Research on Predictive Models for Heart Disease Based on Machine Learning

Zhengkang Li

Duke Kunshan University, No. 8 Duke Avenue, Kunshan, Jiangsu Province, 215316, China
zl458@duke.edu

Abstract. The heart is the core of the human circulatory system, providing a continuous supply of oxygenated blood to different organs in the body. However, when the heart can no longer effectively pump blood to the other organs, heart disease may develop. Therefore, the effective treatment and early diagnosis of heart disease have become the most significant issue in the medical field. Fortunately, the development of artificial intelligence and big data give birth to machine learning, and some statistical models in machine learning are applied to disease treatment and research. By applying machine learning to heart disease prediction, it can get a more accurate prediction with less time, thereby enhancing the reliability of predictive results. This thesis focuses on comparing diagnostic classifiers using five models in machine learning: logistic regression, k-nearest neighbor classification (KNN), decision tree, support vector machine (SVM), and random forest. The performance of the five models is then analyzed using various indicators, including accuracy, precision, recall, and F1 score. By applying different algorithms in machine learning to heart disease prediction, this research aims to achieve a more precise prediction and timely intervention before the outburst of heart disease, helping to reduce the incidence of heart disease.

Keywords: Heart disease prediction; Logistic regression, KNN, SVM, Random forests

1 Introduction

Heart disease, a common cardiovascular condition, is the leading killer around the world, according to the World Health Organization (WHO)'s December 2020 report on the top 10 global death causes. Nowadays, with the prosperity of material life, a variety of unhealthy habits such as staying up late have become increasingly prevalent, which significantly contributes to the rising rate of heart disease. Moreover, heart disease often lacks obvious warning signs, and its sudden onset may lead to severe consequences. Additionally, many heart diseases share similar symptoms before an outburst, which increases the difficulty of diagnosing and predicting accurately, and this issue also leads to an increase in the incidence of misdiagnosis and the rate of heart disease. Therefore, minimizing the predictive biases and achieving accurate heart disease prevention is of great importance.

A. J. Moshayedi (ed.), *Proceedings of the 2025 International Conference on Hybrid Commerce, Human Capital, and Economic Dynamics (ICHCH 2025)*, Advances in Economics, Business and Management Research 374,
https://doi.org/10.2991/978-2-38476-585-0_19

In heart disease prediction model research, in 2019, Dimopoulos et al. compared traditional cardiovascular disease scoring systems with machine learning algorithms and k-Nearest Neighbor (KNN) using established risk assessment tools [1]. The results indicated that these machine learning algorithms performed exceptionally well and can serve as an effective methodological approach for risk prediction research. In 2019, Gokulnath et al. proposed an intelligent Support Vector Machine (SVM) framework for early and accurate diagnosis of heart disease, and this GA simultaneously tuned the SVM hyperparameters and performed wrapper-based feature selection, improving both classification performance and model interpretability [2]. Using the Cleveland database, Receiver Operating Characteristic (ROC) analysis highlighted its potential in enhancing diagnostic accuracy through optimized feature selection. In 2021, Valarmathi et al. used three hyperparameter optimization algorithms to tune and test random forest and extreme gradient boosting algorithms [3]. They found that random forest with random search-based hyperparameter tuning yielded better results for heart disease prediction.

This thesis develops and compares diagnostic classifiers using five different models in machine learning, and then analyzes the corresponding results with various metrics, finally choosing one algorithm with the best performance, which improves the accuracy of heart disease diagnosis and reduces the mortality of heart disease at its origin.

2 Method

2.1 Dataset Resource and Instance Description

The UCI Machine Learning Repository is commonly used when people use machine learning, which is published by the University of California, Irvine. This thesis focuses on the clinical records of heart failure in the dataset and builds different models. This dataset has 12 features and 299 instances [4]. The instance's description is in Table 1.

Table 1. Variables description table

Variable name	Variable type	Remark
age	numerical	1=yes; 0=no
anaemia	categorical	1=yes; 0=no
Creatinine phosphokinase	numerical	
diabetes	categorical	1=yes; 0=no
Ejection fraction	numerical	
High blood pressure	categorical	1=yes; 0=no
platelets	numerical	
Serum creatinine	numerical	
Serum sodium	numerical	
sex	categorical	1=male; 0=female
smoking	categorical	1=yes; 0=no
time	numerical	
DEATH_EVENT	categorical	1=yes; 0=no

2.2 Model Building

This project uses Logistic regression, KNN, SVM, Random forests and Decision trees to build and describe the model.

Logistic Regression. Logistic regression, a classification algorithm, describes the probability of a binary result. Coefficients indicate the influence of a feature on the outcome. Despite its simplicity and interpretability, it shows a linear relationship between features and limits performance on non-linear data.

Decision Tree. A decision tree, a supervised learning algorithm, mainly makes decisions based on the value of different features by building a model that looks like a tree. The hierarchical tree starts from a root node, and then the tree splits data into several branches at internal nodes, eventually reaching the leaf nodes, which represent the results [5]. However, it may overfit if grown too deeply, which requires pruning or setting the depth limits.

Random Forest. Random forest is an ensemble method that combines several decision trees to improve the accuracy of prediction and mitigate overfitting [6]. In the random forests, each tree is based on a random set of features and data. And the predictions are made by aggregating results from all trees via majority averaging or voting. The diversity of this algorithm improves the generalization and reduces the variance.

SVM (Support Vector Machine). SVM is a dualistic classification algorithm, and the decision boundary is determined by support vectors. Not only can SVMs achieve robustness to outliers by minimizing the margin, but they can handle the nonlinear boundaries using kernel tricks to project data into higher-dimensional spaces.

KNN (K-Nearest Neighbors). KNN is an instance-based algorithm that classifies data by finding the k training samples closest to a new instance using distance metrics like Manhattan or Euclidean. KNN requires no training phase, making it adaptable to dynamic data. However, it is sensitive to irrelevant features, the choice of k, and computational complexity.

3 Results and Discussion

3.1 Result

To understand how well a model fits the data, it needs to use relevant metrics to evaluate its performance. These metrics can compare different models and choose the algorithm with the best performance, which is the significance of model performance evaluation.

Five models are constructed using five different algorithms above. Evaluation metrics for these models were calculated, including accuracy, precision, recall, and F1

score. ROC curves were plotted, and AUC values were computed. The evaluation results are in Table 2.

Table 2. Models result comparison

Model	Accuracy	Precision	Recall	F1 Score	AUC
Logistic Regression	0.80	0.93	0.56	0.70	0.82
SVM	0.75	0.86	0.48	0.62	0.82
Decision Tree	0.63	0.58	0.44	0.50	0.61
Random Forest	0.75	0.86	0.48	0.62	0.83
KNN	0.68	0.88	0.28	0.42	0.74

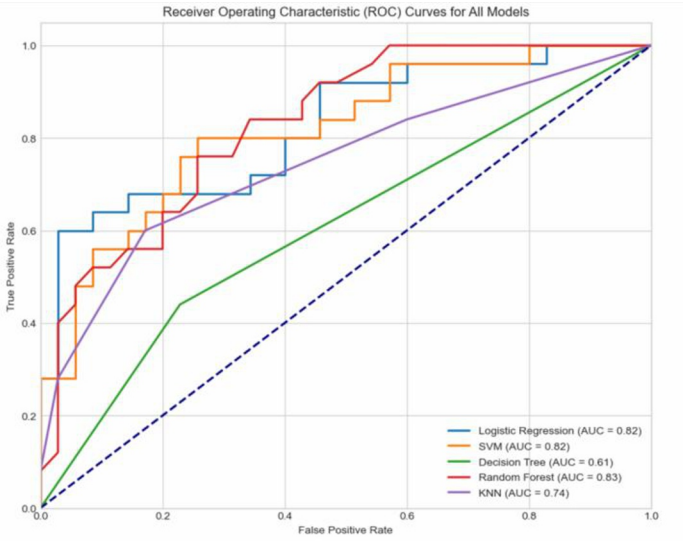

Fig. 1. ROC curve (Photo/Picture credit: Original).

Fig. 1 shows the ROC curve for all models, and this comprehensive comparison of the indicator results shows that for this heart disease dataset, the model fitted and predicted by the random forest algorithm has the best performance. This objectively verifies that the Bagging method of random forest demonstrates excellent performance and has significant advantages over other algorithms.

3.2 Discussion

Compared to other studies that mainly focus on single algorithms, this study uniquely evaluates five different models simultaneously, showcasing the Random Forest's superiority. However, the limitations of this study are also of great significance. Firstly, the instances of this dataset are quite limited (299 instances). The limited scale of the dataset may result in some errors in the study, making the result inappropriate for larger public heart disease studies. With a larger and more diverse dataset, the models could

potentially develop more nuanced relationships between different variables. Secondly, the five models constructed in the thesis are simple and basic. The shallow architecture of machine learning may be insufficient to capture the complex nonlinear feature interactions in cardiovascular disease prediction. For instance, some deep learning models like convolutional neural networks or other sophisticated ensemble methods may offer better performance. Additionally, the metrics used in the evaluation may not fully capture the whole model's performance in the specific clinical context. For example, some aspects like the cost and effectiveness of false positives and false negatives in the medical diagnosis context are not accounted for in the simple accuracy metric used here [7, 8].

Future work should incorporate more advanced machine learning methods like deep learning architectures, which could capture some temporal dependencies missed by traditional machine learning. Kwon et al. demonstrated that a hybrid CNN-RNN model outperformed RF by 9.8% in arrhythmia detection when it was trained on 50,000 ECG records [9]. Moreover, cost-sensitive learning should be explored to align model objectives with clinical priorities, as misdiagnosing heart disease incurs substantial societal costs [10].

4 Conclusion

This paper mainly studies a heart disease prediction model using real heart disease detection data from the UCI dataset. First, the dataset is processed to lay the foundation for the following models. Second, several visual analyses are conducted to find the internal links between different variables. Third, the heart disease data is predicted and classified using five different algorithms. Then the model's performance is evaluated by several metrics, and finally choose Random Forest is chosen as the best model for the prediction. The future research can focus on including more advanced and specialized machine learning methods and finding more evaluation metrics, which could further improve the precision of the research and reduce errors, and this is of great significance to heart disease prediction, for the patients, hospitals, and society.

References

1. Dimopoulos, A.C., Nikolaidou, M., Caballero, F.F., Chrysohoou, C., Pitsavos, C.: Machine learning methodologies versus cardiovascular risk scores, in predicting disease risk. BMC Medical Research Methodology 18(1), 179–180 (2018)
2. Gokulnath, C.B., Shantharajah, S.P.: An optimized feature selection based on genetic approach and support vector machine for heart disease. Cluster Computing 22, 14777–14787 (2019)
3. Valarmathi, R., Sheela, T.: Heart disease prediction using hyper parameter optimization (HPO) tuning. Biomedical Signal Processing and Control 70, 103033 (2021)
4. Heart failure clinical records. UCI Machine Learning Repository. https://doi.org/10.24432/C5GP7T (2020)
5. IBM: Decision trees. IBM Think China. https://www.ibm.com/cn-zh/think/topics/decision-trees, last accessed 2025/07/31

6. Pinheiro, D., Uchôa, A., Bezerra, C., Rodrigues, E., Pires, R., Valente, M.T., Rocha, L., Garcia, A.: Towards an effective refactoring triviality index: A machine learning approach from a developer's perspective. SSRN (2025)
7. Encord: Accuracy vs. precision vs. recall in machine learning: What is the difference? https://encord.com/blog/classification-metrics-accuracy-precision-recall/, last accessed 2025/07/31
8. Zielinski, J.: F1 score in machine learning: Intro & calculation. V7 Labs Blog. https://www.v7labs.com/blog/f1-score-guide, last accessed 2025/07/31
9. Kwon, J.M., Jeon, K.H., Kim, H.M., Kim, S.H., Park, J., Kim, Y.G., …, Oh, B.H.: Deep learning for ECG arrhythmia detection: A multicenter validation. Journal of the American College of Cardiology 77(18), 2384–2395 (2021)
10. Johnson, A.E., Aboab, J., Raffa, J.D., Pollard, T.J., Mark, R.G., Badawi, O.: A machine learning approach to severity assessment in sepsis. JMIR Medical Informatics 4(4), e100 (2016)