Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ30923

Heart Disease Prediction Base on Machine Learning

Lingxiao Ren

School of Business, University College Dublin, Dublin, Ireland
renlingxiao107@outlook.com

Abstract. Heart disease is a major cause of death around the world. Accurate predictions in the early stages can provide additional time for treatment and significantly increase the likelihood of survival. Traditional methods rely on manual diagnosis, which usually occurs when patients already have obvious symptoms. This study uses machine learning to predict heart disease and identify key risk factors, aiming to find a model that can provide accurate and reliable predictions to assist in early clinical diagnosis of heart disease. Among all models, logistic regression performs best with 88.04% accuracy, and its precision, recall, and F1 score also performed the best among the four models. It also identifies the key factors that influence heart disease risk, the research indicates that factors such as sex, type of chest pain, fasting blood sugar, and the slope of the peak exercise ST segment are the main determinants of the risk of heart disease. The results show that this model is reliable for medical risk prediction and decision support.

Keywords: Heart Disease, Machine Learning, Logistic Regression, Prediction

1. Introduction

Heart disease is a major cause of death around the world [1]. Therefore, early and accurate diagnosis is crucial [1]. Traditional clinical methods rely on manual interpretation, which can be slow and subjective [2]. Machine learning (ML) is a method that uses data to guide its analysis, it can promptly assist in making human decisions by identifying potential trends in clinical data, thereby enhancing efficiency.

In the past, many scholars have conducted ML research on heart disease prediction. Kumar and Maben conducted predictions by using random forest (RF), neural network, extreme gradient boosting, and logistic regression (LR) [3]. They evaluated the accuracy of each model and concluded that the RF model was the optimal one [3]. The same conclusion that RF is the optimal model was also reached by Sharma et al [4]. By using and evaluating four models including Naïve Bayes, decision tree (DT), RF, and Support Vector Machine (SVM), they believed that RF could reduce noise and the risk of overfitting [4]. Anbuselvan used seven models such as LR, SVM, and K-Nearest Neighbors (KNN) [5]. After evaluating the accuracy, it was found that RF achieved an accuracy of 86.89%, which was the highest among the seven models [5]. Jindal et al. also employed the RF model [6]. At the same time, they also used the LR and KNN models [6]. They found that the KNN model had the highest accuracy (88.5%), and they identified that the most crucial factors were chest pain and resting blood pressure [6]. However, Rindhe et al. used and evaluated three models -

DOI: 10.54254/2755-2721/2026.TJ30923

SVM, RF, and artificial neural network, and found that the best model was SVM, with its accuracy reaching 84% [7]. Although there have been many studies that have integrated ML into heart disease prediction, most of these studies have focused on improving the accuracy of the models, and few have deeply analyzed the importance of the features and linked the models to clinical significance. This study learns from previous scholars' research to summarize the ML models with generally better performance, to maintain a high level of accuracy in the model. It also uses a composite dataset to ensure that the model is applicable to multiple datasets rather than just a single one, so as to accurately identify the main factors influencing the risk of heart disease, thereby ensuring the reliability of the model in clinical applications.

This research applied four ML approaches—LR, DT, RF, and KNN—to perform heart disease prediction. The aim was to train a model that could accurately and reliably predict heart disease and identify the key factors affecting heart disease, to assisting in clinical decision-making and improving efficiency.

2. Methodology

2.1. Dataset introduction

This dataset is sourced from Kaggle and was formed by Federosioriano by integrating multiple previously separate but uncombined datasets [8]. It contains several commonly used datasets for predicting heart diseases [8]. By integrating the observations and features from multiple datasets, the prediction results become more reliable, to make sure the accuracy of the prediction model in clinical applications [8]. This dataset consists of 918 patient records and 11 features related to physiological indicators, as well as the dependent variable - whether the patient has heart disease [8].

2.2. Data preparation

After initial checks, no missing value was found. Further analysis shows that one recorded RestingBP is 0, and another 172 records have a Cholesterol value of 0, both of which are beyond the range of clinical standards. These data are treated as errors and replaced with medians to reduce deviations, which follows statistical and clinical medical recommendations [9,10].

2.3. Exploratory data analysis

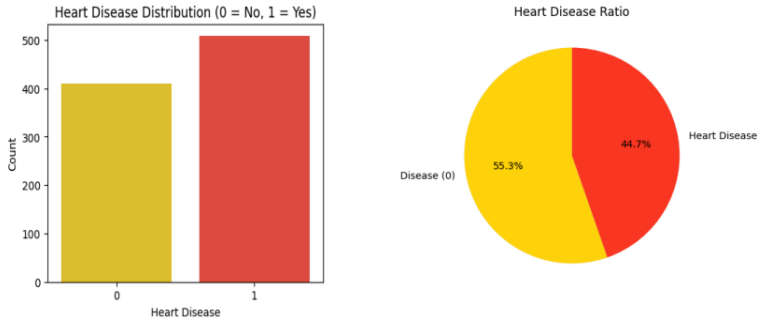
An exploratory data analysis was conducted to better understand the underlying patterns and structure of the dataset, identify distribution patterns, detect potential outliers, and explore potential relationships between predictor and target variables. Statistical summaries and graphical techniques were also used to provide insights that aided in feature selection and model development.

2.3.1. Target variable distribution

Figure 1 shows a bar graph illustrating the distribution of heart disease cases, showing the number of patients classified as having the disease (1) and the number of patients who did not have the disease (0). Pie chart shows the same information in relative proportions. The results show that out of a total of 918 observations, approximately 55.3% of the participants were diagnosed with heart disease, while 44.7% did not have the disease. This indicates that the dataset is relatively balanced. From a modeling perspective, this distribution is advantageous because it reduces the potential bias in the

DOI: 10.54254/2755-2721/2026.TJ30923

classification model and eliminates the need for additional resampling techniques such as oversampling or under sampling.



(a) This is the first part (b) This is the second part

Figure 1. Bar chart and pie chart of heart disease distribution (original)

2.3.2. Correlation analysis of numerical and categorical variables

Figure 2 is a correlation heatmap of numerical and categorical variables, with the aim of finding which features are strongly linked to the target variable. The three features, Exercise Angina, Old peak, and Sex, show strong positive links to the possibility of having heart disease. This indicates that patients with these features (having exercise-induced angina, a higher old peak value, and being male) have more possibility of having heart disease. On the contrary, the three features, ST Slope, MaxHR, and Chest Pain Type, show a strong negative correlation. This means that the ST slope is increasing, the maximum heart rate achieved is higher, and those who reported having typical angina chest pain were more likely to not have heart disease.

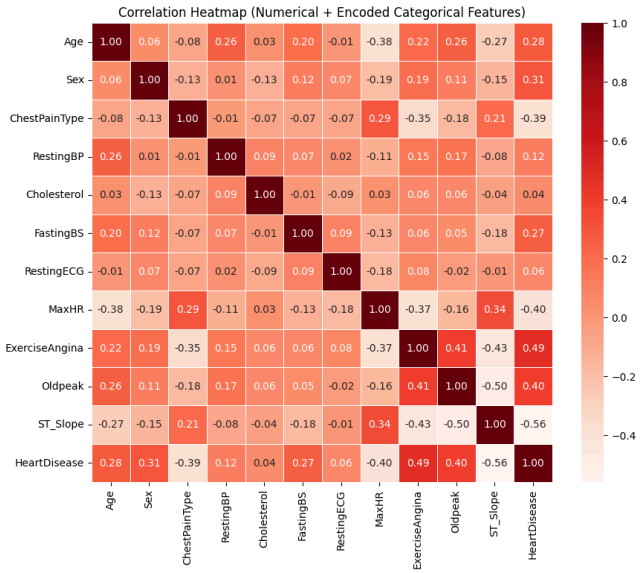


Figure 2. Correlation heatmap of numerical and categorical features (original)

DOI: 10.54254/2755-2721/2026.TJ30923

2.4. Feature engineering

2.4.1. Label encoding

To prepare a dataset for ML, label encoding is first used to convert categorical variables into numerical representations. According to Boslavskaya and Korolev, label encoding performs comparably to one-hot encoding for binary classification tasks, with no significant difference in predictive accuracy [11]. Therefore, label encoding is sufficient and computationally more efficient for binary categorical variables [11].

In this study, label coding was used for binary variables. The variable sex was encoded as 0 for women and 1 for men. Similarly, patients with exercise-induced angina pectoris were coded as yes 1 and no as 0. This transformation ensures that the characteristics of binary variables are correctly interpreted by the model without introducing artificial ordering. Label coding retains the categorical meaning and converts it into a numerical form, allowing algorithms to learn the potential correlation between these predictive variables and the target variables, thereby making the stability and reliability of the model Improve [11].

2.4.2. One-hot encoding

In addition to binary variables, this dataset also contains categorical features with more than two distinct categories: Chest Pain Type, Resting ECG, and ST Slope. To make these variables effectively usable by ML algorithms, one-hot encoding is employed.

This transformation enables the model to capture the independent impact of each clinical attribute on heart disease prediction while avoiding any assumptions about sequential relationships between features. It also ensures that all categorical information is preserved in a mathematically sound form, this helps the model understand the role of each feature and improves its predictive performance in later steps.

2.4.3. Feature scaling (standardization)

After processing the categorical variables, all numerical variables were standardized using the Standard Scaler method. Standardization ensure that each feature receives equal weight during model training. This prevents features with a wide range of values from having an uneven impact on the learning process. Standardization can significantly improve model performance when feature values vary significantly [12].

2.4.4. Variance inflation factor

To examine whether the predictors exhibited multicollinearity, variance inflation factors were computed after all variables had been encoded. High levels of multicollinearity can interfere with model interpretation and reduce the stability of coefficient estimates, particularly in regression-based algorithms. All VIF values were under the common cutoff of 5, and this shows that the independent variables did not have strong multicollinearity. The variable Sex had the highest VIF at 4.49, and the next highest values came from ST Slope-Up (4.42) and ST Slope-Flat (4.23). This finding indicates that the dataset is well suited for model training, without the need for additional feature elimination or dimensionality reduction. By performing this step, the study ensured the reliability and predictive accuracy of the subsequent classification model.

DOI: 10.54254/2755-2721/2026.TJ30923

2.4.5. Train–test split

To evaluate the model's predictive accuracy and generalization ability, the data set was split into a training and a testing set. In accordance with the standard ML convention, we used an 80:20 division ratio. This approach allows trained models to be evaluated with unknown data, resulting in unbiased estimates of predicted performance [13].

In classification tasks, appropriate data division is important. This is to prevent overlearning and allow models to learn actual patterns rather than simply remembering data sets. The random state parameter was fixed to ensure the reproducibility of the results. By adopting this segmentation strategy, this study was able to obtain enough data to train robust models while retaining sufficient samples for reliable performance evaluations. This process lays an important foundation for evaluating the effectiveness of subsequent models in predicting heart disease [13].

2.5. Modelling

After preprocessing the data, four taught learning models were developed to predict heart disease. These models included LR, DT, RF, and KNN. All models were trained using the standardized amount of characteristics in the training set and then tested with unknown data.

LR was selected as the baseline model in this study. It is commonly applied to binary classification tasks and estimates the likelihood that a patient has heart disease. The method assumes a linear association between the predictors and the outcome variable. One of its key strengths is its interpretability. LR makes it possible to identify how each feature influences the prediction, which is particularly valuable in medical decision-making.

DT were used to capture nonlinear patterns. The DT divides the data into smaller groups based on the characteristic quantity, forming a clear and simple set of rules. This makes it easier to interpret. The DT can show how variables such as age, cholesterol levels, and resting blood pressure interact to affect the risk of heart disease.

RF were introduced to improve stability and accuracy. Multiple DT were built and the results were integrated. This reduces over conformity and generates more reliable predictions. RF find the features that have the strongest effect on heart disease prediction and give information that can support later analysis.

The KNN is classified as a non-parametric model. Classify each patient by comparing it with the most similar cases in the data set. Because it is based on distance, it is essential to standardize the characteristics. KNN is simple and intuitive and helps to detect local patterns in the data.

3. Results

Table 1 presents a comparison of the performance results from the four models. The results show that LR has the highest accuracy, precision, recall, F1 score, and cross-validation accuracy. Therefore, LR is the optimal model selected as the final choice.

DOI: 10.54254/2755-2721/2026.TJ30923

Table 1. Model performance comparison

Model	Accuracy	Precision	Recall	F1 Score	CV Accuracy
LR	0.8804	0.8846	0.9020	0.8932	0.8474
KNN	0.8478	0.8700	0.8529	0.8614	0.8297
DT	0.8043	0.8300	0.8137	0.8218	0.8120
RF	0.8696	0.8750	0.8922	0.8835	0.8474

Figure 3 shows feature importance based on the coefficients of the LR model. The absolute value of each coefficient indicates the degree of influence of the feature on the prediction. Whether the coefficient is positive or negative indicates its importance.

Features such as sex, Chest Pain Type-ASY, and FastingBS have high positive coefficients. These variables have a strong link to a higher risk of heart disease in the model's predictions. In contrast, ST Slope-Up and Chest Pain Type-NAP have large negative coefficients, meaning they also influence the results, but in the opposite direction.

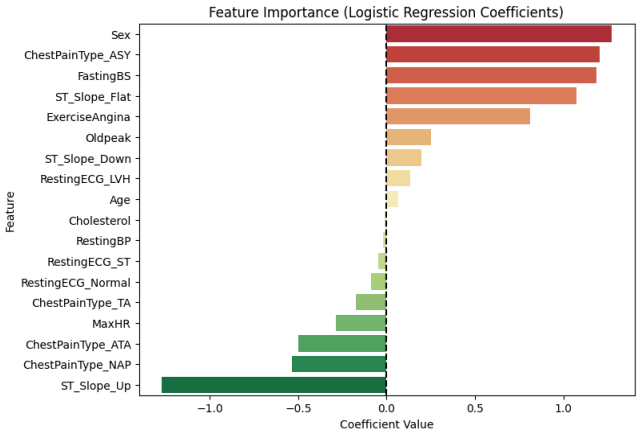


Figure 3. Feature importance (original)

Figure 4 shows how prediction accuracy changes across five key features: Sex, Chest Pain Type-ASY, ST Slope-Up, FastingBS, and ST Slope-Flat. For gender, the model made mostly correct predictions for both males and females. Male samples showed slightly higher accuracy. In Chest Pain Type-ASY, both groups had a large number of correct predictions. For ST Slope-Up and ST Slope-Flat, the model classified most samples correctly, showing steady performance across slope categories. In FastingBS, people with normal blood sugar levels were predicted correctly more often, while those with high blood sugar still showed good accuracy.

DOI: 10.54254/2755-2721/2026.TJ30923

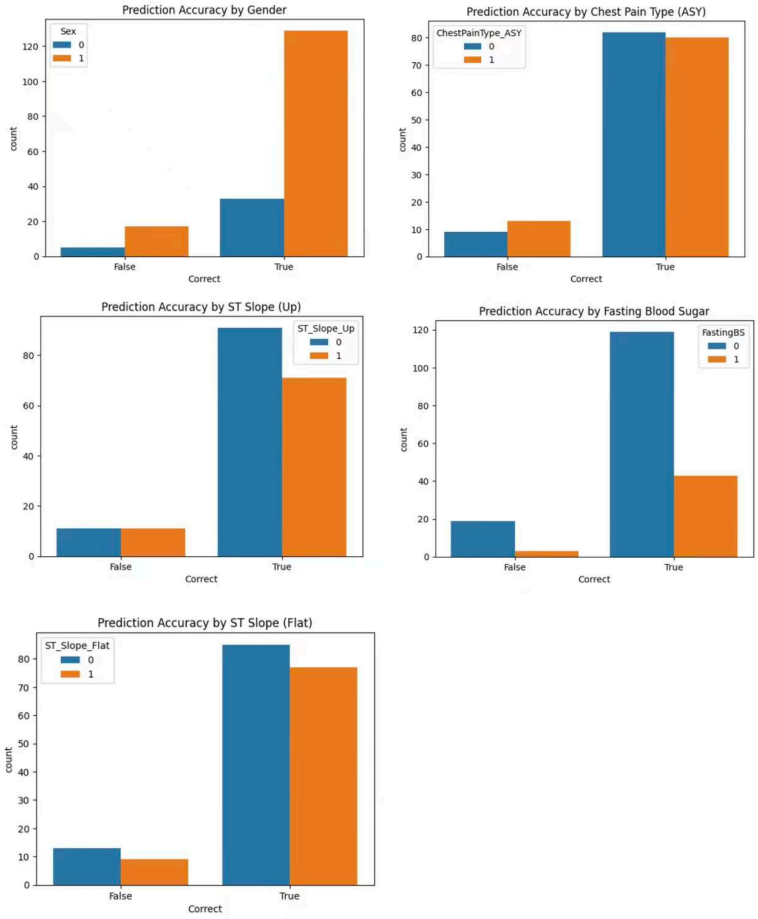


Figure 4. Prediction accuracy distribution across key features using LR (original)

4. Discussion

This research assessed four different ML algorithms for the purpose of predicting heart disease. As a result, the LR model showed the best performance, with the highest accuracy, precision, recall, and F1 score, with an accuracy of 88.04%. This means that this model is easy to understand and performs well in the medical field. LR can also show the impact of each feature on the outcome, so it helps to support clinical decision-making. However, one of the drawbacks is that it assumes a linear relationship between the predictory variable and the resulting logarithm odds, which may miss more complex clinical patterns [14].

To improve model performance, future work can try mixing simple linear models with deep learning methods. Hassan et al. developed a combined modeling approach that integrates pre-trained deep neural networks, principal component analysis, and LR for heart disease prediction [15]. They used the Cleveland dataset for both training and testing [15]. Their results show that the model reached 91.79% accuracy on the training set and 93.33% accuracy on the test set [15].

Despite the encouraging results, the study still has some limitations. The data set used has only 918 observations, which limits the universality of the model and may affect its stability and reliability in clinical applications. Dhiman et al. pointed out that the insufficient sample size will lead to instability in the prediction of new samples, thus limiting the universal applicability of the model [16]. Future research can address these limitations by expanding the data set and evaluating the model based on real hospital data to improve its clinical applicability.

DOI: 10.54254/2755-2721/2026.TJ30923

Previous studies indicate that deep learning models trained on larger heart disease datasets often achieve higher AUC values than conventional linear models [17]. Other studies also show that the combination of deep learning with an interpretable hybrid integration framework can not only achieve strong predictive performance but also ensure clinical transparency [18]. Therefore, expanding the data set and adopting deep learning methods can improve the model's ability to perform well on new data. Future work may explore explainable integrated or hybrid deep learning models to strike a balance between predictive accuracy and real-world clinical usefulness.

5. Conclusion

This study employs four ML methods to predict heart disease based on 11 clinical features. Among these models, LR demonstrates the strongest performance, achieving an accuracy of 88.04%, a precision of 88.46%, and a recall of 90.20%. In addition, the model shows strong cross-validation accuracy, high universality and lack of significant overfitting.

The analysis of the importance of characteristic quantities shows that sex, chest pain type-ASY, fasting blood glucose, ST slope-flat and exercise angina are the most influential variables. These results are consistent with known cardiovascular risk factors, proving the clinical value of the model.

However, this research has some limitations. Due to the relatively small data set used, the robustness of the model may be reduced when applied to a larger or more diverse population. External verification with data from other sources is required to verify the generalization.

Future research may use deep learning or hybrid models to improve prediction accuracy. Adding more physiological and behavioral characteristics can also improve the explanatoryness. In general, this study shows that ML can effectively predict heart disease and help early clinical diagnosis.

References

[1] D i Cesare, M., Perel, P., Taylor, S., Kabudula, C., Bixby, H., Gaziano, T. A., McGhie, D. V., Mwangi, J., Pervan, B., Narula, J., Pineiro, D., & Pinto, F. J. (2024). The Heart of the World. Global heart, 19(1), 11.

[2] Talin, I.A., Abid, M.H., Khan, MM. (2022). Finding the influential clinical traits that impact on the diagnosis of heart disease using statistical and machine-learning techniques. Sci Rep12, 20199

[3] Kumar, A., & Maben, M. (2025). Heart disease prediction using machine learning algorithms. In 2025 Third International Conference on Networks, Multimedia and Information Technology (NMITCON) 1-6. IEEE.

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

[5] Anbuselvan, P. (2020). Heart disease prediction using machine learning techniques. Int. J. Eng. Res. Technol, 9, 515-518.

[6] Jindal, H., Agrawal, S., Khera, R., Jain, R., & Nagrath, P. (2021). Heart disease prediction using machine learning algorithms. In IOP conference series: materials science and engineering (Vol. 1022, No. 1, p. 012072). IOP Publishing.

[7] Rindhe, B. U., Ahire, N., Patil, R., Gagare, S., & Darade, M. (2021). Heart disease prediction using machine learning. Heart Disease, 5(1).

[8] FEDESORIANO. (2021, September). Heart Failure Prediction Dataset. Retrieved October 25, 2025, from www.kaggle.com website: https: //www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

[9] Kwak, S. K., & Kim, J. H. (2017) Statistical data preparation: management of missing values and outliers. Korean journal of anesthesiology, 70(4), 407–411.

[10] Gijs F N Berkelmans, Read, S. H., Soffia Gudbjörnsdottir, Wild, S. H., Franzen, S., van, … J A N Dorresteijn. (2022). Population median imputation was noninferior to complex approaches for imputing missing values in cardiovascular prediction models in clinical practice. ScienceDirect, 145(0895-4356), 70–80.

DOI: 10.54254/2755-2721/2026.TJ30923

[11] Poslavskaya, E., & Korolev, A. (2023). Encoding categorical data: Is there yet anything 'hotter' than one-hot encoding?. arXiv preprint arXiv: 2312.16930.
[12] Pinheiro, J. M. H., de Oliveira, S. V. B., Silva, T. H. S., Saraiva, P. A. R., de Souza, E. F., Godoy, R. V., ... & Becker, M. (2025). The impact of feature scaling in machine learning: Effects on regression and classification tasks. arXiv preprint arXiv: 2506.08274.
[13] Xu, Y., & Goodacre, R. (2018). On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. Journal of analysis and testing, 2(3), 249-262.
[14] Chesnaye, N. C., van Diepen, M., Dekker, F., Zoccali, C., Jager, K. J., & Stel, V. S. (2024) Non-linear relationships in clinical research. Nephrology Dialysis Transplantation, 40(2). https: //doi.org/10.1093/ndt/gfae187
[15] Hassan, M. M., Zaman, S., Rahman, M. M., Bairagi, A. K., El-Shafai, W., Rathore, R. S., & Gupta, D. (2024). Efficient prediction of coronary artery disease using machine learning algorithms with feature selection techniques. Computers and Electrical Engineering, 115, 109130.
[16] Dhiman, P., Ma, J., Qi, C., Bullock, G. S., Sergeant, J. C., Riley, R. D., & Collins, G. S. (2023) Sample size requirements are not being considered in studies developing prediction models for binary outcomes: a systematic review. BMC Medical Research Methodology, 23(1).
[17] Sajeev, S. et al. (2019). Deep Learning to Improve Heart Disease Risk Prediction. In: Liao, H., et al. Machine Learning and Medical Engineering for Cardiovascular Health and Intravascular Imaging and Computer Assisted Stenting. MLMECH CVII-STENT 2019 2019. Lecture Notes in Computer Science, 11794. Springer, Cham.
[18] Hasnat, Md Abrar, Jobayer, M., Hasan, M., & Alam, (2025, November). Interpretable Heart Disease Prediction via a Weighted Ensemble Model: A Large-Scale Study with SHAP and Surrogate Decision Trees. Retrieved from arXiv.org website: https: //arxiv.org/abs/2511.01947?utm_source=chatgpt.com