Dean&Francis
ISSN 2959-6157

Predicting Heart Disease Risk Using Machine Learning

Linkun Qian

Department of Mathematics, King's College London, London, the UK

*Corresponding author: linkun.qian@kcl.ac.uk

Abstract:

Heart disease is considered the world leader in mortality rates among people. A project was created with the goal of deploying a machine learning model for the prediction of cardiac disease using publicly available datasets from the University of California, Irvine Machine Learning Repository. Early detection through prompt care may improve death rates. In this respect, several machine learning algorithms: decision trees, logistic regression, random forests, and XGBoost, were used to identify heart disease patterns and risk factors. These above-mentioned models will be evaluated against the following key performance metrics: precision, accuracy, recall, and F1-score. Of all the algorithms, the XGBoost model performed the best, giving a precision of 89% and an F1-score of 0.87, which was one of the best in predicting heart diseases. These findings emphasize the crucial role of machine learning in further improving the prediction of cardiovascular diseases, possibly allowing for early diagnosis. Such predictive tools will allow healthcare providers to move toward more personalized and preventive treatments in patient care and outcomes.

Keywords: Heart disease prediction; Machine learning; Logistic regression; Random forest; XGBoost

1. Introduction

Cardiovascular disease is among the leading causes of death across the world and accounts for about 17.9 million deaths every year, according to the data set from World Health Organisation [1]. The main targets of the disease are the heart and blood vessels, leading to heart failure, coronary artery disease, and arrhythmias, among others. Consequently, heart conditions have imposed a high burden on healthcare systems the world over, an issue that calls for effective early diagnosis and prevention methods [2].

Early detection of cardiovascular disease can prevent a significant portion of mortality and morbidity by offering appropriate and timely interventions through individualized treatment plans. Traditional assessment methods for cardiovascular disease risk usually rely on statistical models and clinical scoring systems, such as the Framingham Risk Score, which considers age, cholesterol, blood pressure, and smoking status. These approaches, however, do not completely reveal the multi-factorial, nonlinear nature of the association between different risk factors and the

ISSN 2959-6157

susceptibility to heart disease [3].

With technological advancement and an increase in large-scale health data, the development of predictive models in healthcare has considerably been complemented, making machine learning a strong tool. Large amounts of data can be analyzed by machine-learning algorithms to find patterns that human specialists might not notice.The ability constitutes, therefore, a particular value of machine learning in developing predictive models able to precisely assess the risk of heart disease [4].

The objective of the study is, therefore, to come up with a well-considered machine learning model that could use patient data in predicting the risk due to heart ailment.

2. Literature review

Recent studies have emphasized the effective use of machine learning (ML) in heart disease prediction with various models showing higher accuracy than traditional statistical methods. Patel et al. (2023) explored algorithms such as Random Forests and Multiplayer Perception (MLP) and achieved an accuracy of 87.28% with hyper-parameter tuning and cross-validation [5]. This demonstrates the importance of algorithm selection and model optimization for obtaining better predictions .

In a long-term study, Mehrabani-Zeinabad et al. (2023) compared ML with traditional methods, emphasizing the superiority of integrated models such as stacking over traditional methods for cardiovascular risk prediction [6]. This is further supported by Arora et al. (2019), who demonstrated that Random Forest and XGBoost can be effective in reducing over-fitting and improving accuracy when combined with appropriate feature engineering.

Another key advancement is the interpretability of the models; the PLOS ONE (2021) study used Shapley Additive exPlanations (SHAP) to emphasize the importance of individual risk factors and their impact on disease prediction, helping clinicians to better understand the model outputs [7].

Furthermore, Khanna et al. (2019) emphasized the need to use feature selection techniques recursive feature elimination (RFE) to reduce computational complexity while maintaining predictive accuracy [8].

3. Methodology

Fig. 1 below illustrates the complete methodology workflow, followed by a detailed explanation and supplementary information.

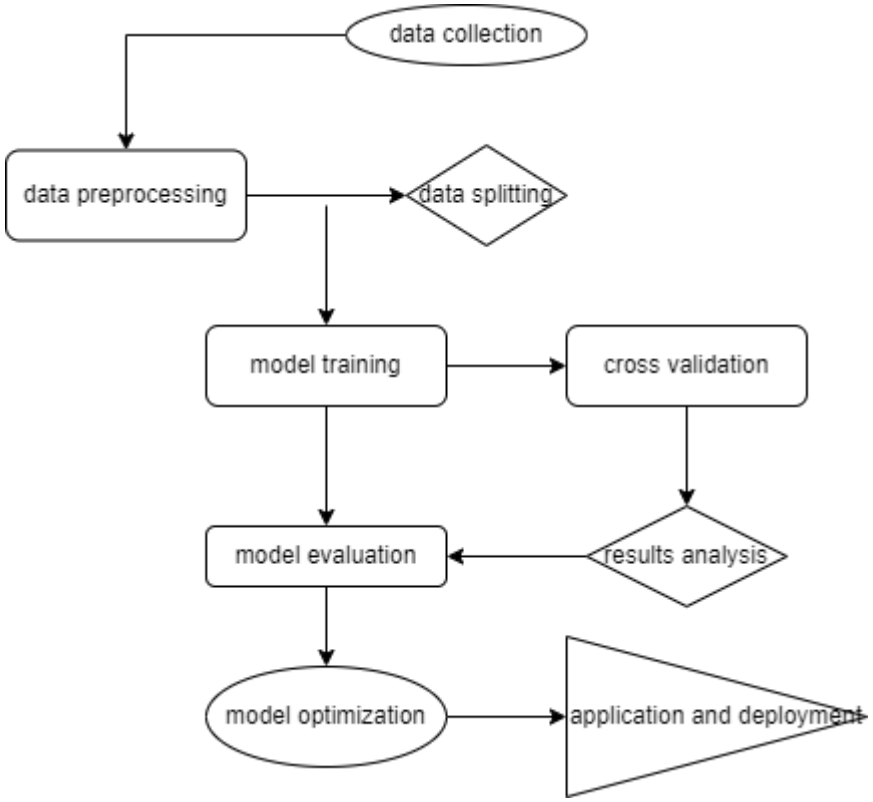

Fig. 1 Flowchart of Methodology

3.1 Data preprocessing

A crucial step in guaranteeing the caliber and dependability of machine learning models is data prepossessing [9]. Several key steps are performed in this process. First, missing values in the data set are handled by removing rows with incomplete data or interpolating missing values using appropriate statistical methods (e.g. median interpolation for numerical variables). The standard scale was used to scale the features evenly in feature space, including blood pressure, cholesterol, and age. For the model, the final data set was divided into 20% testing and 80% training [10]. To further enhance the generalization of the model and avoid over fitting, this study applied 5-fold cross validation. A data set is divided into five equal-sized subsets using K-fold cross-validation; one subset is utilized for validation and the other four are used for training in each iteration. This is repeated five times, and the average performance of those iterations gives a more realistic estimate of the model's ability to generalize new data [11].

3.2 Model Selection

Model selection is a very important building process in any effective machine learning solution in predicting cardiovascular diseases, and there are several considered.

Logistic regression is a simple, transparent linear model that normally serves for binary classification. It offers a linear feature combination that represents the probability of a binary result given certain inputs. In that regard, it is also a rather easy option to utilize for classification problems in which there is a linear relationship between the characteristics and the target variable.

Decision trees are one of the nonlinear tree-based algorithms used to make decisions by recursively partitioning the feature space.It provides an intuitive method of decision-making by building a tree-like model in which each internal node represents a feature, each branch represents a decision rule, and each leaf node represents a result.

The random forest is a learning methodology that integrates improvement on decision trees. It builds a number of trees and then combines the predictions over those trees made during training. This may reduce over fitting and make the model robust due to the collection of trees wherein biases or inconsistencies coming from singular trees are eliminated.

XGBoost, in full, is Extreme Gradient Boosting. The gradient boosting framework serves as the foundation for this sophisticated machine learning technique. Several advanced techniques are availed: parallel processing, regularization, and optimal tree pruning. These enhancements have dual purposes: in order to decrease overfitting and improve forecast accuracy.

3.3 Evaluation Metrics

In the following formula, TP means True Positives, FP means False Positives, TN means True Negatives, FN means False Negatives. This paper applies four evaluation methods: accuracy, precision, recall, and F1-score. The following is an introduction to these methods.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

4. Result

4.1 Data set

The UCI Heart Disease data set was selected [12]. It contains 303 patient records, each with 13 attributes representing potential risk factors for heart disease. These include age, gender, cholesterol level, maximum heart rate, and other medical indicators.

The target variable is a binary outcome which shows the presence of heart disease.

4.2 Results of 4 ML models

The results of the machine learning models are summarized in Table 1:

Table 1. Output of Four Models

Models	Accuacy	Precision	Recall	F1 score
Logistic Regression	81%	83%	79%	0.81
Decision Tree	85%	80%	75%	0.77
Random Forest	85%	89%	82%	0.85
XGBoost	88%	89%	85%	0.87

ISSN 2959-6157

Among the models tested, the XGBoost classifier performed the best, achieving an accuracy of 88% and the highest precision of 89% shown in Table 1. This indicates that XGBoost is highly effective at distinguishing if patients with or without heart disease. The recall value and F1-score for XGBoost were higher than those of the other models, making it the most suitable for heart disease risk prediction.

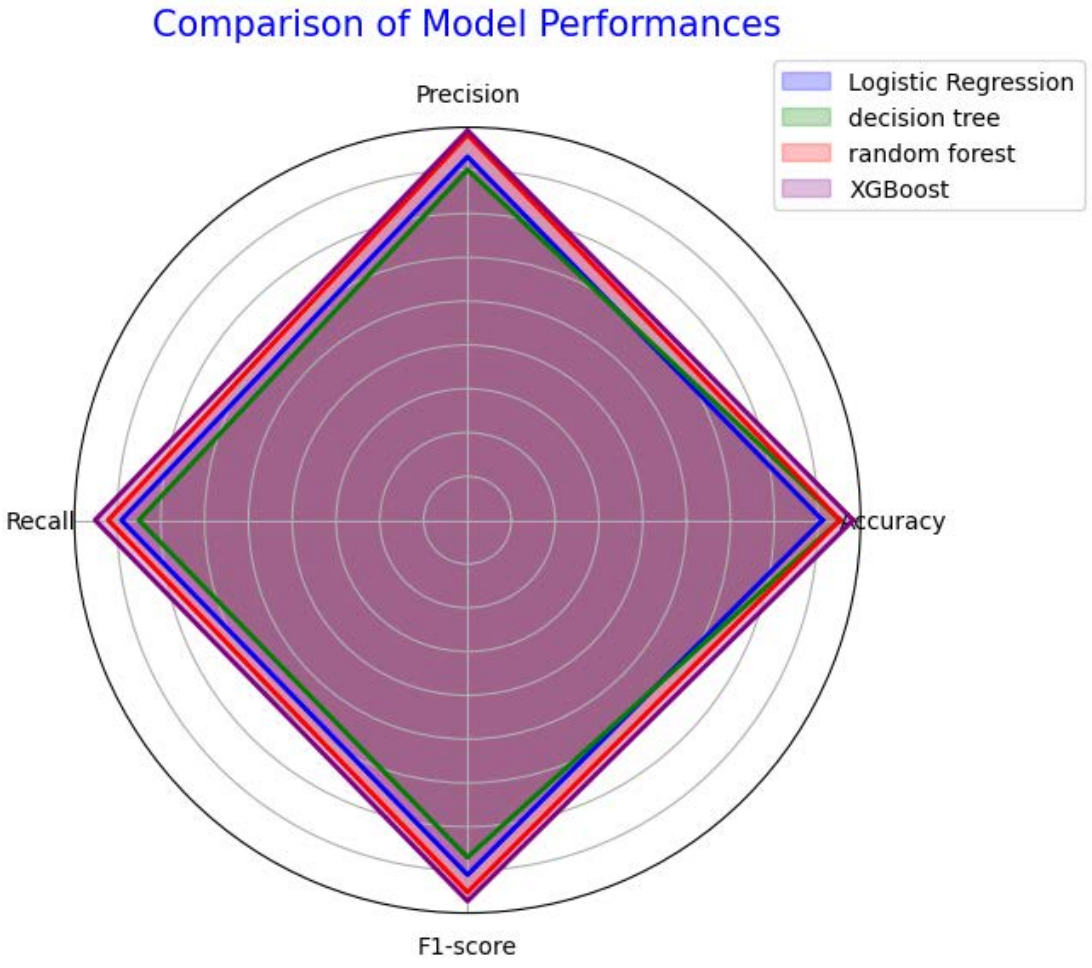

Fig. 2 Radar Chart of Four Models

Fig. 2 clearly show the differences in accuracy, precision, recall and F1 scores between models. By analyzing the shape and size of each radar plot, it is easily to determine which models provide more balanced performance and which excel in specific areas.

XGBoost had the best recall and F1 score, which will come in handy in order to identify more positive heart disease cases, reducing the risk of false negatives. Besides, while the decision tree yields the highest accuracy, its problems with recall mean it can fail in true positive cases. This imbalance is pretty dangerous in a medical scenario because the false negatives lead to misdiagnosis and grave consequences.

F1 scores yield the important balance between accuracy and recall. Models with high F1 scores, such as XGBoost in this case, strike a good balance between correctly predicting positive cases and minimizing false negatives. This balance is of prime importance in cardiac risk prediction since both false positives and false negatives might have substantial impacts on patient outcomes.

Lastly, there is obvious room for improvement for those models that show weak performances for certain metrics. For example, the random forest model, while showing high accuracy and precision, performs very low on recall. These challenges being faced could be dealt with by adjusting hyper-parameters, adjusting decision thresholds, or assigning more weights to positive cases. In these areas of focus, the model's ability to detect true positives will be enhanced hence improving its general reliability in clinical application.

5. Conclusion

The XGBoost model thus exhibited very good predictive performance, showing 88% accuracy, 89% precision, 85% recall, and an F1 score of 0.87, hence turning out to be a very effective model in classifying patients into high- and low-risk categories for heart disease. Besides, its balance between precision and recall, along with regularization techniques used, reduces over fitting and allows good predictions even for unbalanced datasets, thus finding very good applications in healthcare. However, these may be limited to direct clinical applicability because of over fitting on small datasets and complexities in interpreting model outputs. This means that the methodologies of feature engineering and model interchangeability will need refinement in future studies, possibly through hybrid approaches, to enhance their clinical utility. Larger and more diverse datasets could allow the generalization of the model across populations and conditions. This suggests that the XGBoost model may be helpful in a clinical decision support system for early cardiac disease identification, leading to prompt intervention and more individualized treatment regimens, improving patient outcomes and enhancing preventative healthcare tactics.

References

[1] World Health Organization (WHO). Cardiovascular diseases (CVDs). 2021.

[2] Yusuf S, Joseph P, Rangarajan S, Islam S, Mente A, Hystad P, Teo K. Modifiable risk factors, cardiovascular disease, and mortality in 155,722 individuals from 21 high-income, middle-income, and low-income countries (PURE): a prospective cohort study. The Lancet, 2020, 395(10226): 795-808.

[3] Sharma V, Yadav S, Gupta M. Heart disease prediction using machine learning techniques. In 2020 2nd international conference on advances in computing, communication control and networking (ICACCCN). IEEE, 2020: 177-181.

[4] Aljanabi M, Qutqut M H, Hijjawi M. Machine learning classification techniques for heart disease prediction: a review. International Journal of Engineering & Technology, 2018, 7(4): 5373-5379.

[5] Patel P, Chaudhary A, Bhatnagar S. A review of machine learning algorithms for heart disease prediction. Algorithms, 2023, 16(2): 88.

[6] Mehrabani-Zeinabad H, Pourmohammadi K, Rahimi M. Predictive modeling of heart disease using machine learning algorithms. BMC Medical Informatics and Decision Making, 2023, 23: 21.

[7] Podkowinski D, Beka S, Mursch-Edlmayr A S, et al. A swept source optical coherence tomography angiography study: imaging artifacts and comparison of non-perfusion areas with fluorescein angiography in diabetic macular edema. PLoS One, 2021, 16(4): e0249918.

[8] Jindal H, Agrawal S, Khera R, Jain R, Nagrath P. Heart disease prediction using machine learning algorithms. IOP Conference Series: Materials Science and Engineering, 2021, 1022(1): 012072.

[9] Fitriyani N L, Syafrudin M, Alfian G, Rhee J. HDPM: an effective heart disease prediction model for a clinical decision support system. IEEE Access, 2020, 8: 133034-133050.

[10] Ayon S I, Islam M M, Hossain M R. Coronary artery heart disease prediction: a comparative study of computational intelligence techniques. IETE Journal of Research, 2022, 68(4): 2488-2507.

[11] Kavitha M, Gnaneswar G, Dinesh R, Sai Y R, Suraj R S. Heart disease prediction using hybrid machine learning model. 2021 6th international conference on inventive computation technologies (ICICT). IEEE, 2021: 1329-1333.

[12] UCI Machine Learning Repository. Heart disease dataset.