ITM Web of Conferences **84**, 01010 (2026) https://doi.org/10.1051/itmconf/20268401010
ATCIDS 2026

Application Analysis of Machine Learning Models Integrating Multi-Source Clinical Data in the Early Prediction of Heart Disease

Jiajun Li[1*]

[1]Faculty of Science and Engineering, The University of Nottingham Ningbo China, Ningbo, China

Abstract. Since heart disease is one of the main causes of death globally, it makes sense to adopt cutting-edge techniques to increase the prediction accuracy of heart illness, which can lower mortality and save expenses. This research uses data preprocessing, data mining, machine learning, data visualization to generate and evaluate several prediction models, obtain that Random Forest (RF) has highest accuracy but lower ROC-AUC. Although SVM has lower accuracy than RF Model, it has the highest ROC-AUC. Therefore, SVM is the best model for medical scenarios. find the top 5 main factors influencing heart disease are ST_Slope_Up, Oldpeak, ST_Slope_Flat, ChestPainType_ASY and Cholesterol, using data visualization to show the result, which warning people to pay more attention about these physical indicators and seek medical attention promptly in case of mild chest pain or other symptoms. develop a better lifestyle. The research also reflects that there are still many practical challenges of the application of machine learning in data, model, and actual use. The future research could pay attention to improving data qualities, explore better models and gradually enhance practicality.

1 Introduction

Health is one of the most significant issues that affects everyone. Heart disease is closely related to health because the number of people who die from heart disease accounts for 31% of the total global deaths. In addition, with rapid development of socio-economic condition and changes in residents' lifestyles, the number of heart disease patients in our country is on the rise and they are gradually getting younger [1]. Moreover, most patients are missing diagnosed due to atypical early symptoms (such as mild chest pain and blood pressure fluctuations), thus missing the golden window period for intervention treatment and leading to an increase in mortality rate. Therefore, early detection of heart disease can minimize costs and mortality rates [2]. However, traditional diagnosis of heart disease mainly relies on methods such as symptom inquiry, physical examination, electrocardiogram, echocardiography and coronary angiography, due to the drawbacks of strong subjectivity, dependence on the accuracy of detection equipment and inability to provide early warning, the efficiency and accuracy of diagnosis are difficult to meet the demands of large-scale

* Corresponding author's email: ssyjl44@nottingham.edu.cn

screening. With the development of medical informatization and machine learning technologies, predictive models based on supervised learning have provided a new approach for assessing the risk of heart diseases. For instance, data mining technology can use classification algorithms to accomplish this task at low cost and high efficiency, which plays a crucial role in clinical research [2]. This can combine with machine learning to generate models，this can be applied in medical scenarios to increase prediction accuracy of heart disease [3]. This article will apply methods based on data mining and machine learning, train four types of prediction model which are Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), and Decision Tree (DT) then evaluate the best model for medical scenarios. In addition, it will also screen out the most critical factors affecting the incidence of heart disease. Finally, use data visualization techniques to show all the results.

2 Methods

2.1 Data Processing

The dataset used for training models is usually very large and most of data is real so inevitably there will be missing values and outliners which will affect the accuracy of model; therefore, it is necessary to deal with the problem through data preprocessing. For example, instance selection is a data preprocessing technique that seeks to exclude outliers and noisy data from a given dataset while also lowering storage costs and maintaining mining quality. However, this strategy would take a lot of time if the dataset is quite large [4]. Therefore, this article applies code to deal with this problem.

2.2 Model Training

2.2.1 Logistic Regression

The first model is LR, a type of multiple regression method used to examine the relationship between a binary or categorical outcome and multiple influencing factors. Multiple LR, conditional LR, polytomous LR, ordinal LR, and adjacent categorical LR are examples of regression techniques that are frequently used in medical research because they can measure associations, predict outcomes, and control for confounding variable effects [5]. This article applies binary logistic regression to train model. The fundamental presumptions of LR, such as the independence of errors, linearity in the logit for continuous variables, lack of multicollinearity, and absence of very significant outliers, must be taken into consideration while applying this method [6]. The core advantage is that as a linear model, it features fast computation speed and extremely strong interpretability (it can quantify the direction and intensity of the impact of features on disease probability). It has low requirements for data volume and is less prone to overfitting.

2.2.2 Random Forest

The second method is RF, which is an ensemble of DT. Each tree is developed using the repetitive partitioning concept, which starts with the root node and applies the same node splitting process repeatedly until specific stopping rules are satisfied [7]. The sum of weaker learners provides the prediction ability, and when each tree has fewer correlations, the performance will be better [7]. In this article, the performance of RF is relatively balanced, the core positioning is "an anti-overfitting nonlinear prediction tool", combining predictive performance with a certain degree of interpretability.

2.2.3 Support Vector Machine (SVM)

The next is the SVM, a supervised machine learning (ML) technique that can learn from data and make judgments. Vapnik and Chervonenkis first proposed the basic ideas of SVM in a theory in the 1960s [8]. However, following two notable advancements in the 1990s, it begins to garner more interest from the scientific world. First, the SVM can classify highly nonlinear situations thanks to a kernel method. The second allowed the SVM to be expanded to address issues in a regression framework known as support vector [8]. SVM is a typical model used in this task for handling small samples and high-dimensional features. Its core advantage lies in achieving classification through the "maximum margin hyperplane", which can effectively deal with scenarios in medical data where the feature dimension is high and the sample size is moderate (300+). At the same time, it has strong generalization ability and is less prone to overfitting.

2.2.4 Decision Tree

The last one is DT, DT is the basic unit of RF, it is an efficient and reliable decision-making method, which performs high accuracy of classification by presenting the collected information in a simple way. Moreover, it has ability to learn automatically, this makes it suitable for situations which need to make effective and reliable decisions that usually occur in medical settings. Therefore, it has been widely applied in various medical areas [9]. In this article, the core advantage of DT is its low training and usage costs, strong interpretability, and can be directly used for interpreting medical risk factors.

3 Result

3.1 Evaluation Dataset & Experimental Setup

The dataset of the research comes from Kaggle website and also codes run in Kaggle, the sample size is 918, use 12 feature dimensions including Age, Sex, ChestPainType, RestingBP, Cholesterol, FastingBS, RestingECG, MaxHR, ExerciseAngina, Oldpeak, ST_Slope and whether have Heart Disease or not. The training set and test set were randomly divided in an 8:2 ratio, and 5-fold cross-validation was used to reduce overfitting

3.2 Definition of Evaluation Metrics and Data Visualization

The research results are listed in table 1:

Table 1. Performance comparison of different models

Model Type	Accuracy(the higher the better)	ROC-AUC Score(core indicator in medical scenarios, the closer to 1 the better)	Precision	F1-score	Recall

LR	0.9130	0.9316	0.91(Healthy); 0.91(Diseased)	0.89(Healthy); 0.93(Diseased)	0.90(Healthy); 0.92(Diseased)
RF	0.9076	0.9340	0.92(Healthy); 0.90(Diseased)	0.87(Healthy); 0.94(Diseased)	0.89(Healthy); 0.92(Diseased)
SVM	0.8967	0.9456	0.88(Healthy); 0.91(Diseased)	0.88(Healthy); 0.91(Diseased)	0.89(Healthy); 0.90(Diseased)
DT	0.8152	0.8661	0.79(Healthy); 0.82(Diseased)	0.80(Healthy); 0.82(Diseased)	0.80(Healthy); 0.82(Diseased)

Accuracy: One of the main evaluations, the formula is "the number of correctly predicted samples / the total number of samples", which is used to intuitively reflect the overall classification accuracy of the model for positive and negative samples. It is suitable for scenarios where the sample distribution is relatively balanced. A bar chart is generated using data visualization (figure 1):

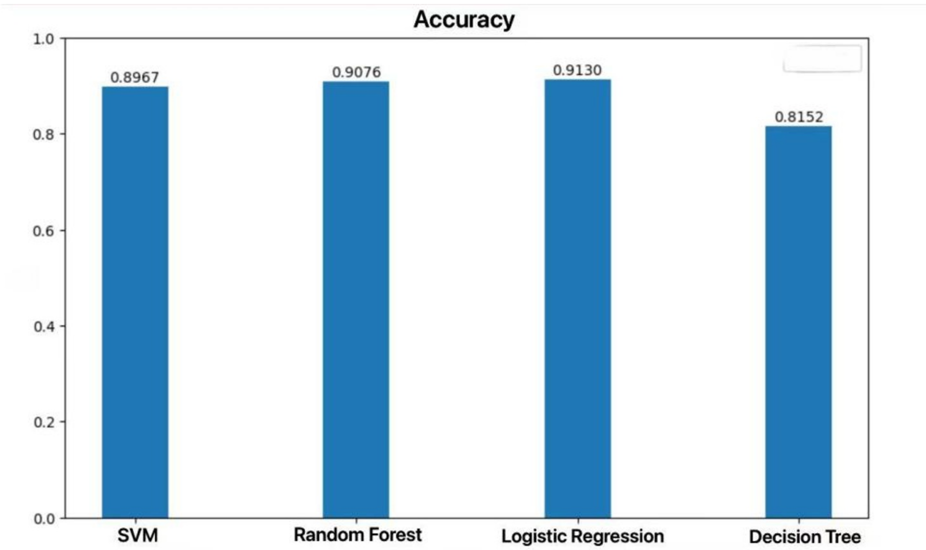

Fig. 1. Comparison of Accuracy of Different Models

Figure 1 shows that the LR has the highest accuracy which reaches 0.9130, all the models' accuracy is over 0.8.

ROC-AUC Score: Another main evaluation, the full name of ROC-AUC Score is Receiver Operating Characteristic - Area Under the Curve, Receiver Operating Characteristic (ROC) is widely used to reflect the performance of a model or algorithm when analyzing binary outcomes [10]. ROC analysis is a tool used to understand and explain models and text results, also is a goal of explainable artificial intelligence [11]. The area under ROC is called AUC, it has been used for medical diagnosis since 1970s [12]. The receiver operating characteristic curve (ROC Curve) and calculating the area under the curve (AUC) can measure he ability of the model to distinguish between positive and negative samples (The closer to 1, the better). Due to the typical class imbalance problem in medical scenarios where "the number of disease samples is usually less than that of normal samples", the ROC-AUC score can better reflect the actual generalization ability of the model than the accuracy rate. Therefore, it is set as the core evaluation indicator. A bar chart is generated using data visualization (figure 2):

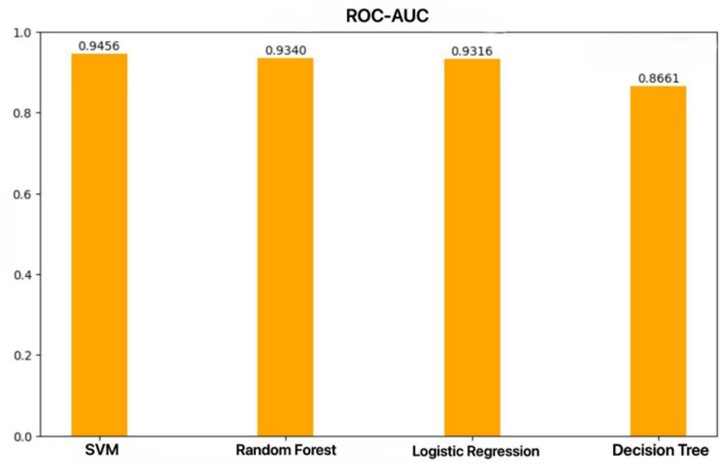

Fig. 2. Comparison of ROC-AUC of Different Models

Figure 2 shows SVM has the highest ROC-AUC scores which reach 0.9456, and all the models' ROC-AUC scores over 0.85.

3.3 Results and analysis

Then combine the figure 1, Figure 2 and do the data visualization (figure 3):

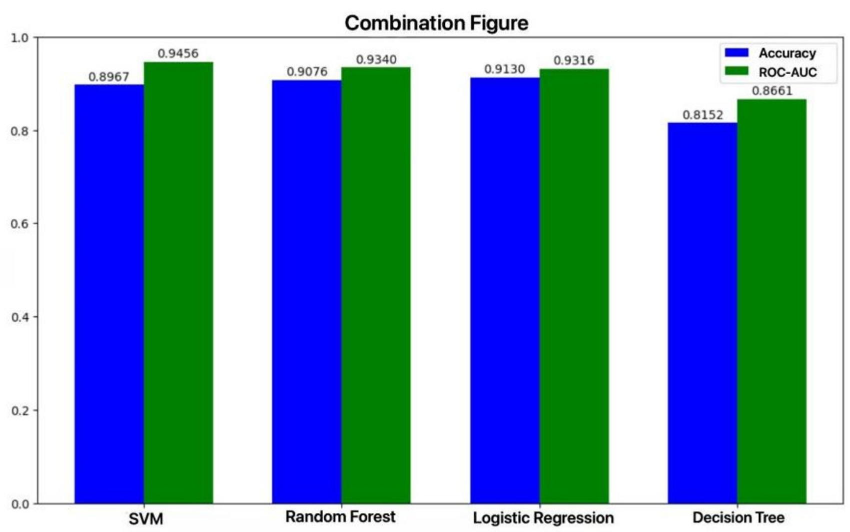

Fig. 3. Comparison of two mixed indicators.

Figure 3 directly shows the accuracy, ROC-AUC scores and comparison of each model. Although RF has the highest accuracy, higher than SVM, SVM has the highest ROC-AUC score, which is the core evaluation, and its accuracy is also reach 0.8967 that is relatively high and enough to use. In general, SVM performed the best and had higher practical value in the medical field.

The research also used SVM to further identify the five most critical factors influencing heart disease. The results are displayed as follows:

Table 2. Top 5 Influencing Factors Medical Interpretation

Influencing Factor	Medical Interpretation
ST_Slope_Up	ST segment slope upward: An upward ST segment on an electrocardiogram is a strong signal of myocardial ischemia, with the highest risk of disease.
Oldpeak	ST segment depression degree: Reflects the range of myocardial ischemia; the larger the value, the higher the risk.
ST_Slope_Flat	This feature is a categorical variable encoding and needs to be analyzed in combination with the original variable.
ChestPainType_ASY	This feature is a categorical variable encoding and needs to be analyzed in combination with the original variable.
Cholesterol	Cholesterol level: High cholesterol leads to atherosclerosis and is a core inducement of coronary heart disease.

Data visualization used to generate a bar chart that can make the result clearly:

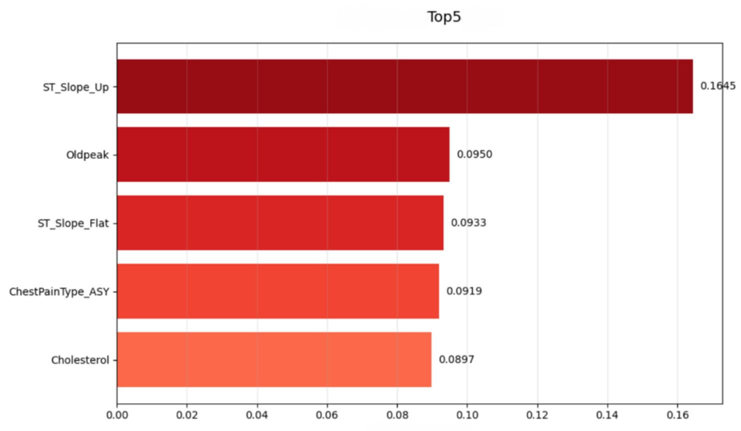

Fig. 4. Visualization of Top 5 Influencing Factors Medical Interpretation

Table 2 and Figure 4 shows that ST_Slope_Up, Oldpeak, ST_Slope_Flat, ChestPainType_ASY and Cholesterol are the top 5 factors influencing heart disease.

4 Conclusion

This research still reflects some practical challenges of the application of advanced methods like machine learning, data mining, and data visualization in medical scenarios. In data aspect, data may have issues such as distribution differences, missing values, and outliers, which could potentially affect the generalization ability of the model since the collection of clinical heart disease data involves multiple centers and multiple devices. In addition, the data may lack coverage for specific groups (such as rare heart disease types, specific age groups), resulting in poor performance in these groups. In model aspect, each model has their own advantages, conversely, they also has their own disadvantages, for instance, RF and other ensemble models have good predictive performance, but they have poor interpretability and are difficult to explain to medical professionals and patients about the key influencing factors of diseases; linear models such as LR have strong interpretability, but they are insufficient in handling complex nonlinear relationships, making it difficult to balance the two. Through data preprocessing, model training and evaluation, the performance of LR, RF, SVM, and DT were compared. The optimal model was selected based on ROC - AUC, and the feature importance of the optimal model was analyzed and interpreted in the context of medicine. At the same time, the top 5 influencing factors medical interpretation were identified. The overall process was standardized, and the training, evaluation of the model, and explanation of key factors were achieved. However, there are still practical challenges in terms of data and models. The future research can start from these aspects: Integrate multi-center and multi-source medical data to build a larger-scale and more representative dataset, covering more special populations and disease stages. At the same time, utilize advanced data cleaning techniques to improve data quality; explore other models which can integrate the advantages of existing models, Moreover, it is also necessary to promote the clinical validation and standardization of the model, and collaborate with medical institutions to conduct multi-center clinical trials to verify the effectiveness and safety of the model in real scenarios, and gradually incorporate it into the clinical auxiliary diagnosis process.

References

1. J. Y. Zhong, T. Y. Zhang, B. X. Chi, et al. Construction of Heart Disease Prediction Model and Visual Decision Support Research Based on Multimodal Data Fusion. Computer Campus, (19):297-299. (2025)
2. M. M. Ali, B. K. Paul, K. Ahmed, et al. Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. Computers in biology and medicine, 136, 104672. (2021)
3. D. Shah, S. Patel. & S. K. Bharti. Heart Disease Prediction using Machine Learning Techniques. SN COMPUT. SCI. 1, 345 (2020)
4. Y. H. Hu, W. C. Lin, C. F. Tsai, S. W. Ke, & C. W. Chen. An efficient data preprocessing approach for large scale medical data mining. Technology and health care: official journal of the European Society for Engineering and Medicine, 23(2), 153-160. (2015)
5. Q. Q. Wang, S. C. Yu, X. Qi, et al. Zhonghua yu fang yi xue za zhi [Chinese journal of preventive medicine], 53(9), 955-960. (2019)
6. J. C. Stoltzfus. Logistic regression: a brief primer. Academic emergency medicine : official journal of the Society for Academic Emergency Medicine, 18(10), 1099-1104. (2011)
7. J. Hu, & S. Szymczak. A review on longitudinal data analysis with random forest. Briefings in bioinformatics, 24(2), bbad002. (2023)
8. D. Valkenborg, A. J. Rousseau, M. Geubbelmans, & T. Burzykowski. Support vector machines. American journal of orthodontics and dentofacial orthopedics : official publication of the American Association of Orthodontists, its constituent societies, and the American Board of Orthodontics, 164(5), 754-757. (2023)
9. V. Podgorelec, P. Kokol, B. Stiglic, & I. Rozman. Decision trees: an overview and their use in medicine. Journal of medical systems, 26(5), 445-463. (2002)
10. J. Muschelli. ROC and AUC with a binary predictor: a potentially misleading metric. Journal of classification, 37(3), 696-708. (2020)
11. A. M. Carrington, D. G. Manuel, P. W. Fieguth, et al. Deep ROC analysis and AUC as balanced average accuracy, for improved classifier selection, audit and explanation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(1), 329-341. (2022)
12. J. Huang, & C. X. Ling. Using AUC and accuracy in evaluating learning algorithms. IEEE Transactions on knowledge and Data Engineering, 17(3), 299-310. (2005)