (photo credit: original)

DOI: 10.54254/2753-8818/2026.PJ30865

3.3. Confusion matrix and model stability

In Figure 5, the confusion matrix of the stacking model on the test data is given. The counts of true positives and true negatives are clearly higher than the counts of false positives and false negatives. This means that both missed high-risk cases and wrongly flagged low-risk cases are kept at a low level, which is important for practical screening [4-6].

In stratified five-fold cross-validation, the AUC values of the stacking model fall in a narrow range. The variance is smaller than that of several single models, which shows that the ensemble structure is stable across different splits of the data [8-10]. Figure 6 reports the average AUC of each model in cross-validation and again shows that the stacking model gives the strongest overall results [8-10].

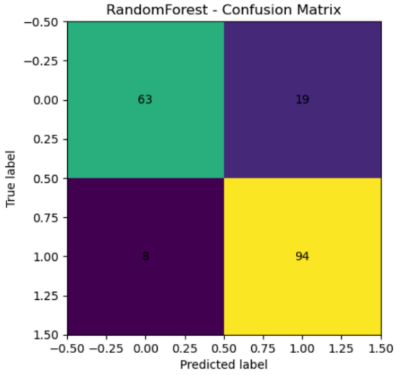

Figure 5. Confusion matrix of the stacking model on the test set (photo credit: original)

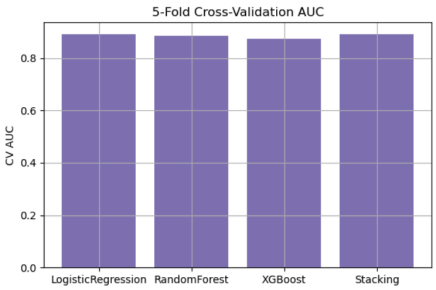

Figure 6. Average AUC of different models in five-fold cross-validation (photo credit: original)

3.4. Comparison between base and enhanced models

With the same model structure, the enhanced feature set brings small but consistent improvements in AUC, precision and F1-score for most models. In the ROC curves, the enhanced models move closer to the upper-left corner. In the PR curves, precision is slightly higher in the high-recall region. These patterns suggest that the gains are systematic rather than random fluctuations. These results indicate that carefully designed feature engineering can provide stable and meaningful improvements on top of standard machine learning models, even when the dataset is not very large.

4. Discussion

This study uses the UCI Heart Disease dataset and evaluates the role of “feature engineering + ensemble learning” in heart disease risk prediction [4-7]. The results show that, under a limited

DOI: 10.54254/2753-8818/2026.PJ30865

sample size, designing enhanced and composite features based on clinical mechanisms brings stable improvements across several models [2,6,11,12,14]. Simply using more complex models without feature engineering gives smaller gains. This is in line with previous work on feature engineering for medical and heart disease prediction [11,12,14].

Compared with single models, the stacking ensemble combines the outputs of different algorithms through a meta-learner [8-10]. In this work, the stacking model achieves higher AUC and better stability than all base learners while keeping a reasonable level of interpretability [4-6,8-10,15,16]. This agrees with recent studies that suggest combining several models and human knowledge is a promising direction for cardiovascular risk prediction and explainable artificial intelligence in medicine [15,16]. The feature importance analysis confirms that engineered features such as Age-Chol ratio, log-transformed cholesterol, composite risk scores, and Rank High have high contribution [11,12,14]. These features describe the joint effect of several key risk factors, such as blood pressure, cholesterol, ST depression, and exercise-induced angina [2,6,14]. The model therefore focuses on core clinical indicators when classifying high-risk individuals, which can help doctors understand and check the outputs [15,16]. This is consistent with the idea of interpretable machine learning for medical decision support [16].

There are still limitations. First, the data comes from a single public dataset with a limited sample size and a relatively simple population structure, so external generalization to other regions and multi-center populations has not been tested [7]. Second, the study uses only structured variables and does not include ECG signals, imaging data or genetic information, which may provide extra information for more precise risk prediction [4-6,15]. Third, the feature engineering process is mainly manual, and some useful combinations may have been missed. Future work could use automated feature engineering and causal feature selection methods and could test the model on larger real-world datasets [11,12,17].

5. Conclusion

In this study, a heart disease risk prediction method is developed that combines feature engineering with a stacking ensemble structure. Based on the UCI dataset, enhanced and composite features are designed on top of the original clinical variables to capture relations between blood pressure, cholesterol, ST depression, chest pain type, and exercise-induced angina. Five common classifiers, including LG, RF, XGBoost, SVM and KNN, are used as base learners, and XGBoost serves as the meta-learner in the stacking framework.

The experimental results show that, overall, feature engineering leads to small but consistent improvements in AUC, F1-score and accuracy across the model set. The stacking model achieves a test AUC of 0.8771 and keeps a good balance between sensitivity and specificity. Feature importance analysis indicates that several engineered features, such as Age-Chol ratio, log-transformed cholesterol and composite risk scores, play a key role in classification.

Overall, the study suggests that, under a limited amount of structured clinical data, reasonable feature engineering and ensemble learning strategies can improve the overall accuracy and stability of the heart disease risk prediction models. In the future, the method can be extended to multi-center real-world datasets and longitudinal follow-up data, and it may be embedded into clinical workflows to support early detection and management of high-risk cardiovascular patients.

DOI: 10.54254/2753-8818/2026.PJ30865

References

[1] World Health Organization. (2023). Cardiovascular diseases (CVDs) – fact sheet. World Health Organization. https: //www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
[2] Mensah, G. A., Roth, G. A., Fuster, V., et al. (2023). Global burden of cardiovascular diseases and risks, 1990–2021. Journal of the American College of Cardiology, 82(13), 1159–1205.
[3] D'Agostino, R. B., Sr., Vasan, R. S., Pencina, M. J., et al. (2008). General cardiovascular risk profile for use in primary care: The Framingham Heart Study. Circulation, 117(6), 743–753.
[4] Chicco, D., & Jurman, G. (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Medical Informatics and Decision Making, 20(1), 16.
[5] Quer, G., Arnaout, R., Henne, M., & Rao, P. (2021). Machine learning and the future of cardiovascular care: A state-of-the-art review. Journal of the American College of Cardiology, 77(3), 300–313.
[6] Krittanawong, C., Virk, H. U. H., Bangalore, S., et al. (2020). Machine learning prediction in cardiovascular diseases: A meta-analysis. Scientific Reports, 10(1), 16057.
[7] Dua, D., & Graff, C. (2019). Heart disease (Cleveland) data set. UCI Machine Learning Repository.
[8] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https: //doi.org/10.1023/A: 1010933404324
[9] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794).
[10] Wolpert, D. H. (1992). Stacked generalization. Neural Networks, 5(2), 241–259. https: //doi.org/10.1016/S0893-6080(05)80023-1
[11] Kuhn, M., & Johnson, K. (2020). Feature engineering and selection: A practical approach for predictive models. CRC Press.
[12] Roe, K. D., Jawa, V., Zhang, X., & Overby Taylor, C. (2020). Feature engineering with clinical expert knowledge: A case study assessment of machine learning model complexity and performance. PLOS ONE, 15(4), e0231300.
[13] Saito, T., & Rehmsmeier, M. (2015). The precision–recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), e0118432.
[14] Yadav, D., Rani, D., & Verma, O. P. (2025). Impact of feature selection and feature engineering in prediction of cardiovascular diseases. Computers in Biology and Medicine, 197(Pt B), 111027.
[15] Bilal, A., Alzahrani, A., Almohammadi, K., Saleem, M., Farooq, M. S., & Sarwar, R. (2025). Explainable AI-driven intelligent system for precision forecasting in cardiovascular disease. Frontiers in Medicine, 12, 1596335.
[16] Molnar, C. (2022). Interpretable machine learning: A guide for making black box models explainable (2nd ed.). Self-published. https: //christophm.github.io/interpretable-ml-book
[17] Chen, Y., Zhang, J., & Qin, X. (2022). Interpretable instance disease prediction based on causal feature selection and effect analysis. BMC Medical Informatics and Decision Making, 22(1), 51.