Evaluating the Performance of Various Classifiers in Heart Disease Detection

July 31, 2024

Contents

Abstract

This study evaluates the performance of multiple machine learning classifiers for heart disease detection. Heart disease remains a leading cause of death globally, and accurate detection is crucial for early intervention. Various classifiers, including Logistic Regression, Support Vector Machines (SVM), Linear Discriminant Analysis (LDA), Decision Trees, K-Nearest Neighbors (KNN), Random Forests, AdaBoost, and an Ensemble Voting method were implemented and evaluated on their performance metrics. This report details the methodologies, results, and code implementations, culminating in a comprehensive analysis of the classifiers' effectiveness in heart disease detection.

Chapter 1

Introduction

Heart disease remains one of the most significant health challenges worldwide. According to the World Health Organization (WHO), heart disease is the leading cause of death globally, accounting for approximately 17.9 million lives each year. This underscores the critical need for early and accurate detection methods to mitigate the risks associated with heart disease.

1.1 Background

Heart disease encompasses a range of conditions that affect the heart, including coronary artery disease, arrhythmias, heart valve problems, and heart failure. The most common cause is the narrowing or blockage of coronary arteries, leading to coronary artery disease, which can result in heart attacks.

1.2 Motivation

The motivation behind this study is rooted in the urgent need for improved diagnostic tools. Traditional methods, such as electrocardiograms (ECGs), echocardiograms, and stress tests, while effective, can be enhanced by integrating machine learning techniques. Machine learning offers the potential to analyze vast amounts of data and identify patterns that might be missed by conventional methods.

1.3 Objectives

The primary objective of this research is to evaluate the performance of various machine learning classifiers in detecting heart disease. The study aims to:

1. Implement and train multiple classifiers on heart disease data.
2. Compare the classifiers based on accuracy, precision, recall, and F1 scores.
3. Identify the most effective classifier for heart disease detection.

Chapter 2

Literature Review

Machine learning in healthcare has garnered significant attention in recent years. Numerous studies have explored the application of various machine learning algorithms for heart disease detection.

2.1 Early Studies

One of the earliest studies in this domain was conducted by Detrano et al. (1989), who applied logistic regression to the Cleveland heart disease dataset. Their work demonstrated the potential of statistical models in predicting heart disease. Subsequent studies have expanded on this by exploring more complex models and larger datasets.

2.2 Recent Advances

Recent advances have seen the application of ensemble methods, such as Random Forests and AdaBoost, which combine the predictions of multiple models to improve accuracy. For instance, Soni et al. (2011) compared the performance of various classifiers, including Decision Trees, SVMs, and ensemble methods, finding that ensemble methods often outperform single classifiers.

2.3 Deep Learning

Deep learning, a subset of machine learning, has also been explored for heart disease detection. Researchers have utilized neural networks to analyze ECG signals and other medical data. Although deep learning models can be highly accurate, they require large amounts of data and significant computational resources.

2.4 Gaps in Research

While many studies have demonstrated the effectiveness of machine learning in heart disease detection, there are still gaps that need to be addressed. For instance, the interpretability of models remains a challenge. Clinicians need to understand the decision-making process of these models to trust and effectively use them in practice.

Chapter 3

Methodology

The methodology of this study involves several steps, from data preprocessing to model evaluation. Each step is crucial for ensuring the reliability and accuracy of the results.

3.1 Data Preprocessing

Data preprocessing involves cleaning the dataset and handling missing values. In this study, the dataset was obtained from the UCI Machine Learning Repository, which contains several missing values. These were addressed by replacing missing values with the median of the respective feature.

3.2 Feature Selection

Feature selection is a critical step in machine learning. It involves identifying the most relevant features that contribute to the prediction of heart disease. In this study, features such as age, sex, chest pain type, resting blood pressure, cholesterol levels, fasting blood sugar, resting ECG results, maximum heart rate achieved, exercise-induced angina, ST depression, slope, number of major vessels, and thalassemia were selected.

3.3 Model Training

Various machine learning classifiers were trained on the dataset, including:

- **Logistic Regression**: A statistical model that predicts the probability of a binary outcome.
- **Support Vector Machines (SVM)**: A model that finds the optimal hyperplane for classifying data points.
- **Linear Discriminant Analysis (LDA)**: A technique used for dimensionality reduction while preserving class separability.
- **Decision Trees**: A model that splits the data into subsets based on the most significant features.
- **K-Nearest Neighbors (KNN)**: A non-parametric method that classifies instances based on their proximity to other instances.

- **Random Forests**: An ensemble method that combines multiple decision trees.
- **AdaBoost**: An ensemble technique that focuses on difficult-to-classify instances.
- **Ensemble Voting**: A method that combines the predictions of multiple classifiers.

3.4 Model Evaluation

The models were evaluated using various metrics, including:

- **Accuracy**: The proportion of correctly classified instances.
- **Precision**: The proportion of positive identifications that were actually correct.
- **Recall**: The proportion of actual positives that were identified correctly.
- **F1 Score**: The harmonic mean of precision and recall, providing a balance between the two.

Chapter 4

Data Description

The dataset used in this study is the Heart Disease Dataset from the UCI Machine Learning Repository. This dataset is widely used in machine learning research due to its comprehensive collection of medical records and its relevance to heart disease detection.

4.1 Dataset Overview

The dataset contains 14 features and 303 instances. The features include:

- **Age**: Age of the patient.
- **Sex**: Gender of the patient.
- **Chest Pain Type (cp)**: Type of chest pain experienced.
- **Resting Blood Pressure (trestbps)**: Resting blood pressure in mm Hg.
- **Cholesterol (chol)**: Serum cholesterol in mg/dl.
- **Fasting Blood Sugar (fbs)**: Fasting blood sugar ¿ 120 mg/dl.
- **Resting ECG (restecg)**: Resting electrocardiographic results.
- **Maximum Heart Rate Achieved (thalach)**: Maximum heart rate achieved during exercise.
- **Exercise-Induced Angina (exang)**: Presence of exercise-induced angina.
- **ST Depression (oldpeak)**: ST depression induced by exercise relative to rest.
- **Slope**: The slope of the peak exercise ST segment.
- **Number of Major Vessels (ca)**: Number of major vessels colored by fluoroscopy.
- **Thalassemia (thal)**: A blood disorder involving lower-than-normal amounts of an oxygen-carrying protein.

4.2 Data Preprocessing

The dataset was preprocessed to handle missing values and scale features. Missing values were replaced with the median of the respective feature, and features were scaled using standard scaling techniques to ensure that they had a mean of 0 and a standard deviation of 1.

4.3 Feature Engineering

Feature engineering involves creating new features or transforming existing ones to improve model performance. In this study, interactions between features, polynomial features, and discretization of continuous variables were explored to enhance the predictive power of the models.

4.4 Data Splitting

The dataset was split into training and test sets using an 80-20 split. The training set was used to train the models, while the test set was used to evaluate their performance. A stratified split was used to ensure that the class distribution was similar in both the training and test sets.

Chapter 5

Model Descriptions

5.1 Logistic Regression

Logistic Regression is a statistical model that predicts the probability of a binary outcome based on one or more predictor variables. It is widely used in binary classification problems and is particularly effective when the relationship between the features and the target variable is linear.

5.2 Support Vector Machines (SVM)

Support Vector Machines (SVM) are supervised learning models that find the optimal hyperplane that separates the data points into different classes. SVMs are effective in high-dimensional spaces and are known for their ability to handle non-linear data using kernel functions.

5.3 Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) is a technique used for dimensionality reduction while preserving class separability. LDA projects the data onto a lower-dimensional space that maximizes the separation between the classes.

5.4 Decision Trees

Decision Trees are non-parametric models that split the data into subsets based on the most significant features. They are easy to interpret and visualize, making them popular in many applications. However, they are prone to overfitting, which can be mitigated by using ensemble methods.

5.5 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a non-parametric method that classifies an instance based on the majority class among its k-nearest neighbors. KNN is simple to implement and understand but can be computationally expensive for large datasets.

5.6 Random Forests

Random Forests are ensemble methods that combine multiple decision trees to improve classification accuracy. Each tree in the forest is trained on a bootstrap sample of the data, and the final prediction is made by averaging the predictions of all the trees.

5.7 AdaBoost

AdaBoost is an ensemble technique that combines multiple weak classifiers to form a strong classifier. It focuses on difficult-to-classify instances by assigning higher weights to them, thereby improving the overall performance of the model.

5.8 Ensemble Voting

Ensemble Voting is a method that combines the predictions of multiple classifiers by averaging their probabilities or taking a majority vote. This approach leverages the strengths of different classifiers to improve the overall accuracy.

Chapter 6

Experimental Setup

The experimental setup involves preparing the data, training the models, and evaluating their performance.

6.1 Data Preparation

The data was prepared by handling missing values, scaling features, and splitting the dataset into training and test sets. Feature engineering techniques were applied to create new features and enhance the predictive power of the models.

6.2 Model Training

Each model was trained on the training set using the default hyperparameters. Hyperparameter tuning was performed using cross-validation to identify the best parameters for each model.

6.3 Model Evaluation

The models were evaluated on the test set using various performance metrics, including accuracy, precision, recall, and F1 scores. The results were compared to identify the best-performing model.

Chapter 7

Results and Discussion

7.1 Training Accuracies

Model	Training Accuracy
Logistic Regression	0.8388
SVM Optimized	0.8595
LDA	0.8471
Decision Tree Optimized	0.8926
KNN	0.8554
Random Forest Optimized	0.9504
AdaBoost Optimized	0.8760
Ensemble Voting	0.8802

Table 7.1: Training Accuracies of Various Classifiers

7.2 Test Accuracies

Model	Test Accuracy
Logistic Regression	0.8033
SVM Optimized	0.8033
LDA	0.8197
Decision Tree Optimized	0.7705
KNN	0.8361
Random Forest Optimized	0.7541
AdaBoost Optimized	0.7705
Ensemble Voting	0.8361

Table 7.2: Test Accuracies of Various Classifiers

7.3 Precision Scores

Model	Precision

Logistic Regression	0.8000
SVM Optimized	0.7838
LDA	0.8056
Decision Tree Optimized	0.8276
KNN	0.8286
Random Forest Optimized	0.7813
AdaBoost Optimized	0.8065
Ensemble Voting	0.8286

Table 7.3: Precision Scores of Various Classifiers

7.4 Recall Scores

Model	**Recall**
Logistic Regression	0.8485
SVM Optimized	0.8788
LDA	0.8788
Decision Tree Optimized	0.7273
KNN	0.8788
Random Forest Optimized	0.7576
AdaBoost Optimized	0.7576
Ensemble Voting	0.8788

Table 7.4: Recall Scores of Various Classifiers

7.5 F1 Scores

Model	**F1 Score**
Logistic Regression	0.8235
SVM Optimized	0.8286
LDA	0.8406
Decision Tree Optimized	0.7742
KNN	0.8529
Random Forest Optimized	0.7692
AdaBoost Optimized	0.7813
Ensemble Voting	0.8529

Table 7.5: F1 Scores of Various Classifiers

7.6 Best Model Analysis

- **Best model on training data**: Random Forest Optimized with an accuracy of 0.9504.
- **Best model on test data**: KNN with an accuracy of 0.8361.
- **Best model based on F1 Score**: KNN with an F1 Score of 0.8529.
- **Best model based on Precision Score**: KNN with a Precision Score of 0.8286.
- **Best model based on Recall Score**: SVM Optimized with a Recall Score of 0.8788.

Chapter 8

Conclusion

The application of machine learning techniques in heart disease detection offers promising avenues for improving diagnostic accuracy and early intervention. In this study, we evaluated the performance of several machine learning classifiers, including Logistic Regression, Support Vector Machines (SVM), Linear Discriminant Analysis (LDA), Decision Trees, K-Nearest Neighbors (KNN), Random Forests, AdaBoost, and Ensemble Voting.

The results of this study indicate that the K-Nearest Neighbors (KNN) classifier demonstrated the highest test accuracy and F1 score, making it the most reliable model for heart disease detection in our dataset. KNN's simplicity and effectiveness in handling the given features highlight its potential for practical implementation in medical diagnostics. Furthermore, the KNN model also achieved the best precision score, indicating its capability in minimizing false positives, which is crucial in medical applications where the cost of false alarms can be significant.

Random Forests, despite showing the highest accuracy on the training data, exhibited signs of overfitting, as indicated by the drop in accuracy when evaluated on the test data. This overfitting suggests that while Random Forests can capture complex patterns within the training data, they may not generalize as well to unseen data without further tuning or more sophisticated approaches to prevent overfitting.

The Ensemble Voting classifier, which combines the predictions of multiple models, also performed well, reinforcing the idea that leveraging the strengths of various classifiers can enhance overall performance. This approach is particularly beneficial in scenarios where individual models have complementary strengths and weaknesses.

In conclusion, this study demonstrates that machine learning classifiers, particularly KNN, can be highly effective tools for heart disease detection. These findings support the integration of machine learning models into clinical decision support systems, potentially leading to more accurate and timely diagnoses, thereby improving patient outcomes. However, it is important to acknowledge that the success of these models depends on the quality and representativeness of the training data, and continuous efforts are needed to validate and refine these models in diverse clinical settings.

Chapter 9

Future Work

One potential direction is the exploration of deep learning techniques. Deep learning models, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have shown significant success in various medical imaging and time-series analysis tasks. Applying these models to heart disease detection, especially with richer datasets that include ECG signals and imaging data, could potentially uncover more nuanced patterns and improve diagnostic accuracy. Additionally, incorporating more diverse and comprehensive datasets could enhance the generalizability of the models. The current study utilized the Cleveland Heart Disease dataset, which, while valuable, may not capture the full spectrum of patient demographics and clinical presentations. Future studies should aim to include larger and more diverse populations to ensure that the models are robust and applicable across different patient groups and healthcare settings.

Another important aspect is the interpretability of machine learning models. Clinicians need to understand how these models make decisions to trust and effectively use them in practice. Future research could focus on developing interpretable models or employing techniques such as SHAP (SHapley Additive exPlanations) values and LIME (Local Interpretable Model-agnostic Explanations) to provide insights into model predictions. Enhancing interpretability could facilitate the integration of these models into clinical workflows and improve clinician acceptance.

Moreover, further hyperparameter optimization and feature engineering could be explored to enhance model performance. Techniques such as automated machine learning (AutoML) could be employed to systematically search for the best model configurations. Additionally, feature selection methods, including genetic algorithms and recursive feature elimination, could be utilized to identify the most predictive features, potentially improving model accuracy and reducing complexity.

Finally, integrating these models into real-world clinical decision support systems presents both opportunities and challenges. Future work should focus on developing and evaluating the implementation of these models in clinical environments, assessing their impact on diagnostic accuracy, workflow efficiency, and patient outcomes. Pilot studies and clinical trials will be essential to validate the practical utility of these models and address any potential barriers to adoption.

In summary, while this study provides a solid foundation for the use of machine learning in heart disease detection, ongoing research and development are crucial to fully realize the potential of these technologies. By addressing the outlined areas for future work, we can move closer to integrating advanced machine learning models into routine clinical practice, ultimately improving patient care and outcomes.

Chapter 10

References

- UCI Machine Learning Repository: Heart Disease Data Set. Available at: `https://archive.ics.uci.edu/ml/datasets/heart+Disease`
- Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J.J., Sandhu, S., Guppy, K.H., Lee, S., and Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. American Journal of Cardiology, 64, 304-310.
- Soni, J., Ansari, U., Sharma, D., and Soni, S. (2011). Predictive data mining for medical diagnosis: An overview of heart disease prediction. International Journal of Computer Applications, 17(8), 43-48.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.