Predicting Heart Disease Using Machine Learning: Analysis and New Insights

Yiming Chen[1, *]

[1]SANTA MONICA COLLEGE, LA, USA

*Corresponding author: chen_yiming01@student.smc.edu

Abstract:

Heart disease remains the leading cause of death worldwide, making the development of efficient diagnostic tools crucial. With the rise of machine learning, data-driven predictive models offer promising avenues for early detection and intervention. This study addresses the challenge by first conducting data preprocessing, including handling missing values, encoding categorical variables, and normalizing numerical features to ensure model quality. Then, exploratory data analysis (EDA) reveals that features such as age, gender, chest pain type, and others are significantly correlated with heart disease. Multiple machine learning algorithms were implemented to compare their performance in heart disease prediction, including Logistic Regression, K-Nearest Neighbors, Decision Tree, Random Forest, and Support Vector Machine (SVM). Finally, feature selection techniques, such as correlation analysis, recursive feature elimination (RFE), and LASSO regression, were employed to optimize the model's input features. Experimental results indicate that SVM achieved the best predictive performance, with an accuracy of 83.61%. This research demonstrates the potential of machine learning models for heart disease prediction in clinical settings, and future improvements can be made by exploring ensemble learning and deep learning techniques.

Keywords: Heart Disease Prediction; Machine Learning; Feature Selection; Data Preprocessing; Support Vector Machine.

1. Introduction

Heart disease, a term encompassing various cardiovascular conditions, remains the leading cause of death globally. The World Health Organization (WHO) estimates that 17.9 million deaths annually result from cardiovascular diseases, particularly heart disease. Early detection of heart disease is essential to reducing mortality rates and preventing complications. Traditionally, heart disease diagnosis relies on a combination of clinical evaluations, including electrocardiograms (ECG), stress tests, and angiograms, alongside patient symptoms and medical history. However, these methods can be subjective, costly, and time-consuming. In recent years, machine learning (ML) has emerged as a powerful tool in healthcare, enabling data-driven insights and predictive analytics. Using historical patient data, ML models can identify patterns and correlations that are not immediately evident through traditional diagnostic techniques. These models offer the potential for earlier diagnosis, personalized treatment, and better patient outcomes. In the context of heart disease, machine learning offers new avenues for more accurate, reliable, and automated prediction tools, which can assist healthcare professionals in making informed decisions [1-3]

The application of machine learning in heart disease prediction holds significant promise. Given the high global mortality rate, accurate predictive models can be critical in early diagnosis and intervention [4]. Implementing ML-based diagnostic tools has the potential to save lives, optimize healthcare resources, and reduce the economic burden associated with heart disease. This study focuses on enhancing heart disease prediction using machine learning models to improve diagnostic accuracy and increase the practical applicability of these models in clinical settings.

The motivation behind this study stems from the ongoing challenge of improving heart disease diagnosis through automated and reliable methods. Traditional clinical methods, though effective, have limitations in terms of speed, cost, and accessibility. Furthermore, many existing machine learning models do not generalize well across diverse patient populations and lack interpretability, which is essential in a clinical setting. By improving machine learning models' accuracy, robustness, and transparency, we hope to bridge the gap between research and real-world application in healthcare.

We begin with data preprocessing, including handling missing values, encoding categorical variables, and normalizing numerical features to ensure consistency across the dataset. Key insights into the data are gained through visualizations, such as histograms, box plots, and heat maps. These tools help identify patterns and relationships between features critical for heart disease prediction. Multiple machine learning algorithms are implemented, including Logistic Regression, K-Nearest Neighbors (KNN), Decision Trees, Random Forest, and Support Vector Machines (SVM) [5-8]. Each model is trained and evaluated using standard performance metrics like accuracy, precision, recall, and F1 score. Advanced feature selection techniques, such as Correlation Analysis, Recursive Feature Elimination (RFE), and LASSO regression, are applied to ensure the models use only the most relevant features, reducing noise and improving efficiency [9,10].

Experiments show that the Support Vector Machine (SVM) performed best, achieving an accuracy of 83.61%. Random Forest was followed closely, with 81.97% accuracy. The advanced feature selection and ensemble learning techniques can significantly improve model performance. Additionally, features such as maximum heart rate (thalach), chest pain type (cp), and the number of major vessels colored by fluoroscopy (ca) consistently emerged as the most predictive factors. This study contributes to the growing body of research on heart disease prediction by demonstrating the effectiveness of machine learning algorithms. Incorporating advanced feature selection methods and ensemble learning enhances predictive accuracy and makes the models more applicable to clinical practice. Future work will explore deep learning techniques and focus on model interpretability and real-world applicability.

The rest of this paper is organized as follows. Section 2 provides an overview of the proposed machine learning methods, including model descriptions and feature selection techniques. Section 3 presents the experimental results, comparing the performance of different models. Section 4 discusses this study's limitations and outlines future research directions.

2. Method

2.1 Section Headings Machine Learning Models

Logistic Regression: Logistic Regression is one of the simplest and most commonly used linear models for binary classification problems. In this model, the relationship between the features and the target variable is modelled using a linear equation, and the result is transformed into a probability using a sigmoid function. The model assigns a probability between 0 and 1 to each prediction. For instance, if the probability exceeds a threshold of 0.5, the model predicts the patient has heart disease. Logistic Regression is efficient and interpretable, making it a good baseline model, though it may struggle with complex, non-linear relationships between features.

Decision Tree: Decision Trees are non-linear models that split the dataset into subsets based on feature values, forming a tree structure. In heart disease prediction, the decision tree starts by selecting the feature that best separates the data (usually based on information gain or Gini impurity) and forms a decision node. The dataset is then split based on this feature, and the process continues recursively until the tree reaches a specified depth or all data points in a node belong to the same class.

Random Forest: Random Forest is an ensemble learning method that improves decision trees by combining the results of multiple trees. Each tree is built using a random subset of the data and a random subset of features, which helps to reduce overfitting and improve generalization. In heart disease prediction, random forest aggregates the results from all the trees using a majority vote. The diversity among the trees allows random forests to capture complex patterns in the data while maintaining robustness against overfitting.

Support Vector Machine: Support Vector Machine (SVM) is a robust classification algorithm that aims to find the optimal hyperplane that separates the data into two classes with the largest possible margin. In cases where the data is not linearly separable, SVM can use kernel functions to project the data into a higher-dimensional space where a hyperplane can be constructed.

Nearest Neighbors: Nearest Neighbors (KNN) is a non-parametric classification method that assigns a class to a data point based on the majority class among its K-nearest neighbours. The model calculates the distance (typically using Euclidean distance) for heart disease prediction between the new data and all the training data points. It then selects the K nearest points and assigns the most common class. KNN is simple and intuitive, but it can be computationally expensive, especially with large datasets, as it requires calculating the distance to every training point during prediction.

2.2 Feature Selection

The first method applied in this analysis is correlation analysis. This method calculates the correlation between each feature and the target variable (whether the patient has heart disease). For the heart disease dataset, features such as maximum heart rate (thalach), chest pain type (cp), and ST segment slope (slope) were found to be highly correlated with the presence of heart disease. The model can focus on the most relevant predictors by identifying and selecting these highly correlated features, improving training efficiency and accuracy. Next, RFE was used to refine the feature set further. RFE is a model-based feature selection method that recursively removes the least important features from the dataset. The process starts by training a model using all available features and then ranks them based on their contribution to the model's prediction performance. The least significant feature is removed, and the model is re-trained. This process repeats until the optimal number of features remains. RFE helps to eliminate noise from the dataset and ensure that only the most impactful features are retained, leading to a more efficient and effective model. Finally, LASSO regression was applied as another feature selection method. LASSO introduces an L1 regularization term to the loss function during model training, which penalizes less important feature coefficients, pushing them towards zero. This effectively removes irrelevant or redundant features from the model. LASSO is especially useful in high-dimensional datasets with many correlated features, as it simplifies the model while retaining the most significant predictors. This method ensures that the final model is not only accurate but also robust and easy to interpret.

3. Literature References Experimental Results and Analysis

3.1 Overview of the Heart Disease Dataset

The Heart Disease dataset used in this analysis is a well-known UCI Machine Learning Repository dataset. It consists of 303 records, each containing 14 attributes that describe various patient features, including the presence or absence of heart disease, as shown in Table 1.

3.2 Dataset Preprocessing

The dataset underwent several preprocessing steps:

(1) Handling Missing Values: The dataset contains missing values, either filled using imputation techniques or handled by removing the corresponding records.

(2) Encoding Categorical Variables: Variables like chest pain type, restecg, and thal were converted to numerical formats using one-hot encoding.

(3) Feature Scaling: Continuous variables like age, trestbps, chol, and thalach were standardized to have a mean of 0 and a standard deviation of 1 to ensure the models perform optimally.

Table 1. Description of Features in the Heart Disease Dataset

Feature name	Description
Age	Age of the patient
Sex	Gender of the patient (1 = male, 0 = female)
Chest Pain Type (cp)	Type of chest pain experienced by the patient (1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic)
Resting Blood Pressure (trestbps)	Resting blood pressure in mm Hg
Serum Cholesterol (chol)	Serum cholesterol in mg/dl

Fasting Blood Sugar (fbs)	Fasting blood sugar > 120 mg/dl (1 = true, 0 = false)
Resting ECG (restecg)	Resting electrocardiographic results (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy)
Maximum Heart Rate (thalach)	Maximum heart rate achieved
Exercise-Induced Angina (exang)	Exercise-induced angina (1 = yes, 0 = no)
ST Depression (oldpeak)	ST depression induced by exercise relative to rest
Slope of ST Segment (slope)	The slope of the peak exercise ST segment (1 = upsloping, 2 = flat, 3 = downsloping)
Number of Major Vessels (ca)	Number of major vessels (0-3) colored by fluoroscopy
Thalassemia (thal)	Thalassemia (3 = normal, 6 = fixed defect, 7 = reversible defect)
Target (target)	Diagnosis of heart disease (1 = heart disease, 0 = no heart disease)

3.3 Results and Analysis

As shown in Table 2, the Support Vector Machine (SVM) emerged as the best-performing model with an accuracy of 83.61%. It showed excellent performance across all metrics, including accuracy, precision and recall. Random Forest followed closely with an accuracy of 81.97% and demonstrated robust performance in handling noisy data and preventing overfitting. Through the feature selection process, the most important predictors of heart disease were identified as maximum heart rate (thalach), chest pain type (cp), and the number of major vessels colored by fluoroscopy (ca). These features consistently showed high predictive power across multiple models. Although the accuracy of Logistic Regression and K-Nearest Neighbors (KNN) was relatively lower at 78.69% and 77.05% respectively, they still provided valuable insights under specific conditions. For example, Logistic Regression is useful for interpreting the linear relationship between features and the target variable, while KNN performs well in small datasets or simpler classification tasks.

Table 2. Performance Comparison of Machine Learning Models for Heart Disease Predication

Model	Accuracy
Logistic Regression (LOG)	78.69%
Support Vector Machine (SV)	83.61%
K-Nearest Neighbors (KNN)	77.05%
Decision Tree (DT)	72.13%
Random Forest (RF)	81.97%
Gradient Boosting Classifier (GBC)	80.33%

4. Discussion

While this study demonstrates the effectiveness of various machine learning models in predicting heart disease, several limitations need to be addressed. First, the dataset used in this study is relatively small, with only 303 records. While machine learning models can still perform well on small datasets, the limited amount of data may hinder the generalization of these models to larger or more diverse populations. In clinical settings, more extensive and varied datasets would be needed to ensure that the models perform reliably across different patient demographics and conditions. Another limitation is the imbalanced nature of some features. Although techniques such as normalization and encoding were applied, some key features (e.g., categorical variables like “chest pain type” and “thal”) may still introduce bias into the models. Additionally, some important clinical features, such as family history, socioeconomic factors, or lifestyle data, are missing from the dataset, which could further improve the model’s prediction capabilities. Finally, while advanced feature selection techniques like Recursive Feature Elimination (RFE) and LASSO were used, not all potential feature interactions were fully explored. Non-linear relationships between features may exist, and a more in-depth exploration using advanced techniques like polynomial features or interaction terms could lead to more accurate predictions.

Future research should focus on addressing these limitations. A key priority would be to utilize larger and more diverse datasets that can improve the model's generalizability and reduce potential biases. Collaborations with medical institutions to access real-world clinical datasets could help create more robust and applicable models. These datasets could include a broader range of clinical factors, including genetic predispositions, lifestyle factors (e.g., diet, smoking habits), and other comorbidities that could significantly impact heart disease prediction. In addition to expanding the dataset, future studies should explore the application of more advanced machine learning techniques. Ensemble learning methods, such as XGBoost or LightGBM, which typically outperform individual models, could be further explored. These methods are designed to handle large datasets, feature interactions, and missing data more effectively, making them ideal for heart disease prediction tasks. Another promising area of research is the integration of deep learning techniques, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), which have shown promise in medical image analysis and time-series data. For example, incorporating ECG data into CNN models may provide further improvements in prediction accuracy. Lastly, model interpretability remains a critical challenge, particularly in medical applications. Models like Support Vector Machines (SVM) and Random Forests can be difficult to interpret, which can hinder their adoption in clinical settings. Future research should explore explainable AI (XAI) techniques, such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations), to provide healthcare professionals with insights into why a model makes specific predictions.

5. Conclusion

This study highlights the potential of machine learning models for predicting heart disease, with Support Vector Machines (SVM) emerging as the top-performing model, achieving an accuracy of 83.61%. Random Forests and Gradient Boosting Classifiers also showed strong results, demonstrating that ensemble models can significantly improve prediction performance by reducing overfitting and capturing complex feature interactions. The significance of this research lies in its contribution to the growing body of work exploring machine learning in healthcare. By improving heart disease prediction, such models can aid in early detection and intervention, potentially saving lives and reducing healthcare costs. However, further research is necessary to improve model generalizability, handle more comprehensive datasets, and ensure model interpretability to transition these models from research to real-world clinical use. Future studies should focus on using larger datasets, exploring advanced machine learning techniques, and applying explainable AI methods to build more reliable and transparent diagnostic tools.

References

[1] Lu Shasha, Jianyu Yang, Yu Gu, et al. Advances in Machine Learning Processing of Big Data from Disease Diagnosis Sensors. *ACS sensors*, 2024, 9(3): 1134-1148.

[2] Sunil Kumar, Harish Kumar, Gyanendra Kumaret, et al. A methodical exploration of imaging modalities from dataset to detection through machine learning paradigms in prominent lung disease diagnosis: a review. BMC Medical Imaging, 2024, 24(1): 30.

[3] Mitra Montazeri, Mahdieh Montazeri. Machine learning models for predicting the diagnosis of liver disease. Koomesh, 2024, 16(1): 53-59.

[4] Arvind Pandey, Borge Akshay Shivaji, Malika Acharya, et al. Mitigating class imbalance in heart disease detection with machine learning. Multimedia Tools and Applications, 2024: 1-26.

[5] Neena Suresh, Binu Thomas, Jeena Joseph. Bibliometric Analysis and Visualization of Scientific Literature on Heart Disease Classification Using a Logistic Regression Model. Cureus, 2024, 16(6).

[6] Rian Oktafiani, Arief Hermawan, Donny Avianto. Max Depth Impact on Heart Disease Classification: Decision Tree and Random Forest. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), 2024, 8(1): 160-168.

[7] Avijit Kumar Chaudhuri, Sulekha Das, Arkadip Ray. An Improved Random Forest Model for Detecting Heart Disease Data-Centric AI Solutions and Emerging Technologies in the Healthcare Ecosystem. CRC Press, 2024: 143-164.

[8] Ayodeji Olalekan Sala, Tsehay Admassu Assegie, Gunjan Chhabra, et al. Heart Disease Detection Model Using Support Vector Machine with Feature Selection. 2024 2nd International Conference on Advancement in Computation & Computer Technologies (InCACCT). IEEE, 2024: 204-207.

[9] Komal Kumar Napa, Angati Kalyan Kumar, Sangeetha Murugan, et al. Early Prediction of Chronic Heart Disease With Recursive Feature Elimination And Supervised Learning Techniques. Int J Artif Intell ISSN, 2024, 2252(8938): 8938.

[10] Chiao-Lin Hsu, Pin-Chieh Wu, Fu-Zong Wu, et al. LASSO-derived model for the prediction of lean-non-alcoholic fatty liver disease in examinees attending a routine health check-up. Annals of Medicine, 2024, 56(1): 2317348.