ISSN 2959-6157

From Data to Diagnosis: Effective Machine Learning-based Heart Disease Prediction

Tianyi Lu

Tandon School of Engineering, New York University, NY, USA
tl3307@nyu.edu

Abstract:

Heart disease is a leading cause of mortality worldwide, contributing to nearly 18 million deaths annually. Early detection is critical but remains a significant challenge due to the limitations of traditional diagnostic methods, which can be prone to human error. This study aims to enhance heart disease prediction using machine learning (ML) by comparing the performance of three ML models: Support Vector Machine (SVM), Random Forest (RF), and XGBoost. A dataset containing 12 features from 918 patients was used, with preprocessing steps such as one-hot encoding for categorical variables and MinMax scaling for numerical features. The models were trained and evaluated using 5-fold cross-validation to ensure robustness. Random Forest demonstrated the highest accuracy at 82.78%, followed closely by SVM (82.67%) and XGBoost (81.58%). The feature importance analysis identified ST_Slope as the most significant predictor of heart disease, providing important insights into which features are most influential in the diagnosis process. While the Random Forest model outperformed the others, this study also highlights the need for better interpretability in ML models, especially in medical applications where understanding the relationships between features is crucial. Future research should focus on improving model transparency to bridge the gap between accuracy and practical application in clinical settings.

Keywords: Machine learning; heart disease; artificial intelligence.

1. Introduction

Heart disease accounts for a large part of mortality worldwide, resulting in nearly 18 million deaths annually [1]. Despite significant advancements in medical science, early detection of heart disease continues to pose substantial challenges. Although traditional diagnostic methods are effective in most cases, they are subject to human error, which makes them less ideal for early intervention. This drawback necessitates innovative solutions to enhance heart disease prediction and diagnosis.

As machine learning demonstrated remarkable capabilities in pattern recognition and predictive analytics, it has increasingly been integrated into various aspects of healthcare, from diagnostic imaging to personalized medicine. For instance, Machine Learning (ML) algorithms have been adopted to analyze medical images for detecting skin cancers, with higher accuracy than traditional methods [2]. Additionally, predictive analytics powered by ML has been employed to enact individualized sepsis treatment, enabling more timely and effective interventions [3]. Moreover, ML models have played a critical role in genomics, where they are used to identify genetic markers associated with specific diseases, paving the way for targeted therapies and personalized treatment plans [4]. These advancements have not only improved diagnostic accuracy but also largely cut the time and cost of traditional medical procedures.

When focusing on heart disease, several studies have applied ML algorithms to enhance diagnosis performance. Researchers have developed models using algorithms such as logistic regression, decision trees, and neural networks to predict the likelihood of heart disease based on clinical data [5]. For example, a study by Johnson et al. used a neural network to predict heart disease outcomes [6]. Similarly, Zhang et al. applied random forest algorithms to a large dataset, identifying key predictors of heart disease with high precision [7]. These studies demonstrate the potential of ML in predicting heart disease with higher accuracy compared to traditional human diagnosis.

Despite the promising results, existing ML models do have several limitations. One of the primary issues is model interpretability. Many high-performing models, are complex and act as “black boxes,” making it difficult for doctors to understand the logic inside. For example, while neural networks have shown high accuracy in heart

disease prediction, their lack of transparency limits their practical application in clinical settings [7]. This study aims to address key research gaps in heart disease prediction by integrating predictive accuracy with model interpretability. While previous studies have utilized machine learning algorithms like Random Forest, SVM, and XGBoost for heart disease prediction, they often focus solely on accuracy, neglecting the importance of understanding the contribution of individual features. Additionally, these studies are frequently based on homogeneous datasets, limiting the models' generalizability across diverse populations. By systematically comparing these algorithms, this research will identify the most effective model, ensuring robustness and accuracy. Moreover, using the best-performing model, feature importance will be analyzed to reveal the most critical factors influencing heart disease outcomes. This approach not only enhances predictive accuracy but also bridges the gap between complex machine learning models and their practical application in clinical settings by providing transparent, interpretable results that can inform and guide heart disease diagnosis.

2. Method

2.1 Dataset Preparation

The dataset utilized in this study is sourced from Kaggle [8]. The dataset collected 12 features from 918 patients. These features include categorical and numerical variables such as sex and age. These features, representing the patient's medical profile, are comprehensive and valuable for heart disease prediction. The dataset's target variable HeartDisease is binary, indicating whether a patient has heart disease (1) or not (0).

Exploratory Data Analysis (EDA) was conducted to understand the distribution of these features and examine correlations between variables. Distribution plots (Fig. 1) and a heat map (Fig. 2) were generated to visualize these relationships:

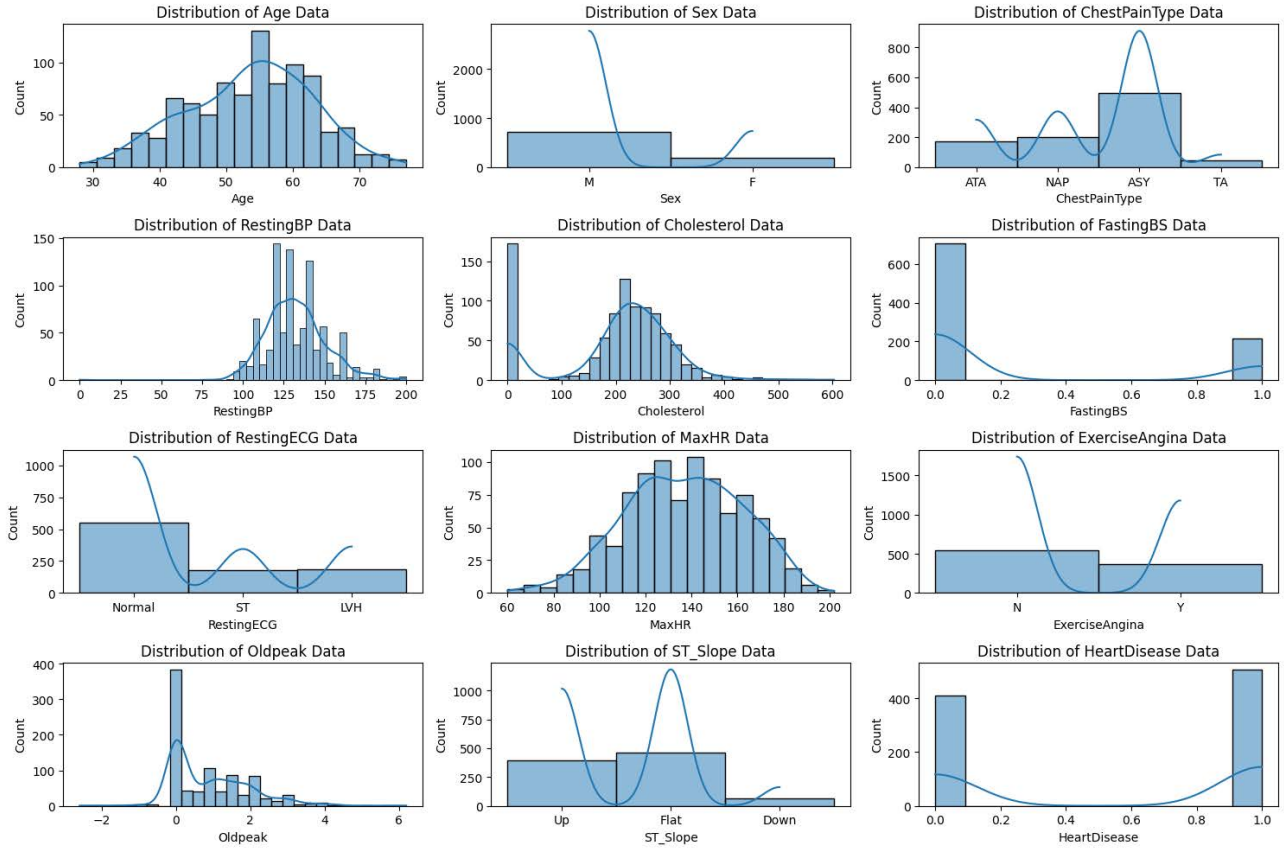

Fig. 1 Distribution of each feature (Photo/Picture credit: Original)

Fig. 2 Heat map (Photo/Picture credit: Original)

2.1.1 Preprocessing

Before applying the machine learning models, several preprocessing steps were conducted. First, categorical variables were encoded using one-hot encoding to convert them into numerical format suitable for the non-tree-based model – Support Vector Machine (SVM). This process ensures that the model can interpret these variables without assuming any ordinal relationship between categories. Next, numerical features were scaled using MinMaxScaler, ensuring that features with larger ranges do not disproportionately influence the model's predictions since the algorithm is distance-based. On the other hand, label encoding was implemented to deal with the categorical variables for the two tree-based models -- Random Forest and XGBoost, as ordinality does not affect the models' performance. The dataset was then split into training and testing sets using a 5-fold cross-validation approach. This method was chosen to ensure that the model's performance is validated across multiple subsets of the data, thereby reducing the likelihood of overfitting and ensuring robust performance.

2.2 Machine Learning Models

In this study, three machine learning models were implemented: SVM, Random Forest (RF), and extreme Gradient Boosting (XGBoost). These models were chosen because of their effectiveness in classification problems. Each model was achieved using Python's scikit-learn library.

To assess the performance of the models, two metrics, average 5-Fold CV accuracy and the confusion matrix, were used. These metrics are essential for the models' evaluation of heart disease prediction, filtering out both poorly performing and overfitting models.

2.2.1 Support vector machine

SVM is primarily implemented for classification tasks and is effective in high-dimensional spaces. It works by finding the hyperplane that best separates the classes in the feature space. SVM includes different kernels as choices for non-linear relationships. In this study, four kernel functions were explored, including the sigmoid kernel and the linear kernel, to determine the optimal model for heart disease prediction. The model's parameters were set to default values.

2.2.2 Random forest

Random Forest constructs multiple decision trees during training and outputting the classification results of individual trees [9], which reduces the risk of overfitting associated with single decision trees. In this study, the number of trees in the forest (n_estimators) was set to 200, and the maximum depth of the trees was set to default. Meanwhile, feature importance was visualized to demonstrate features that had the most significant impact on heart disease predictions. This interpretability is crucial in medical applications, where understanding the factors influencing a diagnosis is as important as the prediction itself.

2.2.3 XGBoost

XGBoost is an advanced implementation of gradient boosting. It constructs a sequential collection of trees, where each tree corrects the errors of its predecessor. XGBoost includes regularization terms in its objective function to prevent overfitting, which is suitable for complex datasets. In this study, the model's hyperparameters, such as the learning rate (eta), maximum tree depth (max_

depth), and the number of boosting rounds (n_estimators), were set to default. Like Random Forest, XGBoost also provides feature importance scores, which were used to showcase critical factors for heart disease prediction.

3. Results and Discussion

3.1 The Performance of Models

The performance of the SVM model was evaluated using four different kernels: linear, polynomial, RBF, and sigmoid. Each kernel was assessed using a 5-fold cross-validation strategy, and the results are summarized in Table 1. The linear kernel demonstrated the highest consistency with an average cross-validation accuracy of 82.672%. It also presented a well-balanced confusion matrix, with 92 true positives, 10 false positives, 67 true negatives, and 15 false negatives. These results indicate that the linear kernel effectively discriminates between classes, making it a reliable choice for heart disease detections.

Table 1. Comparison of performance based on various SVM kernels

Kernel	Average 5-Fold CV Accuracy (%)	True Positives	False Positives	True Negatives	False Negatives	Confusion Matrix Accuracy (%)
Linear	82.672	92	10	67	15	86.413%
Polynomial	82.017	92	12	65	15	85.326%
RBF	82.669	89	10	67	18	84.782%
Sigmoid	77.876	77	12	65	30	77.173%

The polynomial and RBF kernels exhibited a higher number of false positives and false negatives, resulting in lower confusion matrix accuracy (85.326% and 84.782%) compared to the linear kernel (86.413%). The sigmoid kernel, with the lowest accuracy of 77.876%, also showed a significant imbalance in its confusion matrix, with 12 false positives and 30 false negatives, indicating that it is the worst kernel for this dataset.

The linear kernel was proved to be the most effective, likely due to the data's linear separability. The other kernels, particularly the sigmoid, underperformed, possibly due to overfitting or the data's inherent characteristics that did not favor complex transformations. The choice of the linear kernel for the final model is supported by both its superior accuracy and the more balanced confusion matrix.

Table 2. Comparison of performance based on different models

Model	Average 5-Fold CV Accuracy (%)	True Positives	False Positives	True Negatives	False Negatives	Confusion Matrix Accuracy (%)
SVM (Linear)	82.672	92	10	67	15	86.413%
Random Forest	83.651	95	8	69	12	89.130%
XGBoost	81.580	97	13	56	18	83.152%

Building on the decision to utilize the SVM with a linear kernel, further analysis was conducted to compare its performance with two additional models: Random Forest and XGBoost. The Random Forest model slightly outperformed the SVM model with an accuracy of 83.651%, achieving a higher confusion matrix accuracy of 89.130% at the same time shown in Table 2. In contrast, the XGBoost model, with an accuracy of 81.580%, underperformed relative to both SVM and Random Forest.

The Random Forest model is identified as the most effective among the three due to several key advantages. Its ensemble approach, which aggregates the decisions of multiple decision trees, provides robust performance by reducing the risk of overfitting, a common drawback of single-tree models. This characteristic allows Random Forest to capture complex, non-linear relationships within the data more effectively than the linear SVM. While the SVM with a linear kernel excels in cases where the data is linearly separable, it is less adaptable to datasets with non-linear boundaries, leading to its slightly lower performance. On the other hand, XGBoost, although powerful, requires careful hyperparameter tuning and is more sensitive to overfitting if not properly regularized. Given these considerations, Random Forest stands out as the best

model due to its balance between accuracy and resilience to overfitting.

3.2 Feature Importance

After identifying the Random Forest model as the most effective model for this classification task, an analysis of the feature importance was conducted to understand which factors contribute most significantly to the model's predictions.

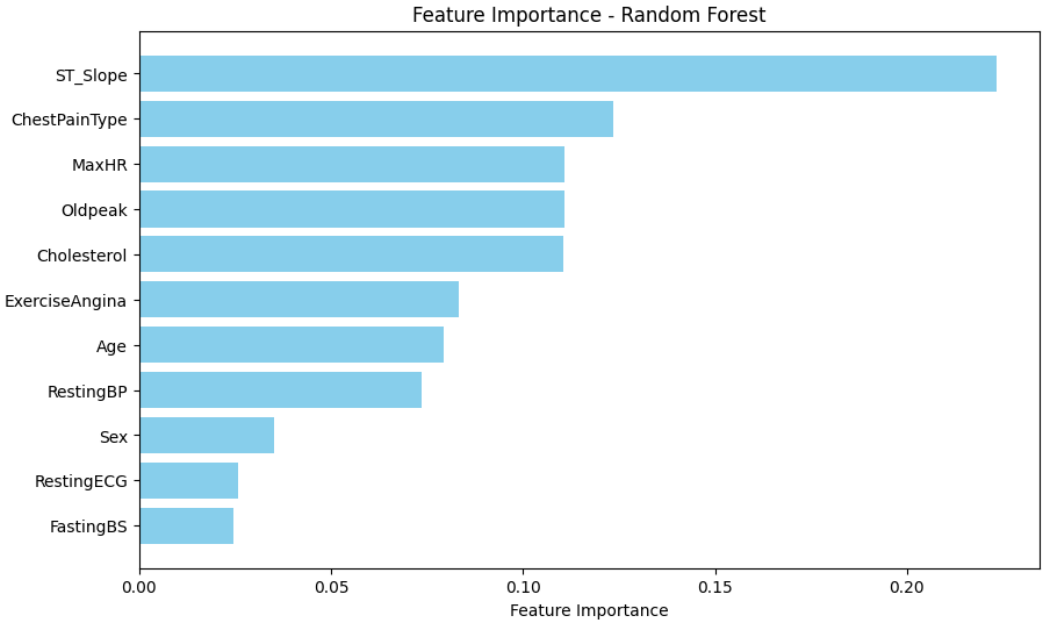

Fig. 3 Feature Importance from RF model (Photo/Picture credit: Original)

The resulting feature importance chart (Fig. 3) indicates that the most influential feature is ST_Slope, followed by ChestPainType, MaxHR, and Oldpeak.

The prominence of ST_Slope as the top predictor is noteworthy, as this feature reflects the slope of the peak exercise ST segment in an electrocardiogram (ECG), a well-known indicator in the diagnosis of heart disease [10]. Medically, this makes sense as an important factor, as abnormalities in ST_Slope often correlate with heart disease. However, the model's emphasis on this feature could suggest that the Random Forest may be overly reliant on this single indicator, potentially at the expense of other important but less prominent features.

Although random forest excels in predictive accuracy, it provides limited insight into the specific relationships between features. For instance, while ST_Slope appears as the most influential factor, the model does not explain how this feature interacts with others or why it outweighs traditionally significant factors like Cholesterol or Age. From a clinical perspective, this lack of transparency may pose challenges when attempting to translate model predictions into actionable medical decisions.

4. Conclusion

This study demonstrated the effective application of machine learning models for heart disease prediction, comparing the performance of SVM with a linear kernel, Random Forest, and XGBoost. The Random Forest model was identified as the most effective, balancing predictive accuracy and robustness. By leveraging its ensemble approach, Random Forest reduces overfitting risks while effectively capturing complex relationships within the data. Although SVM with a linear kernel also performed well, its reliance on linear separability limits its adaptability to more complex datasets. XGBoost, despite its powerful predictive capabilities, requires careful tuning and is more prone to overfitting, which limits its practical application in this context.

The feature importance analysis further highlighted 'ST_Slope' as the most significant predictor, reinforcing the medical relevance of this feature in diagnosing heart disease. However, the Random Forest model's lack of interpretability remains a limitation, as it does not clearly explain the interactions between features, which can be crucial for clinical decision-making. Future work could focus on enhancing the interpretability of machine learning models or integrating them with more transparent models to ensure that predictions are both accurate and actionable in medical settings.

References

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

[Internet]. 2023 [cited 2024 Sep 5]. Available from: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. nature. 2017 Feb;542(7639):115-8.

[3] Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. Nature medicine. 2018 Nov;24(11):1716-20.

[4] Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nature Reviews Genetics. 2015 Jun;16(6):321-32.

[5] Nithya B, Ilango V. Predictive analytics in health care using machine learning tools and techniques. In2017 International Conference on Intelligent Computing and Control Systems (ICICCS) 2017 Jun 15 (pp. 492-499). IEEE.

[6] Johnson KW, Torres Soto J, Glicksberg BS, Shameer K, Miotto R, Ali M, Ashley E, Dudley JT. Artificial intelligence in cardiology. Journal of the American College of Cardiology. 2018 Jun 12;71(23):2668-79.

[7] Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. InProceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining 2015 Aug 10 (pp. 1721-1730).

[8] Soriano F. Heart failure prediction dataset [Internet]. Kaggle. 2021 [cited 2024 Sep 5]. Available from: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction/data

[9] Biau G, Scornet E. A random forest guided tour. Test. 2016 Jun;25:197-227.

[10] Kardashian M, Elamin MS, Mary DASG, Whitaker W, Smith DR, Boyle R, Stoker JB, Linden RJ. The slope of ST segment/heart rate relationship during exercise in the prediction of severity of coronary artery disease. Eur Heart J. 1982 Oct;3(5):449-458.