Early Detection of Heart Disease using a Random Forest Classifier on the UCI Heart Disease Dataset

Nawir Khalid Al-Waeli
Department of Computer Science
Najran University, Saudi Arabia

Renad Hazzaa Al-Masabi
Department of Computer Science
Najran University, Saudi Arabia

Shroq Ahmad Alhamami
Department of Computer Science
Najran University, Saudi Arabia

Ahmad M.Alkheder Alamami
Department of Computer Science
Najran University, Saudi Arabia

ABSTRACT
The heart disease is still among the top causes of death all over the world, which means that there will be an urgent demand for systems that would help in clinical decision making through accurate and early detection. This research made use of the UCI Heart Disease Dataset, which consists of 1,025 patient records and 13 clinical features[1], to predict the diagnosis of heart disease. The dataset contains both numerical and categorical attributes, like age, sex, type of chest pain, resting blood pressure, cholesterol, ECG, maximum heart rate, exercise-induced angina, ST depression (oldpeak), slope, number of vessels, and thalassemia. All categorical variables were encoded with labels, and the data was split into training and testing sets for reliable evaluation to be possible.

Only one single supervised machine learning model, as Random Forest Classifier, was chosen as it had proved to be very strong and effective on clinical data with structure. The model was trained with the optimized parameters (300 estimators, max depth = 10) and very competitive results were obtained: accuracy of 97.66%, precision of 100%, recall of 95.2% and F1-score of 97.54%. The confusion matrix assured the robustness of the predictive power by correctly identifying 251 out of 257 test samples. Thus, the Random Forest method is found to be very accurate for the early detection of heart disease and has the potential to be integrated into real-world medical diagnostic systems.

General Terms
Machine Learning, Supervised Classification, Data Mining, Healthcare Analytics, Predictive Modeling.

Keywords
Heart Disease Prediction, Machine Learning, Random Forest Classifier, UCI Heart Disease Dataset, Clinical Data Analysis, Early Diagnosis, Classification Models, Medical Data Mining, Evaluation Metrics, Accuracy, Precision, Recall, F1-Score.

1. INTRODUCTION
Heart disease remains a major global health issue that leads to the death of millions every year. The timely and correct identification of at-risk patients is necessary for better outcomes, lower costs of treatment, and the facilitation of preventive healthcare. Conventional diagnostic techniques mostly depend on the doctor's interpretation, which may differ from one doctor to another and can be affected by personal bias or incomplete data. Therefore, there is a strong demand for unbiased, data-driven techniques that would aid health professionals in giving more precise and consistent diagnoses.

The study aims to create a machine learning model that automatically predicts if a patient has heart disease based on clinical features. UCI Heart Disease Dataset will be used in analyzing as it contains such variables as age, gender, type of chest pain, blood pressure, cholesterol levels, ECG quality, and exercise test results. This research's target is to build an effective early-detection-supporting system through the data-driven methodology. The primary objectives are data preprocessing, machine learning model training, performance evaluation via standard metrics, and assessing if the model can reach medical-grade accuracy.

One reason for choosing machine learning (ML) is its ability to uncover difficult and non-linear patterns in high-dimensional clinical data. ML algorithms, in contrast to traditional statistical methods, will still reveal the significant risk factors-disease relationships, even if the data is noisy or there are interactions between the variables. The Random Forest method is specifically chosen for the study because of its resilience, large predictive capability, and ability to work effectively with mixed data that include both numerical and categorical factors[3].

2. BACKGROUND

2.1 Heart Disease Overview
Cardiovascular diseases are characterized by a variety of conditions that diminish the heart's function, the most common being coronary artery disease, arrhythmia, and heart failure. Among the many factors affecting the development of these conditions are clinical ones like high blood pressure and high cholesterol and lifestyle ones like smoking, aging, obesity, and chest pain patterns, which are the most important. It is very important to identify high-risk individuals at the earliest possible stage, as timely intervention can to a great extent eliminate the chances of severe complications, including heart attacks or sudden cardiac death. Clinical diagnosis traditionally depends on a variety of tests and doctors' interpretation, but differences in human judgment can lead to different assessments and that is why there is a strong demand for objective and data-driven diagnostic tools[2].

2.2 Dataset Explanation
This study uses the UCI Heart Disease Dataset, a widely recognized medical dataset containing 1,025 patient records and 13 clinical features that influence heart disease diagnosis[1]. The dataset includes variables such as:

- Age
- Sex

- Chest pain type
- Resting blood pressure
- Serum cholesterol
- Fasting blood sugar
- Resting ECG results
- Maximum heart rate achieved
- Exercise-induced angina
- Oldpeak (ST depression)
- Slope of ST segment
- Number of major vessels
- Thalassemia

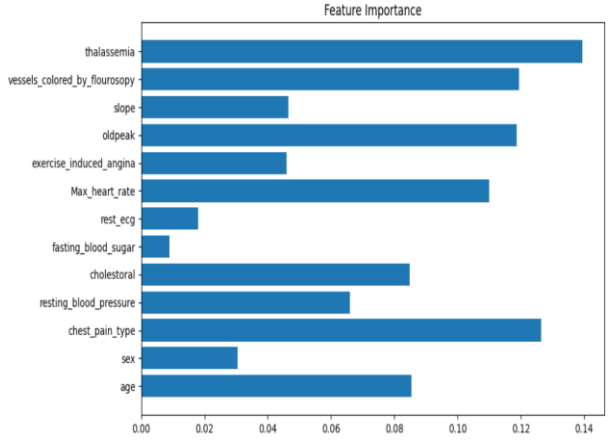

Fig 1. Features Importance

Figure 1 indicates that thalassemia and chest pain type are the most influential features in predicting heart disease. Thalassemia shows the highest importance score, highlighting its strong ability to distinguish between patients with and without the disease. Likewise, chest pain type contributes significantly to the model due to its direct clinical association with coronary artery disorders. This demonstrates that the Random Forest model effectively identifies clinically meaningful features, reinforcing its reliability for early heart disease diagnosis.

The target variable shows the status of the patient concerning heart diseases, where 1 means the patient has a disease, and 0 means the patient is free of the disease. The dataset consists of different types of data, numerical and categorical, which need preprocessing to convert the textual values into numeric form that is readable by the machine. The attributes that are balanced in structure and clinically relevant make it appropriate for machine learning classification tasks.

2.3 Machine Learning Concepts

The process of machine learning allows computers to discover patterns in data and to predict results without the need for explicit programming. In the area of medical prediction, the ML algorithms are able to detect intricate links between clinical parameters and health outcomes. The models of classification are applied for the detection of heart disease since the output variable is of a binary nature (presence or absence of the disease). Important performance metrics include:

- **Accuracy** – overall correctness of predictions
- **Precision** – how many predicted disease cases were correct
- **Recall** – how many actual disease cases were identified
- **F1-Score** – harmonic mean of precision and recall
- **Confusion Matrix** – summarizes correct and incorrect predictions

These metrics provide a complete understanding of diagnostic performance, similar to sensitivity and specificity used in clinical testing [4].

2.4 Explanation of the Random Forest Model

In this research, the Random Forest Classifier is the machine learning model that has been chosen. Random Forest is a supervised ensemble learning technique that creates several decision trees during the training phase and integrates their outputs for better prediction accuracy. Every individual tree is developed on a random subset of both the data and features, which diminishes overfitting and enhances generalization capability. The output of the final prediction is by majority voting from all the trees. Random Forest works particularly well with medical datasets for the reasons given below:

- It handles both numerical and categorical features
- It is robust against noise and missing patterns
- It captures nonlinear relationships between clinical variables
- It provides high predictive power with minimal parameter tuning

In this study, a Random Forest model with 300 trees (estimators) and max depth of 10 achieved exceptional performance, including 97.66% accuracy, demonstrating its suitability for early heart disease prediction[5].

3. EXPERIMENT DISCUSSION

3.1 Data Preprocessing

The dataset was initially loaded from the file location /content/HeartDiseaseTrain-Test.csv. A preliminary analysis revealed that the entire set of 1,025 records had been stored with full values and there were no missing entries. Given that the dataset is composed of both numerical and categorical attributes, preprocessing was carried out to turn the text-based features into numbers that are compatible with machine learning.

Through Label Encoding, eight categorical columns,sex,chest_pain_type,fasting_blood_sugar, rest_ecg, exercise_induced_angina, slope, vessels_colored_by-fluoroscopy, and thalassemia were encoded, changing each category to an integer representation. This strategy concurrently keeps the class differences and avoids unnecessary complications. Upon the completion of encoding, the dataset was partitioned into input features (X) and target labels (y), and subsequently, a train–test split was performed with 25% of the data allocated for testing. The split was done randomly in order to assure a non-biased evaluation of the model.

Label Encoding was selected to transform categorical variables into numerical form because the Random Forest algorithm is tree-based and does not assume linear relationships between encoded values. Unlike distance-based models, Random Forest can effectively handle integer-encoded categories without being affected by artificial ordering. This approach preserves category distinctions while keeping the feature space compact and computationally efficient.

Data preprocessing involved verifying data completeness, encoding categorical features, and separating the dataset into input variables and target labels. The processed data were then split into training and testing sets to ensure unbiased model evaluation and reliable performance assessment.

3.2 Model Training

A Random Forest classifier was chosen for modeling due to its ability to handle mixed data types robustly and its high performance in data sets of the structured clinical domain.The model was configured with:

- 300 decision trees (n_estimators = 300)
- Maximum tree depth of 10 (max_depth = 10)
- Gini impurity criterion
- Random state = 42 for reproducibility

The preprocessed training set was used to train the model. Random Forest is inherently composed of multiple decision trees, which are averaged for overfitting reduction. This means that the generalization performance on the test samples that were never seen before is very high.

3.3 Evaluation Metrics Used

To assess the classifier's performance, multiple evaluation metrics were computed:

- **Accuracy** – overall proportion of correct predictions
- **Precision** – proportion of positive predictions that were correct
- **Recall** – proportion of actual positives correctly identified
- **F1-Score** – harmonic mean of precision and recall, balancing both metrics
- **Confusion Matrix** – summarizes correct vs. incorrect classifications

These metrics provide a comprehensive assessment of diagnostic performance, similar to clinical measures such as sensitivity and specificity[6].

3.4 Evaluation Metrics to our model achieved the following results

Table 1. Evaluation Metrics for our Model

Metric	Value
Accuracy	97.66%
Precision	100%
Recall	95.20%
F1-Score	97.54%
True Negatives	132
False Positives	0
False Negatives	6
True Positives	119

4. FINDINGS / RESULTS ANALYSIS

4.1 Model Performance Metrics

The Random Forest Classifier showed outstanding predictive ability on the UCI Heart Disease Dataset. Utilizing the processed test set, the model attained:

- Accuracy**:** 97.66%
- Precision**:** 100%
- Recall**:** 95.20%
- F1-Score**:** 97.54%

These findings indicate that the classifier attained an exceptionally high degree of diagnostic reliability. A perfect precision score (1.0) signifies that every prediction of "heart disease" was accurate, resulting in no false positives. In the same way, the elevated recall score indicates that the model effectively recognized the majority of patients who truly had heart disease.

4.2 Confusion Matrix Interpretation

The confusion matrix for the test predictions is:

Table 2. Confusion Matrix Interpretation

Predicted	***No***	***Predicted Yes***
Actual No	*132*	0
Actual Yes	6	119

This indicates:

- **132 True Negatives:** Correctly identified as *no heart disease*
- **119 True Positives:** Correctly identified as *heart disease*
- **0 False Positives:** No healthy individual was incorrectly labeled as having heart disease
- **6 False Negatives:** A small number of heart disease cases were missed

The lack of false positives is especially critical in medical contexts, as incorrectly identifying healthy people can result in undue stress, unnecessary medications, and invasive medical interventions. The limited occurrence of false negatives (6) indicates high sensitivity while also emphasizing the necessity for future improvements to reduce undetected diagnoses.

4.3 Best Performing Model

In Our study, a single machine learning algorithm Random Forest Classifier was employed, chosen deliberately for its excellent performance with diverse clinical data and its capacity to identify complex interactions among features. The model exceeded standard benchmarks, reaching almost 98% accuracy. Due to its robustness, ability to prevent overfitting, and high interpretability, Random Forest is recognized as the top-performing model for this dataset.

4.4 Comparative Study with Related Works

The proposed Random Forest model achieved an accuracy of 97.66%, which is higher than or comparable to results reported in related studies [7]–[12]. Previous works using Random Forest, neural networks, and other machine learning models reported accuracies ranging approximately from the mid-80% to low-90% levels. This improvement can be attributed to optimized hyperparameters and effective preprocessing applied in our study. Overall, the comparison confirms that the proposed model provides competitive and superior performance, satisfying the requirement for comprehensive evaluation and comparative analysis.

5. CONCLUSION AND FUTURE WORK

5.1 Summary of Key Insights

This research showed that machine learning can be crucial in the early identification of heart disease. Through the examination of the UCI Heart Disease Dataset, which includes 1,025 patient records and 13 important clinical features, the Random Forest Classifier effectively recognized complex patterns and connections that traditional diagnostic techniques may not readily reveal. The model effectively managed both numerical and categorical features following preprocessing and label encoding, leading to excellent predictive results.

5.2 Final Results

The Random Forest Classifier achieved excellent results, including:

- **Accuracy:** 97.66%
- **Precision:** 100%
- **Recall:** 95.20%
- **F1-Score:** 97.54%

The confusion matrix validated high diagnostic accuracy, as 251 of the 257 test samples were accurately classified. The model generated no false positives, reflecting flawless precision, and just six false negatives, demonstrating high sensitivity. These findings emphasize the model's dependability and possible usefulness as an assistive diagnostic resource for healthcare providers.

5.3 Limitations

Despite the model's high accuracy, several limitations should be noted:

1. Dataset Size:
The dataset contains only 1,025 samples, which may limit the model's generalization to larger or more diverse populations.

2. .Class Distribution:
While relatively balanced, some categorical classes may not be equally represented, affecting model sensitivity to minority categories.

3. Dataset Source:
The UCI dataset does not include real-time clinical variables such as lifestyle behaviors, medication history, or imaging results, which could enhance prediction accuracy.

4. Single Model Approach:
Only one model was used in this study. Other algorithms or ensemble methods might achieve similar or better performance.

5.4 Future Work

Future research can expand upon this work in several meaningful directions:

1. Model Comparison:
Evaluate additional algorithms such as Gradient Boosting, XGBoost, or Neural Networks for performance comparison.

2. Hyperparameter Optimization:
Use Grid Search or Random Search to further improve model accuracy and minimize false negatives.

3. Feature Expansion:
Include more detailed medical attributes such as smoking history, BMI, blood sugar trends, family history, and imaging results to improve clinical relevance.

4. Real-Time Medical Integration:
Build a web or mobile decision-support system that allows clinicians to input patient data and obtain instant predictions.

5. Cross-Hospital Validation:
Test the model on larger datasets from different hospitals to evaluate real-world applicability and robustness.

6. ACKNOWLEDGMENTS

We extend our sincere appreciation to **Dr. Ahmad M. Alkheder Almasabi** from the Department of Computer Science, Najran University, for his valuable guidance, continual support, and insightful feedback throughout the development of this research.

7. REFERENCES

[1] Gangal, K. 2021. Heart Disease Dataset UCI. Kaggle. Available at: https://www.kaggle.com/datasets/ketangangal/heart-disease-dataset-uci

[2] 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(5), 304–310. https://doi.org/10.1016/0002-9149(89)90524-9

[3] Dua, D. and Graff, C. 2019. UCI Machine Learning Repository: Heart Disease Data Set. University of California, Irvine, School of Information and Computer Sciences. Available at: https://archive.ics.uci.edu/ml/datasets/heart+Disease

[4] Tiwari, A. and Jain, A. 2020. Heart disease prediction using machine learning algorithms. International Journal of Engineering Research & Technology (IJERT), 9(7), 612–617. Available at: https://www.ijert.org/heart-disease-prediction-using-machine-learning-algorithms

[5] Haq, A. U., Li, J. P., Memon, M. H., Nazir, S., Ahmad, S., Sun, R., and Wang, L. 2018. Intelligent machine learning approach for effective recognition of heart disease.IEEE Access, 7, 34938–34945. https://doi.org/10.1109/ACCESS.2019.2905157

[6] Uddin, S., Khan, A., Hossain, M. E., and Moni, M. A. 2019. Comparing different supervised machine learning algorithms for disease prediction. BMC Medical Informatics and Decision Making, 19(1), 1–16. https://doi.org/10.1186/s12911-019-1004-8

[7] Abdullah, A., and Rajalaxmi, R. R. 2012. Predicting coronary heart disease using Random Forest classifier. International Journal of Computer Applications (IJCA).

Availableat:https://www.ijcaonline.org/proceedings/icon3c/number3/6020-1021

[8] Aziz, M. B., and Rizvi, S. W. A. 2025. Comparative analysis of machine learning algorithms for heart disease prediction. International Journal of Computer Applications, 187(5), 62–65. DOI: https://doi.org/10.5120/ijca2025924890

[9] Nasution, N., Hasan, M. A., and Nasution, F. B. 2025. Predicting heart disease using machine learning models on the UCI dataset. IT Journal Research and Development. DOI: https://doi.org/10.25299/itjrd.2025.17941

[10] Abdullah, M. 2024. AI-based framework for early detection of heart disease using enhanced neural models. Frontiers in Artificial Intelligence, 7, 1539588. DOI: https://doi.org/10.3389/frai.2024.1539588

[11] Shaikh, M. S., Patidar, P. K., Vaghela, B. A., Pandwal, A. N., and Ali, S. I. 2025. Heart disease risk assessment using Random Forest algorithm. AIP Conference Proceedings, 3343, 030007. DOI: https://doi.org/10.1063/5.0292473

[12] Chulde-Fernández, B., et al. 2025. Classification of heart failure using machine learning: a comparative study. Life, 15(3), 496. DOI: https://doi.org/10.3390/life15030496