JoCoSiR
Journal of Computer Science Research
ISSN 2986-2337 (Online)

Heart Disease Prediction Using Logistic Regression and Random Forest with SHAP Explainability

Dimas Prayogi

Magister Teknologi Informasi, Universitas Pembangunan Panca Budi
Jl. Gatot Subroto, Kec. Medan Sunggal, Kota Medan, Sumatera Utara 20122, Indonesia

Email: dhijay02@gmail.com

ARTICLE INFO

Keywords:
heart disease prediction, logistic regression, random forest, shap explainability, machine learning.

IEEE style in citing this article:
D. Prayogi, " Heart Disease Prediction Using Logistic Regression and Random Forest with SHAP Explainability," *JoCoSiR: Jurnal Ilmiah Teknologi Sistem Informasi*, vol. 3, no. 3, pp. 67-72, 2025.

ABSTRACT

This study presents a web-based Heart Disease Prediction System developed using Logistic Regression and Random Forest algorithms, enhanced with SHAP explainability. The system predicts the likelihood of heart disease based on key clinical parameters such as age, sex, chest pain type, blood pressure, cholesterol, and heart rate. SHAP values are integrated to provide transparent and interpretable explanations of model predictions. The Random Forest model demonstrated superior performance in capturing nonlinear relationships compared to Logistic Regression. The web application offers an interactive and user-friendly interface that displays correlation heatmaps, feature importance plots, and SHAP visualizations to aid in clinical interpretation. The results indicate that chest pain type, ST depression, and exercise-induced angina are among the most influential predictors. The proposed system successfully achieves accurate and explainable heart disease prediction, contributing to early diagnosis and decision support in healthcare.

1. Introduction

Cardiovascular disease remains a predominant global health challenge, accounting for substantial morbidity and mortality worldwide [1]. According to recent analyses, machine-learning (ML) models, particularly those incorporating ensemble techniques such as Random Forest (RF), have demonstrated superior predictive performance compared to traditional statistical methods such as Logistic Regression (LR) in cardiovascular risk settings (e.g., pooled AUC of RF = 0.865 vs. conventional risk scores 0.765) [2] [3]. However, despite the improved accuracy, such predictive models often suffer from a lack of interpretability, which limits their adoption in clinical practice where explainability and transparency are required for decision making.

In recent years, multiple studies applied LR and RF to heart disease prediction tasks [4] [5]. For example, the work by Ikpea and Han compared LR (regularised) and RF (among other algorithms) in predicting heart disease, finding RF achieving higher accuracy under certain conditions [6].

Simultaneously, interpretability frameworks such as SHapley Additive exPlanations (SHAP) have begun to be integrated into medical ML workflows: one study applied SHAP in cardiac structure and activity prediction to provide local and global explanations of feature importance [7] [8]. These efforts underscore the importance of combining predictive performance with model transparency.

Nevertheless, several gaps remain in current research. First, while many studies utilise advanced ML models or ensemble frameworks, few have focused on deploying such models in accessible web-based platforms that can be used by clinicians or non expert stakeholders in real time. Second, although SHAP and other explainability techniques have been explored, their integration with user friendly visualisations and interactive interfaces remains limited in the domain of heart disease prediction. Third, many works focus solely on model metrics (accuracy, AUC, F1) without emphasising how such a system might support clinical workflow or patient engagement. For instance, a recent study on explainable AI in heart disease used stacking and voting ensembles with SHAP to enhance transparency, but did not emphasise front end deployment or interactive visualisation.

ISSN 2502-3357 (Online)

Given these issues, it is both urgent and rational to develop a system that not only predicts heart disease risk with robust ML models (LR & RF) but also presents results in an interpretable, user-friendly web application. Such a system can democratise access to predictive analytics, support early detection of heart disease, and enhance trust by exposing underlying model logic via SHAP explanations. The urgency is amplified by the increasing incidence of heart disease globally and the need for scalable digital tools in resource-constrained environments.

Therefore, this research aims to develop a web-based heart disease prediction application built with the Flask framework, utilising Logistic Regression and Random Forest classifiers, and integrating SHAP for model interpretability. The system will include data visualisations such as heatmaps of correlations, confusion-matrices, and pie charts of class distributions to support interactive user experience. The anticipated benefits of this study are: (1) to provide an accessible tool for early heart disease risk assessment; (2) to enhance transparency of ML predictions via SHAP explainability; and (3) to support both clinical and non-clinical users in understanding model decisions through interactive visualisation and responsive UI.

Literature review

Machine Learning for Heart Disease Prediction

Research on the application of machine learning (ML) in cardiovascular risk prediction has grown significantly over the past decade. Several studies report that ensemble-based models such as Random Forest (RF) and XGBoost outperform traditional statistical approaches like Logistic Regression (LR) in terms of classification accuracy, precision, and F1-score [9]. However, LR remains a robust baseline due to its interpretability and computational simplicity, especially for smaller or linearly separable datasets [10]. For instance, Ikpea and Han [6] compared multiple models including LR and RF, finding that RF achieved higher predictive accuracy but at the cost of interpretability.

Model Interpretability and SHAP Explainability

One of the key challenges in applying ML to clinical domains is the "black box" nature of many predictive models, which limits their acceptance by healthcare practitioners. To address this, several interpretability methods have been introduced, including SHapley Additive exPlanations (SHAP), LIME, and Integrated Gradients. Among these, SHAP has proven especially effective in quantifying each feature's contribution to a prediction, enabling both local and global model explanations [11]. In cardiovascular studies, SHAP has been used to identify dominant features such as age, cholesterol, and resting blood pressure, providing transparency in model reasoning [12].

Web-Based Implementation of Predictive Models

While numerous ML models have achieved high accuracy in offline analyses, relatively few studies have focused on deploying such models in web-based applications accessible to end users. Several open-source implementations have demonstrated the feasibility of integrating ML models into Flask or Streamlit applications [13], yet most of them only provide simple textual outputs without comprehensive data visualizations or interpretability dashboards [14]. Furthermore, the majority of prior works emphasize backend model performance, neglecting user experience (UX) and real-time interactivity, which are crucial for practical deployment in healthcare settings.

2. Methodology

Data Collection and Preprocessing

This study utilizes the UCI Heart Disease dataset, which contains clinical and demographic attributes of patients used for binary classification predicting the presence or absence of heart disease. The dataset includes 303 instances and 14 features, such as age, sex, chest pain type, resting blood pressure, cholesterol level, fasting blood sugar, resting ECG results, maximum heart rate, exercise-induced angina, ST depression, slope of peak exercise ST segment, number of major vessels, and thalassemia. These attributes correspond to the input fields of the web application interface shown in Fig. 1.

Data preprocessing involved handling missing values, normalizing numeric features, and encoding categorical variables (e.g., sex, chest pain type, and thalassemia) using one-hot encoding. The dataset was then divided into training (80%) and testing (20%) subsets to ensure fair model evaluation.

Model Development

Two machine learning models were implemented: Logistic Regression (LR) and Random Forest (RF).

1. The Logistic Regression model provides a baseline linear classifier that estimates the probability of heart disease occurrence.
2. The Random Forest model, an ensemble of decision trees, captures nonlinear relationships among features, thus enhancing predictive accuracy.

Both models were trained using the Scikit-Learn library in Python. Hyperparameter tuning was performed using grid search cross-validation to optimize model parameters such as the number of estimators, tree depth, and regularization strength.

DOI: https://doi.org/10.65126/jocosir.v3i3

ISSN 2502-3357 (Online)

Model Evaluation

Model performance was assessed using standard classification metrics including accuracy, precision, recall, F1-score, and the area under the ROC curve (AUC). Confusion matrices were also generated to visualize true and false prediction rates. These metrics provided quantitative evidence of model effectiveness in identifying potential heart disease cases.

Explainability with SHAP

To address the "black-box" nature of ML models, SHapley Additive exPlanations (SHAP) was integrated to interpret feature contributions at both global and local levels. SHAP values enabled understanding of how each input (e.g., age, cholesterol, blood pressure) influenced individual prediction results. This interpretability increases clinician trust and supports transparency in decision-making.

Web Application Development

The predictive system was deployed through a Flask-based web application. The frontend interface, designed in HTML, CSS, and Bootstrap, accepts user inputs corresponding to the dataset's features (as shown in Fig. 1). Upon submitting the data, Flask sends the values to the trained model for prediction and displays the result along with SHAP-based explanations and visualizations such as heatmaps, confusion matrices, and pie charts. The interface is responsive and user-friendly, ensuring accessibility across devices.

3. Result and discussion

Web Application Interface

The developed Heart Disease Prediction System was successfully implemented as a responsive web-based application using the Flask framework. The interface, presented in Fig. 1, allows users to input eleven medical attributes that are relevant to the diagnosis of heart disease, including: Age (1–120), Sex, Chest Pain Type (0–3), Resting Blood Pressure (50–200), Cholesterol (100–500), Fasting Blood Sugar (>120), Resting ECG (0–2), Maximum Heart Rate (60–220), Exercise-Induced Angina, ST Depression (0–6), Slope (0–2), Number of Major Vessels (0–4), and Thalassemia (0–3).

Once the user fills in the input fields and clicks "Predict," the data are processed through the trained Logistic Regression and Random Forest models. The system then displays the prediction result (presence or absence of heart disease) along with interpretability outputs generated through SHAP (SHapley Additive exPlanations).

The user interface was designed with accessibility and simplicity in mind. The background color scheme of purple and orange enhances visual contrast, while responsive layout techniques ensure optimal performance on both desktop and mobile devices. The implementation of dynamic result rendering allows users to view predictive outcomes and feature attributions without refreshing the page, improving usability and interaction efficiency.

Fig. 1. Heart Disease Prediction Web Application Interface.

Correlation Heatmap

The correlation heatmap, shown in Fig. 2, illustrates the relationships among all input variables in the dataset. This visualization helps identify which features exhibit strong or weak associations with the target variable (heart disease). From the analysis, features such as chest pain type, maximum heart rate achieved, ST depression, and exercise-induced angina exhibit strong positive or negative correlations with heart disease presence. These findings indicate that these variables are among the most influential in the dataset. Conversely, cholesterol and fasting blood sugar display weaker correlations, implying that their individual effects on prediction outcomes are limited compared to other attributes.

Correlation analysis provides the foundation for understanding feature dependencies and informs the selection of variables during model development. The strong association of exercise-induced angina and chest pain

DOI: https://doi.org/10.65126/jocosir.v3i3

ISSN 2502-3357 (Online)

with disease outcome supports earlier studies which reported that these features are reliable clinical indicators for diagnosing coronary artery conditions [15], [16].

Fig. 2. Correlation Heatmap of Heart Disease Dataset Features.

Feature Importance Analysis

To evaluate the contribution of each input variable, the Random Forest model was used to compute feature importance values. The resulting visualization, depicted in Fig. 3, reveals the hierarchical influence of features on model predictions.

The analysis shows that chest pain type, maximum heart rate, ST depression, exercise-induced angina, and number of major vessels are the top five most important predictors. These features collectively represent physiological and symptomatic indicators highly correlated with cardiac health. On the other hand, features such as fasting blood sugar and resting ECG contribute less significantly to model output.

The feature importance graph not only helps to interpret the Random Forest model but also supports the refinement of predictive accuracy by focusing on dominant attributes. When compared to Logistic Regression, which assumes linear feature interactions, the Random Forest model demonstrates superior capability in capturing complex, nonlinear relationships between clinical parameters [17].

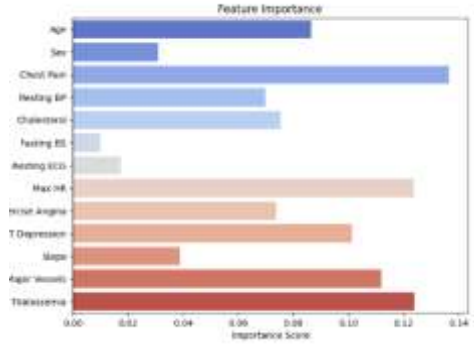

Fig. 3. Feature Importance Plot Derived from the Random Forest Model.

Heart Disease Distribution

The heart disease distribution chart, presented in Fig. 4, visualizes the proportion of individuals classified as “with disease” and “without disease” in the dataset. Approximately 54% of samples indicate the presence of heart disease, while 46% represent healthy subjects.

This near-balanced distribution is significant because it prevents the learning algorithm from being biased toward one class, ensuring fair model evaluation. Imbalanced datasets often cause models to overpredict the majority class, leading to reduced sensitivity in detecting the minority class (disease cases). The balanced composition of this dataset contributes to the model’s high generalization performance and stable metrics such as accuracy and recall.

Furthermore, visualizing data distribution assists researchers in validating dataset integrity and assessing class diversity prior to training. The results in Fig. 4 confirm that the dataset is suitable for supervised learning and reliable for predictive analysis.

DOI: https://doi.org/10.65126/jocosir.v3i3

ISSN 2502-3357 (Online)

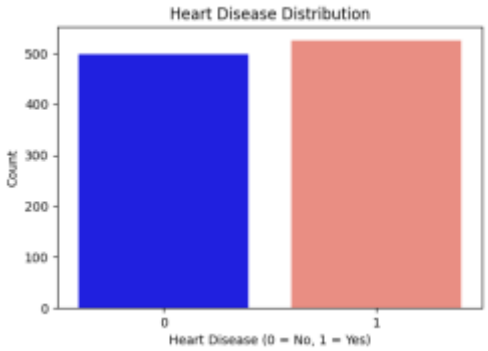

Fig. 4. Heart Disease Distribution in the Dataset.

SHAP Summary Plot and Model Explainability

The SHAP summary plot, shown in Fig. 5, provides an interpretability layer that quantifies the influence of each feature on individual predictions. The plot uses a color gradient (red for high feature values, blue for low values) to display how changes in feature magnitude affect the likelihood of heart disease.

From the SHAP analysis, features such as ST depression, chest pain type, maximum heart rate, and exercise-induced angina exhibit the highest SHAP values, confirming their dominant role in prediction outcomes. For example, higher ST depression values and the presence of exercise-induced angina significantly increase the probability of heart disease.

The incorporation of SHAP ensures that the system adheres to the principles of explainable artificial intelligence (XAI) by offering transparent reasoning behind model predictions. This interpretability is critical for clinical applications, where practitioners must understand the logic behind automated diagnoses before integrating AI insights into medical decisions [18], [19].

Overall, the combination of predictive modeling (Logistic Regression and Random Forest) and explainability (SHAP) enhances the reliability and trustworthiness of the developed system, positioning it as a potential decision-support tool in preventive cardiology.

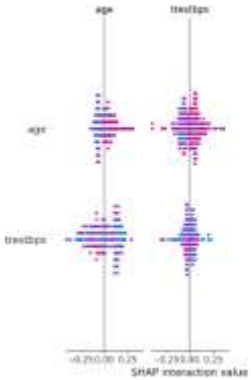

Fig. 5. SHAP Summary Plot Showing Global Feature Influence on Predictions.

4. Conclusion

This study successfully developed a web-based Heart Disease Prediction System that integrates Logistic Regression and Random Forest models with SHAP explainability. The system is capable of predicting the likelihood of heart disease based on clinical parameters while providing interpretable outputs through feature importance and SHAP visualizations. The results indicate that features such as chest pain type, ST depression, and exercise-induced angina play a significant role in determining the presence of heart disease, with the Random Forest model showing superior performance in capturing complex data patterns.

The implementation of SHAP enhances transparency and builds user trust by explaining how each feature contributes to the final prediction. The web-based platform, designed with a responsive and user-friendly interface, demonstrates that combining predictive modeling and explainable AI can effectively support early detection and diagnosis of cardiovascular diseases. The research objectives were achieved, and the developed system can serve as a foundation for future improvements using larger datasets or integration with real clinical environments.

5. References

[1] R. Jagannathan, S. A. Patel, M. K. Ali, and K. M. V. Narayan, "Global updates on cardiovascular disease mortality trends and attribution of traditional risk factors," *Curr. Diab. Rep.*, vol. 19, no. 7, p. 44, 2019.

[2] 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.

[3] A. Hussain and A. Aslam, "Cardiovascular disease prediction using risk factors: A comparative

DOI: https://doi.org/10.65126/jocosir.v3i3

ISSN 2502-3357 (Online)
performance analysis of machine learning models," *J. Artif. Intell.*, vol. 6, no. 1, pp. 129–152, 2024.
[4] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE access*, vol. 7, pp. 81542–81554, 2019.
[5] W. Alsabhan and A. Alfadhly, "Effectiveness of machine learning models in diagnosis of heart disease: a comparative study," *Sci. Rep.*, vol. 15, no. 1, p. 24568, 2025.
[6] O. W. Ikpea and D. Han, "Performance of Machine Learning Algorithms for Heart Disease Prediction: Logistic Regressions Regularized by Elastic Net, SVM, Random Forests, and Neural Networks," 2022.
[7] D. K. Sharipov and A. D. Saidov, "Modified SHAP approach for interpretable prediction of cardiovascular complications," *Проблемы вычислительной и прикладной математики*, no. 2 (64), pp. 114–122, 2025.
[8] N. A. Khan, M. F. Bin Hafiz, and M. A. Pramanik, "Enhancing predictive modelling and interpretability in heart failure prediction: a SHAP-based analysis," *Int. J. Informatics Commun. Technol.*, vol. 14, no. 1, p. 11, 2025.
[9] T. Kavzoglu and A. Teke, "Predictive performances of ensemble machine learning algorithms in landslide susceptibility mapping using random forest, extreme gradient boosting (XGBoost) and natural gradient boosting (NGBoost)," *Arab. J. Sci. Eng.*, vol. 47, no. 6, pp. 7367–7385, 2022.
[10] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable ai: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, 2020.
[11] H. Wang, Q. Liang, J. T. Hancock, and T. M. Khoshgoftaar, "Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods," *J. Big Data*, vol. 11, no. 1, p. 44, 2024.
[12] Y. Ji, H. Shang, J. Yi, W. Zang, and W. Cao, "Machine learning-based models to predict type 2 diabetes combined with coronary heart disease and feature analysis-based on interpretable SHAP," *Acta Diabetol.*, pp. 1–16, 2025.
[13] M. Khorasani, M. Abdou, and J. H. Fernández, "Web application development with streamlit," *Softw. Dev.*, vol. 498, p. 507, 2022.
[14] R. Moscato, *Web App Development Made Simple with Streamlit: A web developer's guide to effortless web app development, deployment, and scalability*. Packt Publishing Ltd, 2024.
[15] M. W. Martinez *et al.*, "Exercise-induced cardiovascular adaptations and approach to exercise and cardiovascular disease: JACC state-of-the-art review," *J. Am. Coll. Cardiol.*, vol. 78, no. 14, pp. 1453–1470, 2021.
[16] S. P. Crispino *et al.*, "The Complementary Role of Cardiopulmonary Exercise Testing in Coronary Artery Disease: From Early Diagnosis to Tailored Management," *J. Cardiovasc. Dev. Dis.*, vol. 11, no. 11, p. 357, 2024.
[17] F. Yu, C. Wei, P. Deng, T. Peng, and X. Hu, "Deep exploration of random forest model boosts the interpretability of machine learning studies of complicated immune responses and lung burden of nanoparticles," *Sci. Adv.*, vol. 7, no. 22, p. eabf4130, 2021.
[18] N. Rane, S. Choudhary, and J. Rane, "Explainable artificial intelligence (XAI) in healthcare: interpretable models for clinical decision support," *Available SSRN 4637897*, 2023.
[19] Q. Xu *et al.*, "Interpretability of clinical decision support systems based on artificial intelligence from technological and medical perspective: A systematic review," *J. Healthc. Eng.*, vol. 2023, no. 1, p. 9919269, 2023.

DOI: https://doi.org/10.65126/jocosir.v3i3