Journal of Computers, Mechanical and Management
Journal homepage: www.jcmm.co.in
e-ISSN: 3009-075X
AN PUBLISHING

Volume 3 Issue 5
Article Number: 24123

Comparative Analysis of Random Forest and Logistic Regression for Heart Attack Risk Prediction

Nilakshman Sooriyaperakasam* [1], Hamid Emami[2], Parinaz Entezam[2], and Chisom Ezekiel[2]

[1]Department of Mechanical Engineering, University of Moratuwa, Colombo, 10400 Sri Lanka
[2]Department of Biomedical Engineering, University of Oulu, 90014 Finland

Abstract

Cardiovascular diseases, particularly heart attacks, are leading causes of global mortality, highlighting the need for enhanced early detection and intervention strategies. This study evaluates the effectiveness of two machine learning algorithmsRandom Forest (RF) and Logistic Regression (LR)in predicting heart attack risk using diverse patient data sets. The focus is on uncovering subtle patterns and risk factors that traditional methods may overlook, while also assessing the accuracy and performance of both models. A critical aspect of the study is the interpretability of these algorithms, addressing a significant gap in current research. Additionally, the issue of dataset imbalance, which is prevalent in medical data, is examined, and solutions are proposed to improve model reliability in real-world applications. These findings contribute to the discourse on optimizing machine learning in healthcare, advocating for tailored approaches that balance predictive power with interpretability. By analyzing the strengths and weaknesses of RF and LR in heart attack prediction, this study aims to provide valuable insights for clinicians and researchers, ultimately enhancing decision-making processes in cardiovascular care and interventions.

Keywords: Machine Learning; Heart Attack Prediction; Random Forest; Logistic Regression; Interpretability

1. Introduction

Cardiovascular diseases, including heart attacks, represent a major public health concern throughout the world, contributing to substantial morbidity and mortality [1]. Despite advances in medical science, early detection and prediction of heart attacks remain a challenge, with delays often leading to life-threatening complications [2]. Detecting individuals at risk early could significantly improve outcomes by enabling prompt intervention [3, 4]. In this context, machine learning techniques offer a promising avenue to enhance predictive accuracy in heart attack detection [5]. These models can process large, complex datasets to uncover subtle patterns that traditional statistical methods might overlook [6]. Several studies have applied machine learning to heart disease prediction, exploring various algorithms from Random Forest (RF) to Logistic Regression (LR) models [7]. However, this study distinguishes itself by providing a direct comparative analysis of the RF and LR models, focusing on interpretability and performance evaluation in real world contexts. Furthermore, unlike previous studies that often overlook the implications of unbalanced datasets or do not address the interpretability of the model in a comprehensive way [8], the present research underscores these challenges, offering novel insights into optimizing predictive performance.

*Corresponding Author: Nilakshman Sooriyaperakasam (nilakshman.sooriyaperakasam@student.oulu.fi)
Received: 19 May 2024; Revised: 25 Oct 2024; Accepted: 30 Nov 2024; Published: 30 Nov 2024

DOI: 10.57159/jcmm.3.5.24123.

2. Related Works

This section provides an overview of the key findings and methodologies employed in previous research related to the application of machine learning models for the prediction of heart attacks. Ruby Hasan (2021) [9] conducted a comparative analysis of various machine learning algorithms to predict heart disease using patient medical data. The study evaluated models including K-Nearest Neighbors (KNN), Decision Trees, Gaussian Naïve Bayes, Logistic Regression, and Random Forest. Each algorithm was assessed based on its ability to predict heart disease effectively, and the analysis highlighted the strengths and limitations of each approach. The findings demonstrated the potential of machine learning tools to enable early detection of heart disease, thus facilitating timely interventions. This work underscores the importance of selecting appropriate algorithms and optimizing models for medical data to improve predictive accuracy and reliability. Abdellati et al. (2022) [10] introduced a supervised infinite feature selection method combined with an improved weighted random forest algorithm to address class imbalance and high-dimensional data challenges in heart disease detection. This approach demonstrated superior performance compared to traditional methods, highlighting the potential of advanced feature selection and ensemble learning techniques in medical diagnostics. Pathan et al. (2022) [11] focused on the application of machine learning models for the prediction of heart disease, highlighting the importance of data pre-processing and model optimization. The study underscored that integrating domain knowledge with machine learning approaches can significantly improve predictive outcomes, thus aiding in early diagnosis and treatment planning. Zhaozhao Xu et al. (2020) address the challenges of imbalanced medical data classification with a hybrid sampling algorithm called RFMSE, which integrates Misclassification-Oriented SMOTE (M-SMOTE) and Edited Nearest Neighbor (ENN) techniques with Random Forest (RF). Unlike traditional methods that solely rely on static imbalance rates, RFMSE dynamically updates sampling rates based on RF's misclassification rates. The hybrid approach enhances minority class recognition by combining over-sampling to generate synthetic samples and under-sampling to remove noise from majority class data. Experimental results on ten UCI datasets demonstrate RFMSE's superiority over classical algorithms, achieving notable improvements in F-value and Matthews Correlation Coefficient (MCC). This method also effectively balances time complexity and classification performance, making it highly suitable for medical applications like missed abortion prediction [6]. Lingyu Li and Zhi-Ping Liu (2022) [12] proposed a connected network-regularized logistic regression (CNet-RLR) model for feature selection in high-dimensional datasets, particularly focusing on genomic data. The model uniquely integrates Lasso penalties for sparsity, graph Laplacians for smoothness, and network connectivity constraints to ensure features form connected subnetworks. The model achieves efficient and interpretable feature selection by reformulating the Lasso penalty into a convex optimization problem and employing an interior-point algorithm. Empirical validation using synthetic and real-world cancer genomics datasets demonstrated superior classification accuracy and feature interpretability compared to existing models like Lasso-RLR and Elastic Net. This work highlights the importance of embedding domain-specific network structures into feature selection algorithms to enhance their practical utility. These recent advancements in heart attack prediction research have emphasized the integration of RF and LR models with complementary techniques such as deep learning, ensemble learning, and feature engineering. At the same time, its also noted that, efforts have been directed towards addressing real-world challenges such as data imbalance, model interpretability, and scalability.

3. Methods

3.1. ML models

The selection of machine learning models in this study was guided by the dataset's characteristics, including its size, feature space, and complexity of relationships among variables. Random Forest (RF) and Logistic Regression (LR) were chosen as the primary models for evaluation due to their complementary strengths in predictive modeling and interoperability. The random forest is an ensemble learning method that creates multiple decision trees during training and predicts the most common class. It effectively captures complex, non-linear relationships in data, making it ideal for medical datasets. Furthermore, it is resistant to over-fitting, especially with smaller datasets, as it averages predictions from various trees to reduce variance [13]. Logistic regression is an efficient statistical method for binary classification, particularly suited for large datasets. Its linear nature allows it to scale effectively with the size of the data, making it ideal for real-time predictions and high-dimensional applications. In addition, it provides clear information on feature-target relationships, which is vital for interpretability in medical contexts [14]. The combination of these two models promises to provide a comprehensive analysis, balancing the interpretability and simplicity of Logistic Regression with the robustness and flexibility of Random Forest. Thus, the dual approach ensures that the study findings are both practically relevant and grounded in methodologically sound modeling strategies.

3.2. Data collection and analysis

The dataset consisted of 303 samples with 14 features, sourced from Kaggle's "Heart Attack Analysis & Prediction Dataset" (https://www.kaggle.com/datasets/sonialikhan/heart-attack-analysis-and-prediction-dataset) as described in Table 1.

The entire model data curation, model preparation, and classification were done in Jupiter Notebook IDE. The feature selection was performed later using Pearson's correlation analysis to identify features with the strongest linear relationships to the target variable. This method was chosen for its simplicity and effectiveness in identifying predictive characteristics in datasets where linear relationships may play a significant role [15].

3.3. Data preprocessing

Data preprocessing was performed to address missing values and prepare the dataset for machine learning models. Missing values were identified but not imputed intentionally to simulate real-world scenarios where complete data may not always be available. This approach allows the models to handle incomplete datasets, reflecting practical clinical applications. Additionally, normalization was applied to continuous features to ensure that they were on comparable scales, improving model performance and stability.

3.4. Model training and validation

The dataset was split into training and testing sets, with 80% of the samples used for training and 20% reserved for testing. Both RF and LR models were trained on the preprocessed dataset, and their performance was evaluated using metrics such as accuracy, precision, recall, and F1 score. These metrics were selected to provide a comprehensive evaluation of the models, especially given the potential class imbalance in the dataset.

Table 1: Feature selection for ML training

Feature	**Description**
age	The age of the patient in years.
sex	The gender of the patient (1 = male, 0 = female).
cp	The chest pain type experienced by the patient (0 = typical angina, 1 = atypical angina, 2 = non-anginal pain, 3 = asymptomatic).
trtbps	The resting blood pressure of the patient in mm Hg.
chol	The cholesterol level of the patient in mg/dl.
fbs	The fasting blood sugar level of the patient (1 = fasting blood sugar > 120 mg/dl, 0 = otherwise).
restecg	The resting electrocardiographic results (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy).
thalachh	The maximum heart rate achieved by the patient.
exng	Exercise-induced angina (1 = yes, 0 = no).
oldpeak	The ST depression induced by exercise relative to rest.
slp	The slope of the peak exercise ST segment (1 = upsloping, 2 = flat, 3 = downsloping).
caa	The number of major vessels (0-3) colored by fluoroscopy.
thall	The thallium stress test result (1 = normal, 2 = fixed defect, 3 = reversible defect).
output	The diagnosis of heart disease (1 = presence of heart disease, 0 = absence of heart disease).

4. Results and Discussion

The implementation of Pearson correlation analysis reduced the initial 13 features to 7, focusing on those with the highest correlation coefficients, as shown in Table 2. For instance, features like `cp` (chest pain type) and `exng` (exercise-induced angina) demonstrated the highest correlations, with coefficients of 0.432 and 0.436, respectively.

The results of this study indicate that both the Random Forest (RF) and Logistic Regression (LR) models achieved comparable test accuracy rates of 86%. Despite the higher training accuracy of RF 100%, its test performance did not exceed that of LR, suggesting the possibility of overfitting. This section delves into the potential reasons for this similarity, examines additional performance metrics, and evaluates the models' real-world applicability.

Table 2: Pearson coefficient analysis for selected features.

Feature	Pearson Coefficient
cp	0.432080
thalachh	0.419955
exng	0.435601
oldpeak	0.429146
slp	0.343940
caa	0.408992
thall	0.343106

Several factors may explain the comparable performance of RF and LR models:

- **Feature Selection Overlap**: Pearson's correlation analysis identified features (`cp`, `exng`, etc.) that were highly predictive for both models. This shared feature space may have led to similar predictive outcomes.
- **Dataset Characteristics**: With only 303 samples, the limited size of the dataset may have restricted RFs ability to take advantage of its complexity advantages over LR. Furthermore, the linear relationships in the dataset between certain features and the target variable may have favored the simpler modeling approach of LR.
- **Class Balance**: Although the dataset exhibited some imbalance, both models likely handled it effectively due to their inherent robustness and the moderate size of the dataset.
- **Noise and Missing Data**: The intentional decision not to impute missing values may have influenced both models similarly, as they had to learn patterns from incomplete data.

4.1. Evaluation Using Additional Metrics

Relying solely on accuracy as a performance metric can mask a model's limitations, particularly in the context of imbalanced datasets. To provide a more comprehensive evaluation, metrics such as precision, recall, F1 score, and ROC-AUC were calculated (Table 3).

Table 3: Performance metrics for RF and LR models.

Metric	Random Forest	Logistic Regression
Precision	0.88	0.87
Recall	0.84	0.85
F1-Score	0.86	0.86
ROC-AUC	0.90	0.89

Both models demonstrated high precision and recall values, indicating their effectiveness in identifying positive cases while minimizing false positives. The F1 scores reinforce their balanced performance in precision and recall, and the ROC-AUC values highlight their strong ability to distinguish between classes. These results further support the robustness of both models in heart attack prediction.

4.2. Real-World Implications and Limitations

The findings emphasize the importance of thorough feature selection and preprocessing when working with clinical datasets. LRs simplicity and interpretability make it an attractive option for real-time applications where computational efficiency is critical. RF, with its ability to capture non-linear relationships, may excel in scenarios with larger datasets or more complex interactions. However, the observed similarity in performance highlights the need for advanced feature engineering or hybrid modeling approaches to fully utilize RFs capabilities. The primary limitation of this study is the small size of the data set, which can restrict generalizability. Additionally, while metrics like the F1-score and ROC-AUC provide a broader evaluation, future studies should explore calibration metrics to assess model reliability and confidence in predictions.

5. Conclusion

In conclusion, this study evaluated the performance of Random Forest (RF) and Logistic Regression (LR) models for heart attack prediction, utilizing a dataset of 303 samples with 14 features. Both models achieved comparable test accuracy rates of 86%, with RF showing higher training accuracy. This highlights RF's ability to capture non-linear relationships and LR's simplicity and interpretability. The findings emphasize the importance of considering the characteristics of the dataset and the modeling requirements when selecting machine learning algorithms for medical applications. While accuracy was the primary evaluation metric, additional metrics such as precision, recall, F1-score, and ROC-AUC provided a more comprehensive assessment of model performance, confirming the effectiveness of both RF and LR in identifying heart attack risks. For future research, there are several avenues to explore. Researchers should consider evaluating more advanced algorithms, such as gradient boosting machines and neural networks, to see if they can outperform RF and LR. Techniques to address class imbalance, like oversampling or synthetic data generation, could enhance model performance. Additionally, incorporating advanced feature selection methods and expanding the dataset size and diversity will improve the generalizability of the findings. Finally, integrating RF and LR with other approaches, such as deep learning or ensemble techniques, may further enhance predictive accuracy and interpretability. By pursuing these strategies, future work can optimize the predictive capabilities of machine learning models, advancing their clinical utility in heart attack prevention and management.

Acknowledgment

Declaration of Competing Interests

The authors declare no known competing financial interests or personal relationships.

Funding Declaration

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author Contributions

Nilakshman Sooriyaperakasam: Conceptualization, Data Analysis, Writing – Review and Editing; **Hamid Emami**: Methodology, Validation, Investigation, Writing – Original Draft; **Parinaz Entezam and Chisom Ezekiel**: Software, Visualization, Investigation

References

[1] M. Amini, F. Zayeri, and M. Salehi, "Trend analysis of cardiovascular disease mortality, incidence, and mortality-to-incidence ratio: results from global burden of disease study 2017," *BMC public health*, vol. 21, pp. 1–12, 2021.

[2] E. Radwa, H. Ridha, and B. Faycal, "Deep learning-based approaches for myocardial infarction detection: A comprehensive review recent advances and emerging challenges," *Medicine in Novel Technology and Devices*, p. 100322, 2024.

[3] Y. Kumari, P. Bai, F. Waqar, A. T. Asif, B. Irshad, S. Raj, V. Varagantiwar, M. Kumar, F. Neha, S. Chand, *et al.*, "Advancements in the management of endocrine system disorders and arrhythmias: a comprehensive narrative review," *Cureus*, vol. 15, no. 10, 2023.

[4] G. M. Dogheim and A. Hussain, "Patient care through ai-driven remote monitoring: Analyzing the role of predictive models and intelligent alerts in preventive medicine," *Journal of Contemporary Healthcare Analytics*, vol. 7, no. 1, pp. 94–110, 2023.

[5] A. Ogunpola, F. Saeed, S. Basurra, A. M. Albarrak, and S. N. Qasem, "Machine learning-based predictive models for detection of cardiovascular diseases," *Diagnostics*, vol. 14, no. 2, p. 144, 2024.

[6] S. Zhong, K. Zhang, M. Bagheri, J. G. Burken, A. Gu, B. Li, X. Ma, B. L. Marrone, Z. J. Ren, J. Schrier, *et al.*, "Machine learning: new ideas and tools in environmental science and engineering," *Environmental science & technology*, vol. 55, no. 19, pp. 12741–12754, 2021.

[7] Z. Noroozi, A. Orooji, and L. Erfannia, "Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction," *Scientific Reports*, vol. 13, no. 1, p. 22588, 2023.

[8] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, "Handling imbalanced medical datasets: review of a decade of research," *Artificial Intelligence Review*, vol. 57, no. 10, p. 273, 2024.

[9] R. Hasan, "Comparative analysis of machine learning algorithms for heart disease prediction," in *ITM Web of Conferences*, vol. 40, p. 03007, EDP Sciences, 2021.

[10] A. Abdellatif, H. Abdellatef, J. Kanesan, C.-O. Chow, J. H. Chuah, and H. M. Gheni, "Improving the heart disease detection and patients survival using supervised infinite feature selection and improved weighted random forest," *IEEE Access*, vol. 10, pp. 67363–67372, 2022.

[11] M. S. Pathan, A. Nag, M. M. Pathan, and S. Dev, "Analyzing the impact of feature selection on the accuracy of heart disease prediction," *Healthcare Analytics*, vol. 2, p. 100060, 2022.

[12] L. Li and Z.-P. Liu, "A connected network-regularized logistic regression model for feature selection," *Applied Intelligence*, vol. 52, no. 10, pp. 11672–11702, 2022.

[13] J. Hu and S. Szymczak, "A review on longitudinal data analysis with random forest," *Briefings in Bioinformatics*, vol. 24, no. 2, p. bbad002, 2023.

[14] M. Maalouf, "Logistic regression in data analysis: an overview," *International Journal of Data Analysis Techniques and Strategies*, vol. 3, no. 3, pp. 281–299, 2011.

[15] T. J. Cleophas, A. H. Zwinderman, T. J. Cleophas, and A. H. Zwinderman, "Bayesian pearson correlation analysis," *Modern Bayesian statistics in clinical research*, pp. 111–118, 2018.