

Performance Analysis and Comparison of Heart Disease Prediction Models

Jiayi Jin¹, Chengyun Zhao^{2, *}

¹ College of Biological Sciences, University of California, Davis, CA 95616, USA

² School of Physical Sciences, University of California, Irvine, CA 92697-4675, USA

* Corresponding Author Email: zhaoc21@uci.edu

Abstract. Since cardiovascular diseases (CVDs) are the leading cause of death, it is significant for people to detect them early and take certain precautions. As a result, the paper performs the study of heart disease prediction with statistical models. In this study, the researcher analyzed a dataset from four different regions, including Cleveland, Hungarian, Switzerland, and Long Beach. Each of the regions is used as one testing dataset and the rest of the three regions are used as one training dataset, with a total of four sets of training and testing datasets. The goal is to predict heart disease with various variables using three main statistical techniques, lasso regression, random forest, and logistic regression (with feature selection). The paper will compare aspects such as accuracy, confusion matrix, and roc-auc for the models created. The results show that the models tested on the Hungarian dataset performed best overall, while the random forest algorithm achieved the highest prediction accuracy.

Keywords: Heart Disease Prediction, Cardiovascular diseases, Lasso Regression.

1. Introduction

Cardiovascular diseases (CVDs) are the leading cause of mortality worldwide, accounting for approximately 17.9 million deaths annually. This accounts for 31% of all global deaths. With such a wide impact, people focus more on early detection and management where a statistical model can be of great help [1]. Fortunately, the development of clinical coronary heart disease has shown a solution for using programming and data science to accomplish accurate prediction. In the last century, Daniel B. Mark and his colleagues successfully used the treadmill score as a useful and valid tool to help clinicians determine prognosis and decide whether to refer outpatients with suspected coronary disease for cardiac catheterization [2]. In recent years, more researchers have focused on the new possibility of applying machine learning to heart disease prediction and have started to identify the ML-based and data-driven recent trends and techniques in heart disease diagnosis with imbalanced data [3]. Further, Ulloa-Cerna, A. E. and his colleagues have proven that clinically significant heart disease among high-risk populations can be successfully predicted by a solely ECG-based machine learning model that uses a composite endpoint [4].

By leveraging the regression model, a fundamental way under machine learning, to analyze these associations, healthcare providers can improve early diagnosis and intervention strategies, ultimately reducing the burden of CVDs.

2. Method

2.1. Data Description

The initial data set is taken from the intersection of two similar datasets on Kaggle, also on the UCI ICS website for machine learning. Excluding dummy values and regions, the researcher found that 11 of these variables were associated with the incidence of heart disease. Therefore, we set these 11 variables as x variables, which is the independent variable, and the incidence of heart disease as the dependent variable. The detailed information will be presented in 3.1. The researcher categorizes the entire dataset by region. Of those, 303 observations were from Cleveland, 270 observations from

Hungarian, 139 observations from Switzerland, and 121 observations from Long Beach VA. The Heart Failure Prediction Dataset is a combination of five smaller datasets from regions Cleveland, Hungarian, Long Beach, Switzerland, and Stalog. The dataset is found in Kaggle, also from the UC Irvine machine learning website. After comparing the two websites, the fifth dataset is excluded from the original dataset since it does not specify its region, it is only a dataset with common variables as the first four. There are a total of twelve variables in the dataset, including one dependent variable that indicates whether having heart disease or not and eleven independent variables such as age, sex, chest pain type, resting blood pressure, cholesterol level, fasting blood sugar, resting ECG, maxHR, exercise angina, old peak, and ST slope.

Cholesterol level is excluded for all the models since observations from Switzerland are all 0, and 27.3% of observations from Long Beach are 0. After converting the negative values of the old peak to zeros, the distribution of it looks right-skewed. Consequently, the log old peak variable is added and used in model construction later. After all the exclusion and manipulation, there are 833 observations and 12 variables.

2.2. Statistical Analysis

The distribution analysis of all the variables is based on region. Before training models, the dataset is divided into a training dataset with three regions and a testing dataset with the left one region, as shown by Table 1.

Table 1. Training dataset and testing dataset

Training Dataset (region)	Testing Dataset (region)	Model Name
Hungarian, Switzerland, Long beach	Cleveland	Cleveland
Cleveland, Switzerland, Long beach	Hungarian	Hungarian
Cleveland, Hungarian, Long beach	Switzerland	Switzerland
Hungarian, Switzerland, Long beach	Long beach	Long beach

Then, the models are built using the lasso regression, random forest, and logistic regression with the four training datasets respectively. Each model has the advantage of analyzing data. Lasso regression (Least Absolute Shrinkage and Selection Operator) is a linear regression technique that includes an L1 regularization term. This term adds a penalty equal to the absolute value of the coefficients. The goal is to minimize the residual sum of squares while also constraining the sum of the absolute value of the coefficients. The Lasso model has two major advantages it can shrink some coefficients to zero, effectively performing variable selection and simplifying the model, and be less prone to overfitting, especially in high-dimensional datasets. For lasso regression, the tuning parameter is selected based on cross-validation results.

Random forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (for classification) or the mean prediction (for regression) of the individual trees. It utilizes bagging (bootstrap aggregating) and random feature selection. The random forest model is good at handling overfitting, managing missing data without requiring imputation, and providing insights into feature importance (helping identify the most influential variable). For the random forest, the number of trees is 100.

Logistic regression is used for binary classification problems. It models the probability that a given input point belongs to a certain class by using the logistic function to constrain the output between 0 and 1. The model estimates the relationship between the dependent binary variable and one or more independent variables. Through logistic regression, coefficients can be easily interpreted as log odds, making it straightforward to understand the influence of predictors. In addition, it provides probabilistic outputs, allowing for better decision-making based on the likelihood of outcomes. Logistic regression is computationally efficient and works well with smaller datasets or when the relationship between the features and the target variable is approximately linear. For logistic regression, all variables are included at first.

After feature selection based on the lasso, and random forest model selected variables are included in the models. For each method, the best model is selected based on accuracy, confusion matrix, and ROC-AUC. After choosing the best model for each method, the best method will be selected based on the same aspects mentioned previously.

3. Result

3.1. Lasso Regression

To facilitate quantifiable predictions, some categorical data are converted to numerical data, including: Sex: M->1, F->0; Exercise Angina: Y->1, N->0; ChestPainType: ASY->0, NAP->1, ATA->2, TA->3; RestingECG: Normal->1, LVH->2, ST->3; ST_Slope: Up->0, Flat->1, Down->2. Based on the feature importance of the four models, Sex, FastingBS, ExerciseAngina, logOld peak, ChestPainType, and ST_Slope were positively or negatively highly correlated with heart disease. According to the results of model testing, the Hungarian model has the highest prediction accuracy (0.86). In the confusion matrix, Hungarian had the highest kappa (0.86), sensitivity (0.86), and specificity (0.87) when compared with the other three models (Figure 3). The Hungarian model also performs best in the AUC-ROC chart, being closest to the top left corner of the chart and having a larger AUC area (0.92) than the other three models (Figure 4).

3.2. Random Forest Classification

The Feature Importance plot below shows that variable MaxHR has the highest importance as its MeanDecreaseGini is the largest (Figure 1). From the Model Accuracy by Region table below of 4 different models, the Hungarian model has the highest accuracy with a value of 0.893(Figure 2). Based on the Confusion Matrices by Region below of each of the 4 different models, the Hungarian model also performs the best. The model correctly predicted 93 cases of heart disease (1) when the actual class was heart disease (1). The model correctly predicted 148 cases of no disease (0) when the actual class was no disease (0). The sensitivity is 0.8757, and the specificity is 0.9208, which is the best of the 4 models, indicating its strong ability to correctly predict patients with heart disease (Figure 3). The Kappa value is 0.7764, which is the highest compared to the other 3 models (Figure 3). According to the ROC Curves for Random Forest by Region plot below, the ROC of the Hungarian model is also the best of these 4 models indicating strong predictive performance, which is the nearest to the top left corner signifying a high true positive rate and a low false positive rate with the highest AUC value of 0.948 while the AUC of each of the other 3 models is only 0.8.

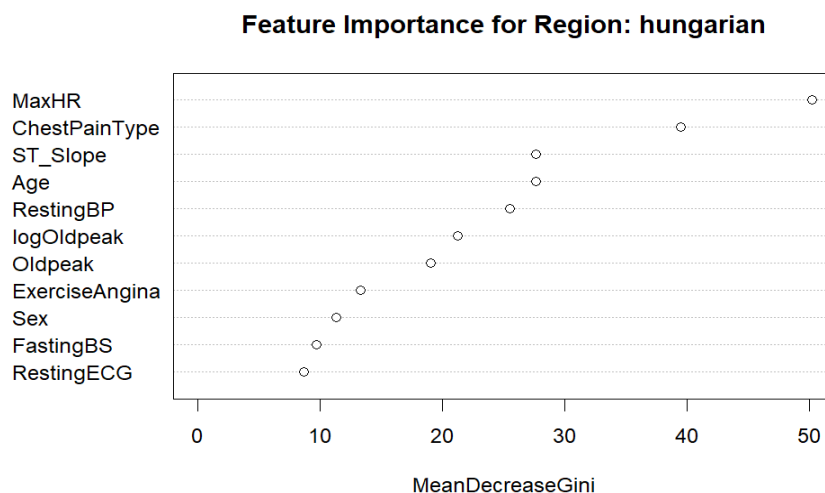


Figure 1. Feature importance for region: Hungarian (Photo/Picture credit: Original).

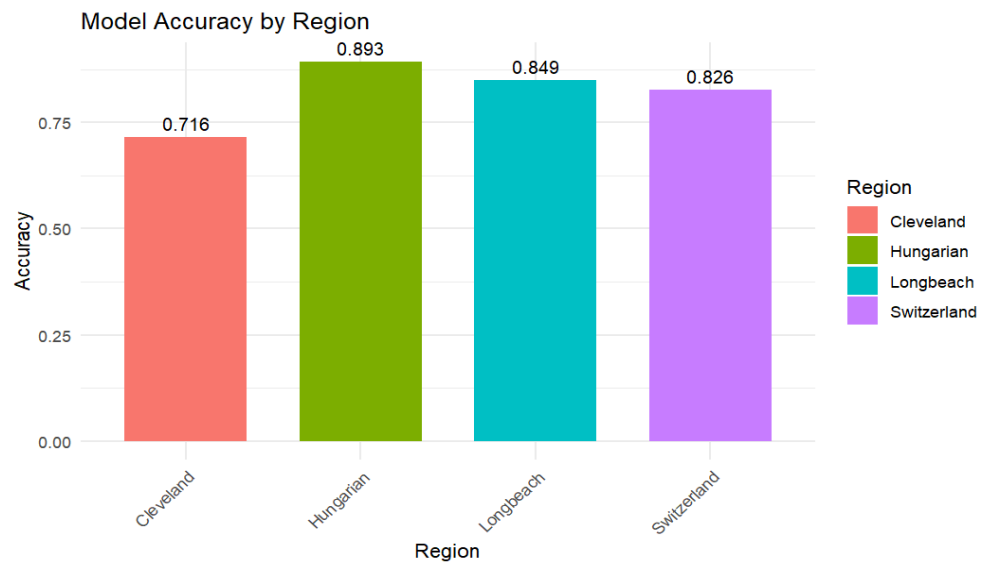


Figure 2. Model Accuracy for RF by Region (Photo/Picture credit: Original).

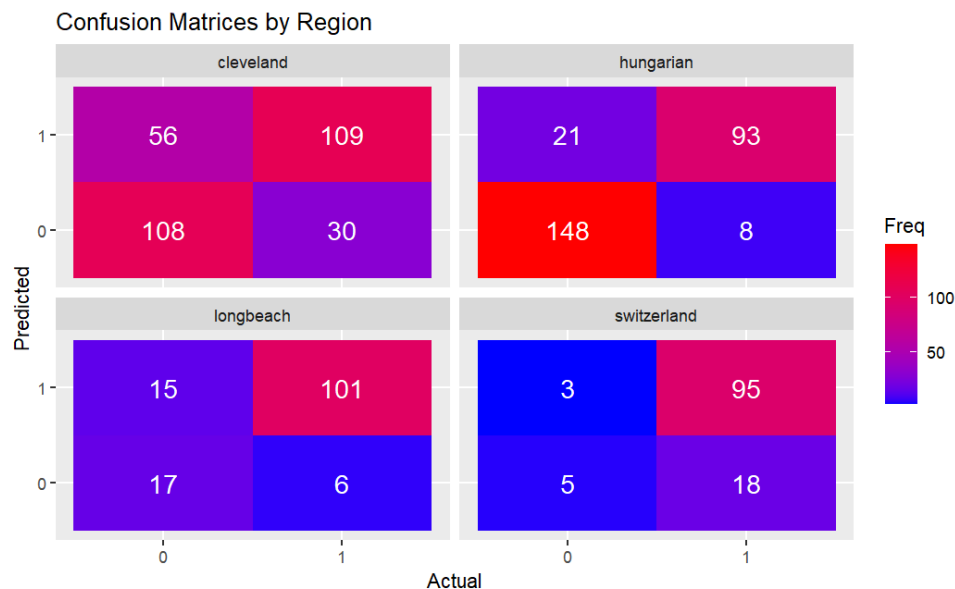


Figure 3. Confusion Matrices for RF by Region (Photo/Picture credit: Original).

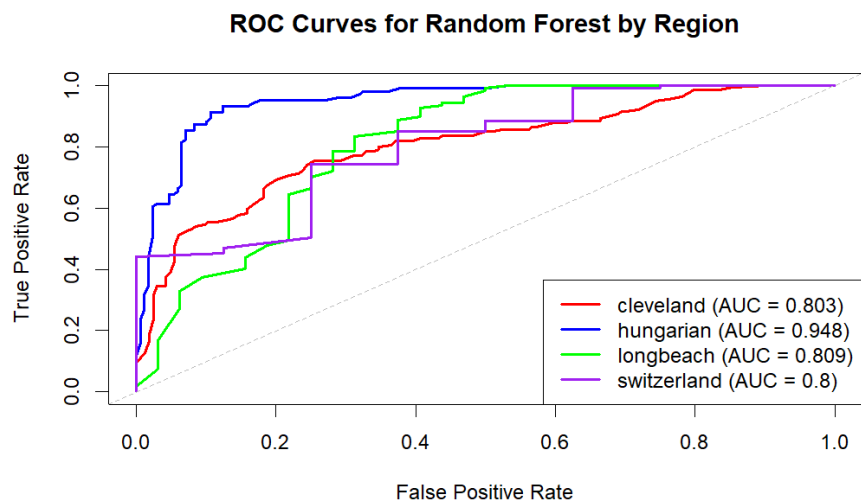


Figure 4. ROC Curves for RF by Region (Photo/Picture credit: Original).

3.3. Logistic Regression

Firstly, all the variables introduced in 3.1 are used for the four models. Based on accuracy, Hungarian is the best model. In the confusion matrix, Hungarian has the highest sensitivity, and Long Beach has the highest specificity. However, according to Figure 3, Hungarian (0.73) performed much better than Long Beach (0.47). Based on ROC and AUC in Figure 4, Hungarian is still the best among the four models. Then, some feature selection was applied based on the lasso regression results. In the Cleveland model, resting BP and resting ECG were excluded. In the Hungarian model, resting ECG was excluded. The other two models remain the same. After comparing these six models, both Hungarian models perform the best. Nevertheless, considering the model's simplicity, the best model in logistic regression is the updated Hungarian model with selected features.

3.4. Overall Comparison

All the methods mentioned above turn out that Hungarian is the best model. Random forest is the best while evaluating the accuracy of the three methods. When comparing the confusion matrix, all three methods have the same specificity (0.87) (Figure 3). Random forest has the highest sensitivity (0.92), while lasso has the highest kappa (0.87) (Figure 3). As a result, the random forest and lasso models are better than logistic regression. When comparing the ROC and AUC, random forest and logistic regression models (0.95) are better than lasso. Thus, based on all three aspects, the random forest with Hungarian as a testing dataset is the best model (Figure 4).

4. Discussion

4.1. Model Performance Analysis

In this study, the Random Forest model significantly outperformed other models, achieving the highest accuracy and AUC scores, indicating its superior adaptability in predicting heart disease. Besides the superiority of RF over the logistic model in this research, RF also shows notably higher sensitivity 91.7% and specificity rates 91.7% than the logistic model in the Wang et al research for recognizing macrosomia. Both results align with findings drawing from Breiman's concept of "bagging" and the random selection of features, Random Forest (RF) can achieve greater accuracy. Cutler has noted that, in comparison to other classifiers, RF offers several advantages: (1) exceptional classification accuracy; (2) the capability to assess variable importance; (3) versatility in handling various types of data analysis; and (4) the ability to model intricate interactions among explanatory variables [5]. However, despite its strong performance, the Random Forest model exhibited certain limitations, particularly in handling imbalanced data, where some regions still had slightly lower sensitivity, possibly due to data heterogeneity across regions.

For the Logistic Regression model, while it also delivered a strong AUC score of 0.95, its performance lagged Random Forest in accuracy and feature interaction capabilities. This finding is consistent with's research, which indicates that while Logistic Regression performs well for simpler datasets, it struggles with more complex relationships.

4.2. Improvement Suggestions

To further enhance the Random Forest model, introducing hyperparameter tuning or additional ensemble techniques like boosting could potentially improve accuracy. Introduced by Jerome H. Friedman from Stanford University in 1999, the boosting has since been enhanced and refined through multiple open-source libraries, including Scikit-Learn and R [6]. Google Scholar indicates that Friedman's two papers on gradient boosting have received around 10,000 citations, and their applications can be found across nearly all scientific disciplines [6]. Zhang et al showed that incorporating gradient boosting increases the precision in similar classification tasks that accompany complicated features and unsure relationships between features [7]. Likewise, integrating cross-

validation or oversampling methods may mitigate the imbalance issue and strengthen sensitivity across all regions.

4.3. Study Limitations

One limitation of this study is the lack of using the variable Cholesterol in the dataset. All the values of this variable in Switzerland are 0, and 27.3% (38/139) values of Long Beach are 0. So, this variable is not included in all models. Rouzrokh pointed out that the bias may hinder the model's ability even if the model has been trained relatively perfectly before [8]. The study didn't explore potential interactions between cholesterol and other features.

4.4. Future Research Directions

It is often difficult for a model to be broadly applicable to all data by training it on just one data source according to Doshi-Velez [9]. Future research should consider expanding the dataset to include more varied regions and demographics to enhance model generalization. Moreover, incorporating deep learning techniques could capture more complex relationships in the data, as supported by recent advancements in cardiovascular disease prediction studies. In Yang's research, traditional regression, which relies on a simplified relationship between predictor variables and outcomes, acted far worse than the validated precise deep learning that they created aimed at predicting mortality rate [10]. Further investigations into model interpretability would also benefit medical professionals by providing clearer insights into feature impacts and decision-making processes.

5. Conclusion

This study compared the performance of several machine learning models in predicting heart disease, with Random Forest achieving the highest accuracy (89%) and AUC (0.95), outperforming Logistic Regression (Accuracy: 87%, AUC: 0.95) and Lasso Regression (Accuracy: 86%, AUC: 0.92). The results demonstrate the strengths of ensemble methods, particularly in handling complex data and improving predictive power.

To further enhance model accuracy and generalizability, several improvements can be made. Adding features such as cholesterol levels, which were excluded in this analysis, could improve prediction accuracy. Including data from more regions and diverse populations would allow the models to generalize more effectively. Additionally, utilizing more recent datasets would ensure the models reflect current health trends and conditions. Exploring other machine learning techniques like Support Vector Machines (SVM) or deep learning approaches could provide alternative perspectives and potentially improve performance, depending on the data's complexity. Overall, the findings suggest that while Random Forest performed best in this analysis, future efforts to expand the dataset and explore other algorithms could yield further improvements.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

- [1] Fetal origins of coronary heart disease. (n.d.). The BMJ. Retrieved September 14, 2024, from <https://www.bmj.com/content/early/recent>.
- [2] Mark, D. B., Shaw, L., Harrell, F. E., Hlatky, M. A., Lee, K. L., Bengtson, J. R., McCants, C. B., Califf, R. M., & Pryor, D. B. Prognostic value of a treadmill exercise score in outpatients with suspected coronary artery disease. *New England Journal of Medicine*, 1991, 325 (12), 849 – 853.
- [3] Ahsan, M. M., & Siddique, Z. Machine learning-based heart disease diagnosis: A systematic literature review. *Artificial Intelligence in Medicine*, 2022, 128, 102289.

- [4] Ulloa-Cerna, A. E., Jing, L., Pfeifer, J. M., Raghunath, S., Ruhl, J. A., Rocha, D. B., Leader, J. B., Zimmerman, N., Lee, G., Steinhubl, S. R., Good, C. W., Haggerty, C. M., & Fornwalt, B. K. Recommend: An ECG-based machine learning approach for identifying patients at increased risk of undiagnosed structural heart disease detectable by echocardiography. *Circulation*, 2022, 146 (1), 36 – 47.
- [5] Yaghoobi V, Martinez-Morilla S, Liu Y, Charette L, Rimm DL, Harigopal M. Advances in quantitative immunohistochemistry and their contribution to breast cancer. *Expert Rev Mol Diagn*. 2020, 20 (5): 509 – 22.
- [6] Wang F, Wang Y, Ji X, Wang Z. Effective macrosomia prediction using random forest algorithm. *Int J Environ Res Public Health*. 2022.
- [7] Zhang, Z., Zhao, Y., Canes, A., Steinberg, D., & Lyashevskaya, O. Predictive analytics with gradient boosting in clinical medicine. *Big-data Clinical Trial Column*, 2019, 7 (7).
- [8] Rouzrokh, P., Khosravi, B., Faghani, S., Moassefi, M., Vera Garcia, D. V., Singh, Y., Zhang, K., Conte, G. M., & Erickson, B. J. Mitigating bias in radiology machine learning: 1. Data handling. *Radiology: Artificial Intelligence*, 2018.
- [9] Doshi-Velez, F., & Kim, B. Considerations for evaluation and generalization in interpretable machine learning. In H. J. Escalante, S. Escalera, I. Guyon, et al. (Eds.), *Explainable and interpretable models in computer vision and machine learning* (pp. 3 – 17). Cham, Switzerland: Springer, 2018.
- [10] Yang, R., Huang, T., Wang, Z., Huang, W., Feng, A., Li, L., & Lyu, J. Deep-learning-based survival prediction of patients in coronary care units. *Journal Name*, Article 5745304, 2021.