Multidimensional Assessment of Heart Disease Risk: Associations Between Changes in Physiological Parameters and Morbidity

Jiaxin Dong *

University of Massachusetts Amherst, Amherst, Massachusetts, 01003, The United States

* Corresponding Author Email: jiaxindong@umass.edu

Abstract. Heart disease is one of the main reasons for death worldwide, with coronary heart disease, hypertensive heart disease, and cardiac arrhythmias being prevalent. Modern lifestyle changes (e.g., unhealthy diets and irregular work schedules) have increased the frequency of heart disease and impacted people of all ages, so it is important to accurately analyze its risk factors. However, there is still a research gap in assessing multiple factors' effects on heart disease through different models. In this paper, LDA (Linear Discriminant Analysis) is used as the core research method to model and analyze multiple health indicators (e.g., blood pressure, sex, age, heart rate, etc.). Compared with different models, like Logistic Regression, and QDA, in accuracy, sensitivity, and specificity, to get the best model. The study shows that LDA has high accuracy in classifying and identifying high-risk groups. Although it also cannot classify every single variable, after removing irrelevant variables, it also shows the best accuracy of the dataset. Therefore, the conclusion states that the LDA model performs well in heart disease risk assessment and can provide strong support for early analysis and accurate assessment.

Keywords: Heart disease, prediction, model compare.

1. Introduction

In Dalen's article The Epidemic of the 20th Century: Coronary Heart Disease, it said that heart disease once a rare disease, began to decline after an outbreak and has maintained steady numbers to date [1].

In the wake of the new coronary epidemic, global public health concerns have risen significantly, with cardiovascular health in particular provoking increased concern. Cardiovascular diseases, particularly coronary heart disease (CHD), continue to be one of the main reasons for death and disability worldwide. According to the World Health Organization, cardiovascular diseases are causing more than 17 million deaths annually, accounting for 31% of all deaths globally. The incidence of cardiovascular diseases continues to rise with the aging of the global population and lifestyle changes. This phenomenon is particularly pronounced in the wake of epidemics, which expose patients with chronic diseases to a higher risk of death and serious illness, further emphasizing the need for prevention and early intervention.

As one of the major health problems worldwide, the potential dangers of heart disease cannot be ignored. Many patients do not have obvious symptoms before a heart attack, resulting in a more serious condition by the time it is detected. Therefore, it is important to establish accurate predictive models for heart disease. Predictive modeling can help doctors identify problems in the early stages of the disease and take timely treatment measures to avoid worsening the condition. Through these models, more patients will be able to detect changes in their heart health in advance during routine medical checkups, allowing for early diagnosis and treatment. Early detection reduces the risk of disease progression and provides more time for patients and physicians to make aggressive interventions.

Risk factors for coronary heart disease: implications of gender written by Lennep can be concluded that the effect of gender on heart disease exists [2]. Traditionally, statistical models such as logistic regression (LR) have been used for risk assessment of cardiovascular diseases, but in recent years, with the expansion of datasets and advances in computational technology, more and more studies

have begun to explore emerging methods such as machine learning and LDA to improve the accuracy and effectiveness of prediction.

Based on the public's concern for cardiovascular health after the epidemic, this study aimed to predict the risk of heart disease incidence using LDA models and to analyze in depth the relative importance of each risk factor for disease prediction by combining multiple variables (e.g., age, gender, type of chest pain, etc.). This not only effectively improves the accuracy of the prediction model so that it more accurately reflects the disease development situation. It also provides reliable data support and a scientific basis for the formulation and improvement of public health policies, thus better-promoting disease prevention, control, and health management.

This study has important practical implications by using the LDA model to analyze and predict the risk of developing heart disease. In addition, this study combines several variables, such as age, gender, type of chest pain, etc.

This paper aims to analyze and predict whether the probability of developing heart disease can be predicted by changes in different indicators of different causative factors so that patients can be detected and treated promptly.

2. Methodology

2.1. Introduce Dataset

This study used data from Kaggle, and there are 13 variables in the dataset to see what factors contribute to heart disease [3]. These variables include age, sex, chest pain type, resting blood pressure, serum cholestoral in mg/dl, fasting blood sugar > 120 mg/dl, resting electrocardiographic results (values 0,1,2), maximum heart rate achieved, exercise-induced angina, oldpeak = ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels (0-3) colored by flourosopy, thal: 0 = normal; 1 = fixed defect; 2 = reversable defect.

Data cleaning: the dataset was checked for missing values and outliers and appropriate processing strategies were adopted for the missing data, such as filling in using the median or the mean, to guarantee the comprehensiveness of the data and the accuracy of the analyzed results.

Type conversion: to make the data more suitable for model training, certain qualitative indicators, such as gender, type of chest pain, and resting ECG results, were digitally coded and converted to numerical data to facilitate model calculation and analysis.

Variable selection: based on the a priori knowledge and the results of exploratory data analysis, the most influential variables for heart disease prediction were selected. For example, the number of major blood vessels and maximum heart rate were highly correlated with the risk of heart disease and were therefore included in the final model.

Outcome variable processing: the representation of heart disease status was standardized, and records with 1 for no heart disease and 2 for heart disease in the raw data were converted to a binary format of 0 and 1, where 0 represents no heart disease and 1 represents heart disease. This step was done to adapt most machine learning algorithms for easy training and prediction of binary classification problems.

Data standardization: to eliminate the possible effects of variables of different magnitudes, all numerical data were standardized to ensure the fairness of the weights of the variables during model training.

Through these preprocessing steps, the dataset is organized into a clear, standardized format suitable for building and training heart disease prediction models. These treatments not only enhanced the predictive power of the models, but also ensured the reliability and validity of the study results. In addition, the in-depth understanding and preparation of the data provided a solid foundation for revealing the interactions and effects of different variables, which is crucial for an in-depth exploration of the pathogenesis and risk factors of heart disease.

2.2. Purpose of the Study

The main objective of this study is to analyze the major risk factors for heart disease through the LDA model and to assess the accuracy of the model in predicting the probability of heart disease. Specific objectives include:

- To identify the important risk factors affecting the development of heart disease.
- Demonstrating the classification effect of the model through visualization and explaining the contribution of each variable to the classification of the model.
- Evaluating the predictive performance of the LDA model and suggesting possible improvements.

2.3. Model Selection Fundamentals

The relationship between blood pressure and coronary heart disease risk changes with aging [4]. The Framingham Heart Study, written by Franklin showed a set of different classification algorithms, this algorithm demonstrates multiple classification algorithms and LDAs and presents an innovative hybrid feature selection method for diagnosing coronary heart disease (CHD).

On this basis, the data results in this dataset are again binary, it considers using LR (Logistic Regression) and LDA as two models to predict the presence or absence of heart disease. Each of the two models has its own advantages and application scenarios. The following are their basic principles and application considerations:

LR:

LR is a widely used binary model that predicts the probability of an outcome variable, specifically the probability of an event occurring. The model assumes that the dependent variables follow a binomial distribution and estimates the model parameters by maximizing the likelihood function of the observations.

LDA:

LDA is a classical linear classification method that aims to find a linear combination of features that maximizes the between-class variance over the within-class variance. In the case of dichotomous classification, LDA determines a decision boundary by calculating the mean difference and scatter of the two classes.

2.4. Model Selection Decision

It performed a dual parallel model of LR and LDA modeling for prediction. The LDA model was constructed to classify the risk of heart disease as a response variable using multiple predictor variables. Finally, visualization techniques to show the classification effect of the model. In Figure 1, each graph shows whether these different factors have any effect on heart disease.

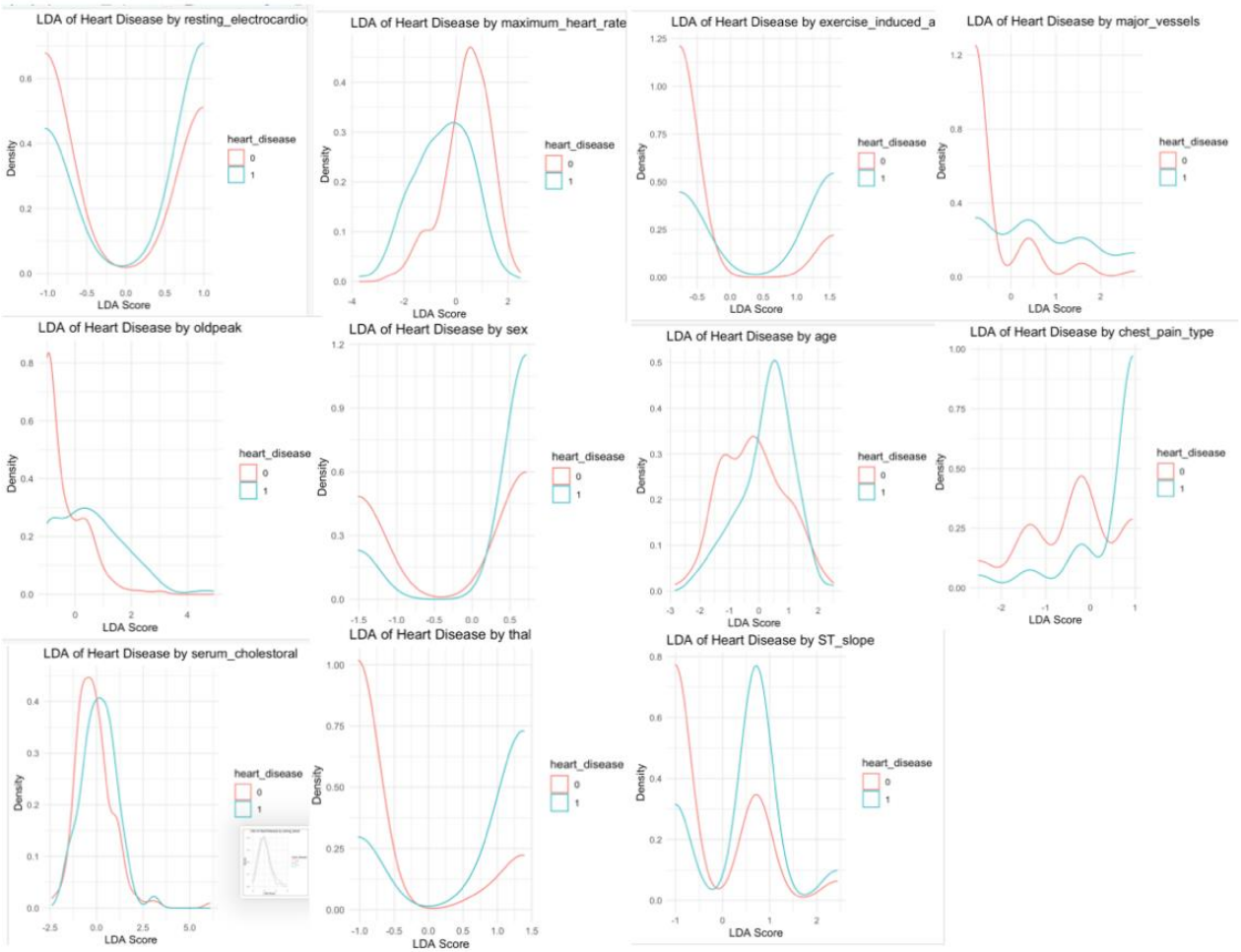


Figure 1. LDA model analysis of significant differences in variables (Photo/Picture credit: Original).

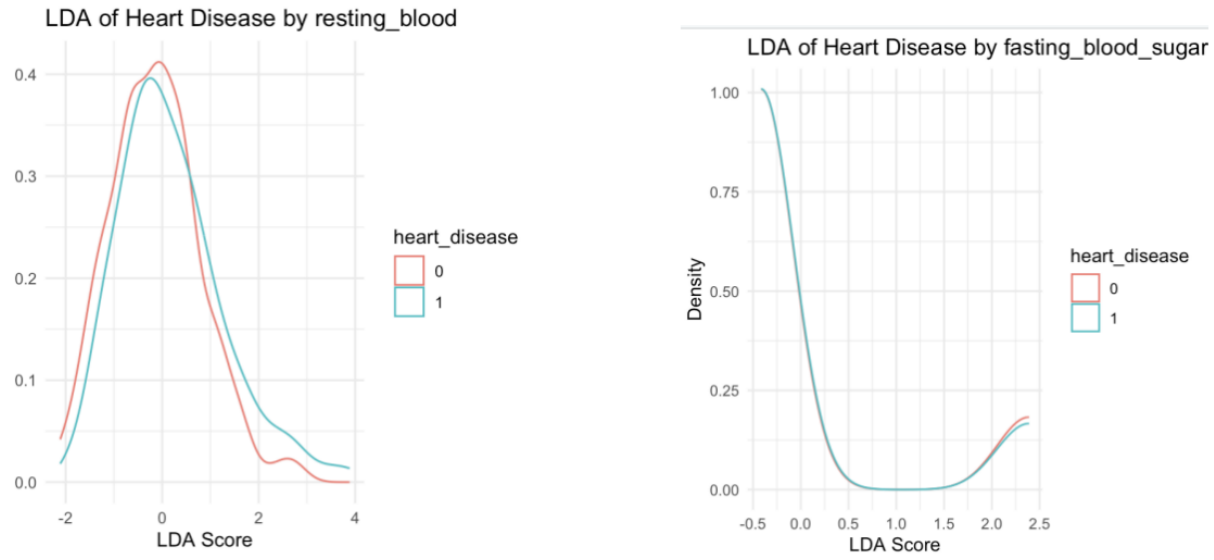


Figure 2. LDA model analysis of significant similarity in variables (Photo/Picture credit: Original).

Although LDA exhibits some data overlap on the graph, like the two variables in Figure 2, the red line, and the green line are overlapped, it is possibly because the data are not perfectly linearly divisible, the LDA model has a higher prediction accuracy than LR, this may indicate that LDA performs better in dealing with the intrinsic structure of the data.

3. Result

The performance of the four models, LR, Naïve Bayes, LDA, and QDA was evaluated using accuracy, sensitivity, and specificity metrics as shown in Table 1, LDA had the best overall performance with the highest accuracy (0.8852), sensitivity (0.9333), and specificity (0.8250) suggesting that it is very effective for this dataset. In contrast, QDA had the lowest accuracy (0.7407) and specificity (0.6250), suggesting that it may not be suitable for this dataset. LR was slightly better than Naive Bayes in terms of sensitivity and specificity, but both models performed similarly overall. Table 1 clearly show the accuracy, sensitivity, and specificity of different models.

Table 1. Comparison of the performance of different models

Model	Accuracy	Sensitivity	Specificity
Logistic	0.7963	0.8667	0.7083
Naive bayes	0.7778	0.8333	0.7083
LDA	0.8852	0.9333	0.8250
QDA	0.7407	0.8333	0.6250

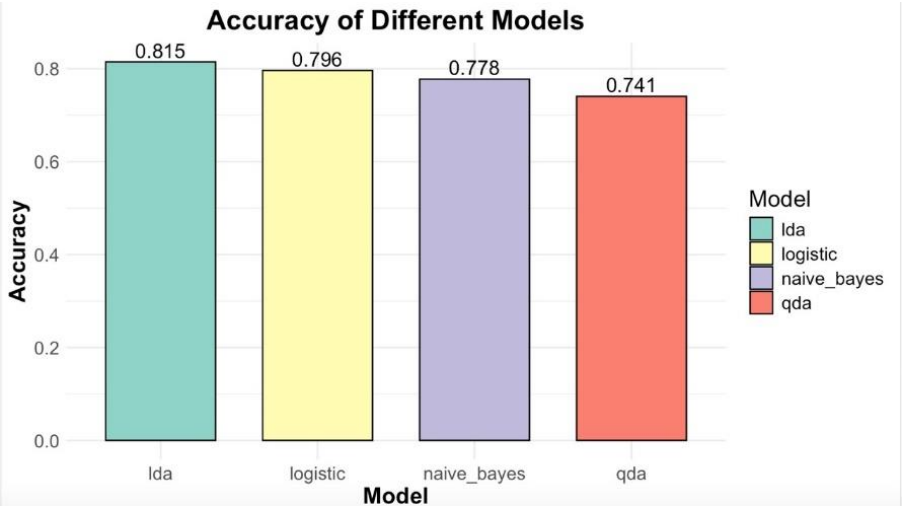


Figure 3. Accuracy of different models (Photo/Picture credit: Original).

It is easy to find that the accuracy of the LDA model is higher than others from Figure 3.

3.1. Possible Cause Analysis

Quadratic discriminant analysis (QDA) is a univariate statistical method that is used for constructing a model based on groups that consider the observed effective factors [5]. Unlike LDA, QDA finds a group membership consisting of a square n × n matrix (where n is the number of variables) and a linear combination of these variables [6]. Therefore, the likely reason for LDA's superior performance is that the dataset may not exhibit the complex decision boundaries that QDA assumes. The linear decision boundaries used by LDA are simpler and may generalize better for this dataset. In contrast to LDA, QDA can deal with different covariance values of the classes, and this can be considered as one of its advantages over LDA [5]. Although QDA is more flexible and useful in more complex data distributions, it may lead to overfitting in this case, especially given the relatively small sample size and limited feature variance.

In summary, LDA's simplicity and stability make it a better choice for this dataset, while QDA's added complexity seems to hinder its performance.

3.2. Strengths and Limitations

The strengths of LDA are its simplicity, efficiency, and stability, making it well-suited to small datasets.

LDA uses linear decision boundaries, the model is relatively simple, easy to understand and interpret, and is particularly suitable for dealing with linearly divisible data. Also, LDA shows good stability when dealing with small datasets, especially when the covariance matrices between categories are similar, LDA can generate stable classification results. LDA is simple, robust, and has a high prediction accuracy [5].

Larger and more diverse datasets could improve the robustness of the model.

Another limitation is the weak effect of certain predictors such as fasting glucose and serum cholesterol. Future improvements could focus on directions such as feature selection.

4. Conclusion

The research methods of this article include the use of a dataset from Kaggle, which contains 13 variables related to heart disease. The study ensured the completeness and accuracy of the data through steps such as data cleaning, type conversion, and variable selection. The outcome variables were standardized into binary format to adapt to machine learning algorithms. The main purpose of the study was to identify the main risk factors for heart disease through the linear discriminant analysis (LDA) model and evaluate the predictive accuracy of the model. Finally, the logistic regression (LR) and LDA models were compared, and the classification effect was displayed using visualization techniques to reveal the risk factors for heart disease and their impact.

From the results of the different models, it is easy to find that both accuracy (0.8852), sensitivity (0.9333), and specificity (0.8250) in the LDA model are higher than others. Expanding the dataset and incorporating more medical indicators (e.g., genetic or lifestyle factors) could improve model performance. In addition, exploring other machine learning methods (e.g., random forests) alongside the use of LDA could provide deeper insights into heart disease prediction.

References

[1] Dalen JE, Alpert JS, Goldberg RJ, Weinstein RS. The epidemic of the 20th century: coronary heart disease. The American journal of medicine. 2014, 127 (9):807 – 12.

[2] Roeters van Lennep JE, Westerveld HT, Erkelens DW, van der Wall EE. Risk factors for coronary heart disease: implications of gender. Cardiovascular research. 2002, 53 (3): 538 - 49.

[3] UCI Machine Learning. 2019. https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset.

[4] Kolukisa B, Hacilar H, Goy G, Kus M, Bakir-Gungor B, Aral A, Gungor VC. Evaluation of classification algorithms, linear discriminant analysis, and a new hybrid feature selection methodology for the diagnosis of coronary artery disease. In2018 ieee international conference on big data (big data), 2018, 2232 - 2238.

[5] Franklin SS, Larson MG, Khan SA, Wong ND, Leip EP, Kannel WB, Levy D. Does the relation of blood pressure to coronary heart disease risk change with aging? The Framingham Heart Study. Circulation. 2001, 103 (9): 1245 - 9.

[6] Hong H, Naghibi SA, Moradi Dashtpagerdi M, Pourghasemi HR, Chen W. A comparative assessment between linear and quadratic discriminant analyses (LDA-QDA) with frequency ratio and weights-of-evidence models for forest fire susceptibility mapping in China. Arabian Journal of Geosciences. 2017, 10: 1 - 4.