Prediction of Cardiovascular Diseases Based on Mainstream Machine Learning Algorithms

Peiyuan Liu

School of Economics and Management, Beijing Forestry University, Beijing, China
liupeiyuan@bjfu.edu.cn

Abstract. Cardiovascular diseases (CVDs) are one of the hottest issues in present medical research due to their status as the leading cause of mortality worldwide. Studies have achieved certain achievements in early detection tool development. However, there is still a research gap in accurate, efficient and widely applicable predictive models for CVDs or models that could integrate multiple clinical indicators. This study attempts to predict cardiovascular diseases based on mainstream machine learning algorithms. Firstly, this study collected a clinical dataset including 11 feature variables (such as Age, Sex, ChestPainType, RestingBP) and a target variable (HeartDisease). Secondly, this study finished the dataset's preprocessing and analysis. Finally, this study constructed and trained five mainstream machine learning models including Gradient Boosting, K-Nearest Neighbors (KNN), Logistic Regression, Random Forest and Support Vector Machine (SVM). The experimental results show that the KNN model got the highest accuracy of 0.8913, recall of 0.9118 and F1-score of 0.9027, while the Gradient Boosting model got AUC of 0.94 and it was ranked first in generalization ability. This study concludes that mainstream machine learning algorithms, especially KNN and Gradient Boosting, could improve the accuracy of CVD prediction and could be used as an accurate and reliable tool in clinical early screening of cardiovascular diseases.

Keywords: Cardiovascular Diseases, Machine Learning, Predictive Model, K-Nearest Neighbors, Gradient Boosting.

1 INTRODUCTION

1.1 Background

Cardiovascular diseases (CVDs) are one of the leading causes of mortality worldwide and they bring a huge challenge to public health [1-3]. Although studies have achieved certain achievements in early detection tool development, there are still research gaps in accurate, efficient and widely applicable predictive models development for CVDs or multiple clinical indicators integrating models development [4-6]. Traditional linear models could not capture the nonlinear relationship between CVD risk factors and target variables, such as age, cholesterol and electrocardiogram features [7, 8]. This study attempts to predict cardiovascular diseases based on mainstream machine

K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_54

learning algorithms and it is expected to provide reliable support in clinical early screening [9, 10].

1.2 CVD Prediction via Machine Learning

The research process can be outlined as follows: Firstly, this study describes and preprocesses cardiovascular disease dataset; Secondly, zero value for Cholesterol and normalization for other continuous variables are applied; Thirdly, Gradient Boosting, KNN, Logistic Regression and Support Vector Machine(SVM), Random Forest models are constructed and trained on the cardiovascular disease dataset; Finally, accuracy, precision, recall, F1-score and AUC(Area under the ROC curve) are used to evaluate the five models, and confusion matrix and ROC curve are used to plot the performance of the five models; At last, the advantages and disadvantages of these models are compared and discussed. The significance of the results for clinical CVD screening is also discussed.

The purposes of this study are: (1) To construct an integrated ML-based prediction framework for CVD; (2) To find the best mainstream ML model for CVD prediction; (3) To explore the key clinical features that affect CVD risk and provide useful references for clinical CVD prevention.

2 Dataset

2.1 Descriptive Introduction

The dataset used in this study is publicly available cardiovascular disease dataset (cardiovascularDiseases.csv) which includes 918 samples and 12 variables. The dataset is composed of 11 feature variables and 1 target variable HeartDisease. The description of the 11 feature variables is shown in Table 1.

Descriptive statistics of continuous variables (Age, RestingBP, Cholesterol, MaxHR, Oldpeak) are presented in Table 2. The range of ages of patients is from 28 to 77 years with the average of 53.5 years, which means that the data comes mostly from middle-aged and elderly people, who are more likely to suffer from CVDs. The average RestingBP is 132.5 mm Hg, which is more than the range of normal BP (120 mm Hg), showing that there are many hypertensive samples. From the Cholesterol variable, this study can see that its range is large (0-603 mg/dL), and there are 177 samples whose value is 0, which is considered as missing data since it is impossible that serum cholesterol is 0 in clinical practice. The average MaxHR is 136.8 bpm and the average Oldpeak is 1.02, which means that most patients have mild ST depression.

Table 1. Description of Variables in the Cardiovascular Disease Dataset.

Variable Name	Type	Description
Age	Continuous	Age of the patient (years)
Sex	Categorical	Gender of the patient (M: Male; F: Female)
ChestPainType	Categorical	Type of chest pain (ATA: Atypical Angina; NAP: Non-Anginal Pain; ASY: Asymptomatic; TA: Typical Angina)
RestingBP	Continuous	Resting blood pressure (mm Hg)
Cholesterol	Continuous	Serum cholesterol level (mg/dL)
FastingBS	Binary	Fasting blood sugar (0: ≤120 mg/dL; 1: >120 mg/dL)
RestingECG	Categorical	Resting electrocardiogram results (Normal: Normal; ST: ST-T wave abnormality; LVH: Left ventricular hypertrophy)
MaxHR	Continuous	Maximum heart rate achieved (bpm)
ExerciseAngina	Binary	Exercise-induced angina (Y: Yes; N: No)
Oldpeak	Continuous	ST depression induced by exercise relative to rest
ST_Slope	Categorical	The slope of the peak exercise ST segment (Up: Upsloping; Flat: Flat; Down: Downsloping)
HeartDisease	Binary	Target variable (0: No heart disease; 1: Has heart disease)

Table 2. Descriptive Statistics of Continuous Variables

Variable	Mean	Std. Deviation	Minimum	Maximum
Age	53.5	9.2	28	77
RestingBP	132.5	18.4	80	200
Cholesterol	244.6	51.8	0	603
MaxHR	136.8	22.3	63	202
Oldpeak	1.02	1.18	-2.6	6.2

2.2 Data Analysis and Processing

Data preprocessing is an important process to guarantee the reliability of the model training. In this study, the following preprocessing steps were performed.

This study handles missing values because the median is less sensitive to outliers than the mean, to avoid the imputed value distorting the results.

It is impossible to use this kind of variable directly as the input of machine learning models. Thus, this study uses one-hot encoding for the variable with multiple categories (ChestPainType, RestingECG, ST_Slope) and label encoding for the binary variable (Sex: M=1, F=0; ExerciseAngina: Y=1, N=0). For example, the ChestPainType

variable with the study's categories was converted into the study's binary variables (ChestPainType_ATA, ChestPainType_NAP, ChestPain Type_ASY, ChestPainType_TA), where each variable took a value of 1 if the original ChestPainType applied and 0 otherwise.

Features often have different units and ranges, so feature scaling is necessary. In some cases, the models' performance depends on the scale of the variables (e.g., KNN, SVM). Therefore, in the study, standardization (Z-score) was applied to normalize these variables. The standardization formula is as follows.

$$z=\frac{x-\mu}{\sigma} \tag{1}$$

X is the original value, μ is the mean of the variable, and σ is the standard deviation.

A Train-Test Split was performed into training and test sets using stratified sampling. Stratified sampling ensures that the proportion of samples with heart disease (=1) in the training and test sets is almost the same (HeartDisease=1 about 55%). Stratified sampling prevents bias caused by an uneven ratio of the number of samples in each class.

3 Models

The study chose five types of mainstream machine learning algorithms for CVD prediction, which have different advantages and disadvantages. The following are details of each model.

3.1 Gradient Boosting

Gradient Boosting is an ensemble learning method that creates a strong classifier from a sequence of weak decision tree learners, such that this study minimizes the logistic loss (for logistic regression) by adding trees that focus on the examples that were misclassified by the ensemble formed previously. XGBoost is used to implement the gradient boosting machine, with 5-fold cross validation to tune n_estimators=100, max_depth=3, eta=0.1, and subsample=0.8. These parameters help prevent overfitting while maintaining complexity. Additionally, early stopping (after 10 rounds with no improvement on the validation loss) results in more efficiency, resulting in the highest AUC (0.94) for generalization on the CVD prediction task.

3.2 K-Nearest Neighbors (KNN)

KNN is an instance-based "lazy learner" that classifies new samples for CVD by voting from its k nearest training neighbors. Since distance is measured in terms of Euclidean space, features must be standardized to prevent bias towards more variable features. This model was tuned using 5-fold cross-validation (k=5) to properly capture local patterns without overfitting. Since this model has not trained a fixed model, it makes use of the labeled training data directly during prediction. This simplicity, combined with its ability to capture non-linear relationships, allowed it to attain the highest accuracy (0.8913) and F1-score (0.9027) among the five models on the CVD screening prediction task.

3.3 Logistic Regression

Logistic Regression is another linear classification algorithm, which models the probability of belonging to a certain output (occurrence or non-occurrence of heart disease) using the logistic function. The algorithm estimates the coefficients of the feature variables in order to maximize the likelihood of the observed data. For the study's implementation of the Logistic Regression model, this study implemented it with L2 regularization (Ridge) to prevent overfitting. The strength of the L2 regularization was tuned to C=1.0. The model was implemented using the scikit-learn library. This study visualized the coefficients of the features.

3.4 Random Forest

Random Forest is an ensemble algorithm made up of multiple decision tree learners, trained on bootstrapped training datasets using random subsets of the features. The class mode of predictions made by the trees is used for the final classification. The randomness in both the data and features used helps to prevent overfitting. For the study's implementation, the model was tuned using 5-fold cross-validation with n_estimators=100, max_depth=5, and max_features="sqrt" (implemented via scikit-learn). Using the Gini impurity as the split criterion helps to provide a good balance of model complexity and generalization on the CVD prediction task with integrated clinical data. The attained accuracy was 0.8750 and the recall was 0.8922.

3.5 Support Vector Machine (SVM)

SVM is a maximum-margin classification method. It finds the hyperplane that provides maximum separation margin between the CVD cases and non-cases. It uses an RBF kernel to map features into a higher dimensional space to capture non-linear relationships. It models the relationship between input-output using the SMO optimizer. Tuned by 5-fold cross validation, SVM sets C=1.0 (regularization parameter) and gamma=0.1 (kernel coefficient) to maximize margin while minimizing errors. Using the scikit-learn package, SVM works in a high dimensional space (e.g., 11 clinical features in the study) and achieves the same accuracy as KNN (0.8913) and the highest recall (0.9510) for CVD cases.

4 Experiments

4.1 Experimental Configuration

All experiments were conducted on a computer with Intel Core i7-10700K CPU (3.8 GHz), 32 GB memory and NVIDIA GeForce RTX 3070 GPU, using Python 3.9 and the following libraries: pandas, numpy, scikit-learn, XGBoost, matplotlib and seaborn to complete data processing, model implementation and visualization, respectively. For the binary classification of cardiovascular diseases, key training parameters of the five models are configured for the sake of reproducibility. Gradient Boosting used binary cross entropy as the loss function, adopts a built-in tree-boosted optimizer, sets the learning rate to 0.1 (tuned via 5-fold cross-validation), sets the number of estimators to 100, and applies early stopping after 10 rounds of stagnant validation loss. KNN is a

lazy learner that stores standardized data, selects the number of neighbors as 5 (determined via 5-fold cross-validation), and uses proportionate Euclidean distance. Standardization guarantees an unbiased distance to another point, even if the two have different units. Logistic Regression uses binary cross entropy as the loss function, employs L-BFGS as the optimizer, incorporates L2 regularization, and sets C to 1.0 (tuned via 5-fold cross-validation). Adaptive steps ensure convergence. Random Forest adopts the mode of class for prediction, uses Gini impurity as the criterion, sets n_estimators to 100, max_depth to 5, and max_features to sqrt (all tuned via 5-fold cross-validation). It does not guarantee reaching the global optimum, but the use of max_depth and max_features makes it less likely to overfit. SVM with RBF kernel uses hinge as the loss function, employs SMO as the optimizer, sets the regularization parameter to 1.0 (tuned via 5-fold cross-validation), sets the kernel coefficient to gamma=0.1 (tuned via 5-fold cross-validation), and relies on adaptive steps to ensure convergence. Accuracy, Precision, Recall, F1-Score and AUC were used to evaluate models. The definitions were consistent to guarantee comparability.

4.2 Experimental Results and Analysis

Model Performance Metrics. The performance of the five models on the test set is presented in Table 3. From Table 3, it can be observed that the KNN model's accuracy is 0.8913 and F1-score is 0.9027, with recall of 0.9118, which is very excellent. In addition, the SVM model has the same accuracy as the KNN model, but recall of SVM is 0.9510 and F1-score is 0.9080. Gradient Boosting's accuracy and F1-score are ranked second, and their values are 0.8859 and 0.8965, respectively. The accuracy of Random Forest is 0.8750. As for Logistic Regression, although its accuracy is relatively low (0.8696), recall is 0.9314. KNN and SVM perform remarkably well, while Gradient Boosting and Random Forest also perform well, and Logistic Regression has advantages in recall.

Confusion Matrices. As for Class 0 (Actual 0), Gradient Boosting correctly predicts 72 cases and misclassifies 10 as Class 1; KNN correctly predicts 71 cases and misclassifies 11 as Class 1; Logistic Regression correctly predicts 65 cases and misclassifies 17 as Class 1; Random Forest correctly predicts 70 cases and misclassifies 12 as Class 1; SVM correctly predicts 67 cases and misclassifies 15 as Class 1.

Table 3. Performance Metrics of Five Machine Learning Models.

Model	Accuracy	Precision	Recall	F1-Score
Gradient Boosting	0.8859	0.9010	0.8922	0.8965
KNN	0.8913	0.8942	0.9118	0.9027
Logistic Regression	0.8696	0.8482	0.9314	0.8930
Random Forest	0.8750	0.8835	0.8922	0.8878
SVM	0.8913	0.8661	0.9510	0.9080

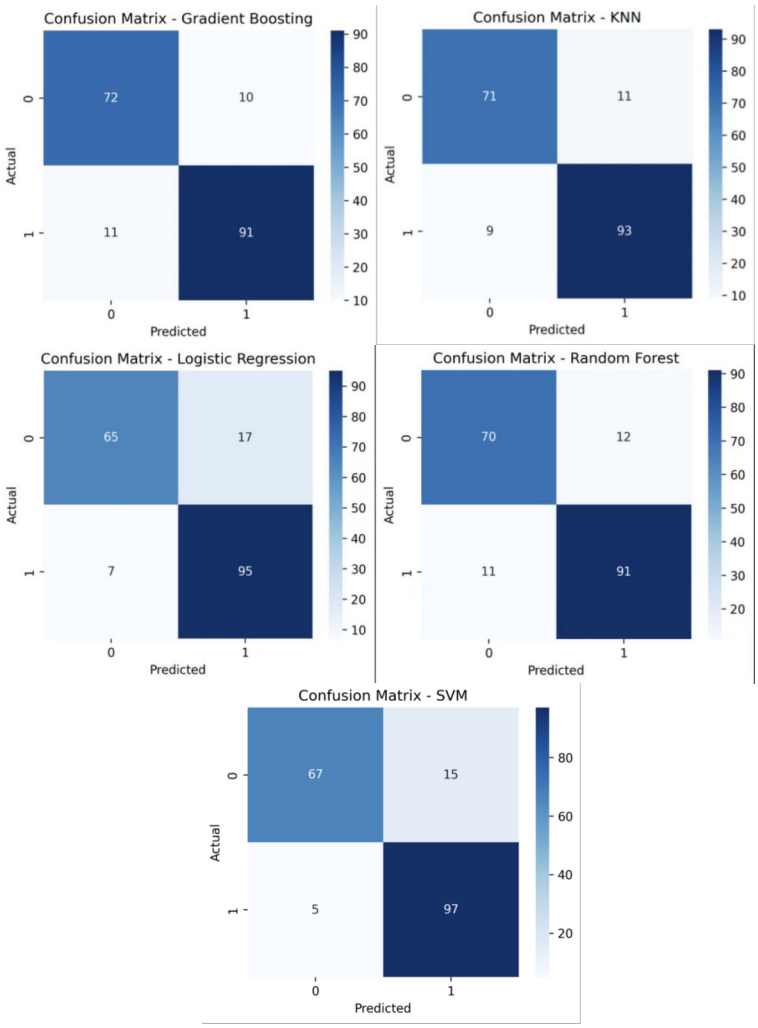

Fig. 1. Confusion Matrices. (Picture credit: Original)

Figure 1 shows the confusion matrices. As for Class 1 (Actual 1), Gradient Boosting correctly predicts 91 cases and misclassifies 11 as Class 0; KNN correctly predicts 93 cases and misclassifies 9 as Class 0; Logistic Regression correctly predicts 95 cases and misclassifies 7 as Class 0; Random Forest correctly predicts 91 cases and misclassifies 11 as Class 0; SVM correctly predicts 97 cases and misclassifies 5 as Class 0.

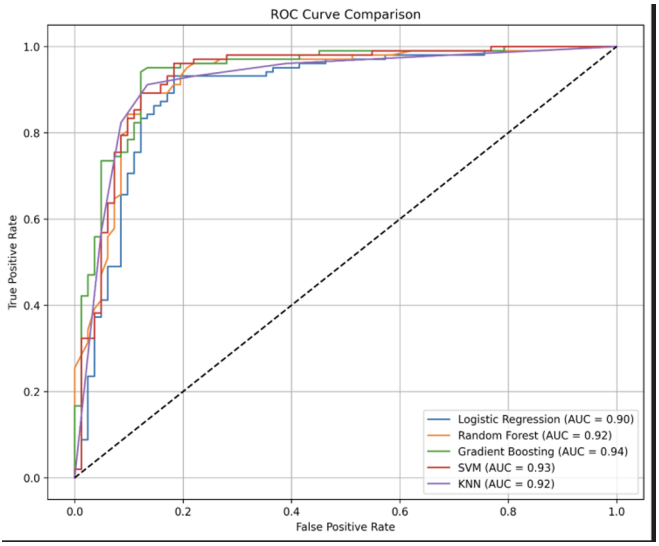

Fig. 2. Comparison of ROC Curves and AUC Values for Various Classification Models. (Picture credit: Original)

It can be seen that SVM performs very well in predicting Class 1, with 97 correct predictions, and Logistic Regression also performs very well in predicting Class 1, with 95 correct predictions. In addition, Gradient Boosting and KNN perform better than the other models in predicting Class 0.

Confusion matrices can give us a more intuitive comparison of how these models distinguish between Class 1 and Class 0, as well as the differences in true positives, true negatives, false positives, and false negatives across the models.

ROC Curve and AUC Analysis. As illustrated in Figure 2, this study draws ROC curves of five models and labels their AUC values. The ROC curve is the plot of true positive rate (recall) versus the false positive rate (FPR = FP / (FP + TN)) for a bunch of classification thresholds. From the ROC curve analysis, It can be observed that the AUC of Gradient Boosting achieves the highest 0.94 in all models, which means that Gradient Boosting has the strongest ability to distinguish the heart disease cases from non-heart disease cases; The AUC of SVM reaches 0.93 and it has good generalization ability; The AUC of KNN and Random Forest are both 0.92, which is slightly lower than Gradient Boosting and SVM, but it also means that they have excellent ability to distinguish the heart disease cases from non-heart disease cases; Logistic Regression

has an AUC of approximately 0.90, which means it has strong discriminative power in distinguishing between positive and negative cardiovascular disease cases.

Discussion of Results. Three main conclusions can be drawn from the experimental results. The KNN model is the best choice for clinical screening because it achieves the highest accuracy (0.8913) and recall (0.9118). In other words, it has the highest ability to identify heart disease cases correctly, while the overall prediction accuracy is also high. It is important to note that the high overall prediction accuracy is critical in clinical applications, because allowing cases of cardiovascular disease to be missed could result in cardiovascular disease patients not receiving timely treatments and an increase in mortality rate. The second reason why the KNN model performs well may lie in the standardized features, because it ensures that the distances are not dominated by different scales of different variables. The Gradient Boosting model has the best generalization ability; 1) The AUC of Gradient Boosting is 0.94, which means that it is the most robust model over all classification thresholds. Therefore, the Gradient Boosting model is recommended when the cost of false positives and false negatives may vary in different situations, i.e., the screening protocols for high-risk and low-risk populations. Linear models are not suitable for CVD prediction. The performance of Logistic Regression is not ideal, which may be caused by the nonlinear relationship between cardiovascular disease and different risk factors. For example, the risk of cardiovascular disease increases rapidly with age. Another example is the relationship between cholesterol and blood pressure; the relationship is not linear and the interaction is complicated. This study suggests that nonlinear machine learning algorithms are more suitable for building predictive models for cardiovascular diseases.

5 Conclusion

In this study, this study developed and evaluated five different machine learning models (Gradient Boosting, KNN, Logistic Regression, Random Forest, SVM) for CVD prediction using a clinical dataset with 918 samples and 12 variables. The experimental results showed that the KNN model achieved the highest accuracy (0.8913) and recall (0.9118) and was therefore the most suitable model for clinical CVD screening. The Gradient Boosting model got the highest AUC (0.94) and had the best generalization ability.

The importance of this study lies in two aspects. First, this study developed and compared five mainstream machine learning models for CVD prediction. Clinicians or studies can choose the most suitable tool according to their specific situation. Second, this study standardized the dataset, which improved the performance of different models. In addition, It can be observed that the raw dataset contained missing cholesterol values, which may affect the results.

This study still has some limitations that could be further explored in future studies. First, the dataset used in this study is limited (918 samples in total). In the future, studies will use larger multi-center datasets to train the study's models. Second, this study did not conduct any feature selection to find the most impactful features that affect the CVD

risk factors. In the future, this study will use methods like mutual information or SHAP values to understand what fraction of the impact of each feature contributes to the CVD prediction. Third, this study only used traditional machine learning algorithms. In the future, studies will also explore deep learning models (e.g., neural networks) to further improve the prediction accuracy.

References

1. Ford, E.S.: Risks for all-cause mortality, cardiovascular disease, and diabetes associated with the metabolic syndrome: a summary of the evidence. Diabetes Care 28(7), 1769–1778 (2005)
2. Kraus, W.E., et al.: Physical activity, all-cause and cardiovascular mortality, and cardiovascular disease. Medicine & Science in Sports & Exercise 51(6), 1270–1281 (2019)
3. Jagannathan, R., Patel, S.A., Ali, M.K., Narayan, K.M.V.: Global updates on cardiovascular disease mortality trends and attribution of traditional risk factors. Current Diabetes Reports 19(7), 44 (2019)
4. Altintas, Z., Fakanya, W.M., Tothill, I.E.: Cardiovascular disease detection using bio-sensing techniques. Talanta 128, 177–186 (2014)
5. Celermajer, D.S., Chow, C.K., Marijon, E., Anstey, N.M., Woo, K.S.: Cardiovascular disease in the developing world: prevalences, patterns, and the potential of early disease detection. Journal of the American College of Cardiology 60(14), 1207–1216 (2012)
6. Cohn, J.N., et al.: Screening for early detection of cardiovascular disease in asymptomatic individuals. American Heart Journal 146(4), 679–685 (2003)
7. Kaplan, G.A., Keil, J.E.: Socioeconomic factors and cardiovascular disease: a review of the literature. Circulation 88(4), 1973–1998 (1993)
8. Ueshima, H., et al.: Cardiovascular disease and risk factors in Asia: a selected review. Circulation 118(25), 2702–2709 (2008)
9. Damen, J.A., et al.: Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ 353, i2416 (2016)
10. Stoner, L., Lucero, A.A., Palmer, B.R., Jones, L.M., Young, J.M., Faulkner, J.: Inflammatory biomarkers for predicting cardiovascular disease. Clinical Biochemistry 46(15), 1353–1371 (2013)