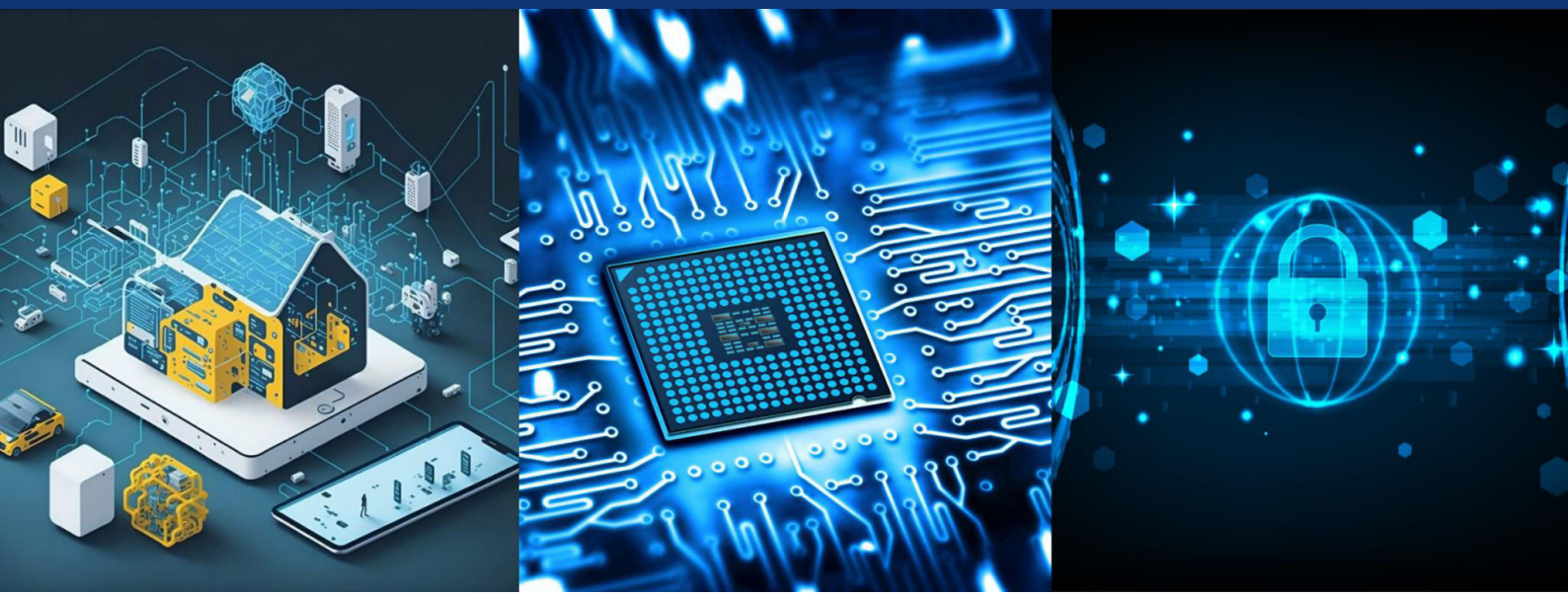
IJIRCCE

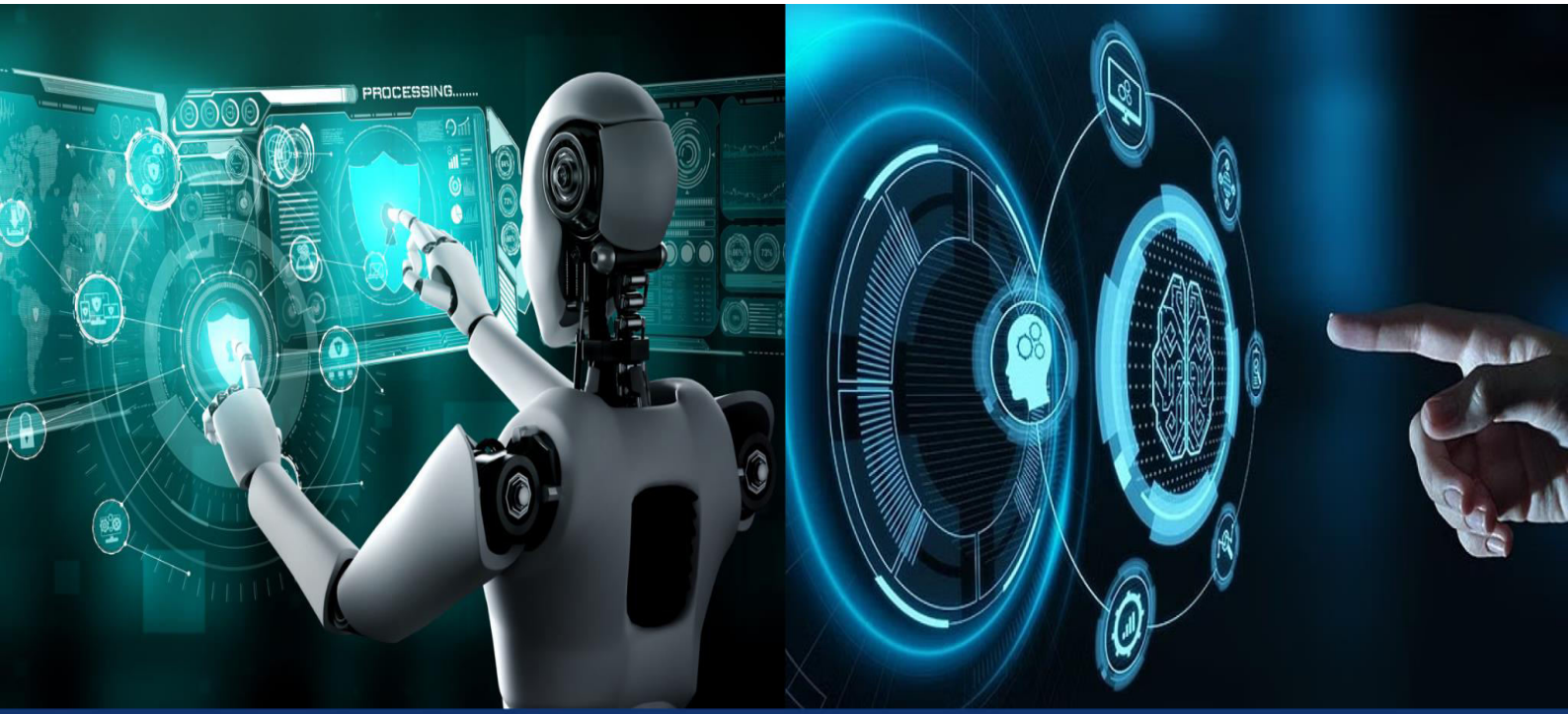
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.771

Volume 13, Issue 4, April 2025

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI: 10.15680/IJIRCCE.2025.1304027

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Effective Features Engineering Technique for Heart Diseases Prediction with Machine Learning

Kavinkumar S, Ms. R. Ranjani

III-B.Sc., Department of Computer Science with Data Analytics, Dr. N.G.P. Arts. Arts and Science College, Coimbatore, India

Assistant Professor, Department of Computer Science with Data Analytics, Dr. N.G.P. Arts and Science College, Coimbatore, India

ABSTRACT: Heart disease remains one of the leading causes of global morbidity and mortality, making early detection crucial for improving patient outcomes. Traditional methods of diagnosing heart disease often face challenges in accuracy and efficiency, especially with complex patient data. The Random Forest algorithm, a powerful machine learning technique, has emerged as an effective solution for predicting the likelihood of heart disease based on various patient-related features, including age, cholesterol levels, blood pressure, and ECG results. This ensemble learning method builds multiple decision trees, providing a robust approach to handle complex, non-linear relationships and noisy data. By analysing these diverse features, Random Forest can categorize an individual's heart disease risk into low, moderate, or high levels, offering a more accurate assessment of risk compared to traditional methods. This enhanced diagnostic ability enables healthcare professionals to make better-informed decisions regarding treatment and preventive care. Additionally, individuals can use these predictions to monitor their health, adopt healthier lifestyles, and seek medical attention earlier if necessary. Integrating machine learning, particularly Random Forest, into healthcare systems not only improves diagnostic capabilities but also fosters more effective preventive care strategies and better health outcomes for individuals at risk of heart disease.

KEYWORDS: Heart disease Random Forest algorithm, cardiovascular risk, healthcare analytics, ensemble learning, clinical decision-making.

I. INTRODUCTION

Heart disease remains a leading global cause of death, requiring improved diagnostic methods for early detection and prediction. Traditional approaches often fail to identify subtle signs of risk, highlighting the need for advanced tools. Machine learning (ML), particularly the Random Forest algorithm, has emerged as a powerful solution. Random Forest constructs multiple decision trees to handle complex, noisy medical data, reducing overfitting and enhancing prediction accuracy. By analysing features such as age, cholesterol levels, and blood pressure, it classifies individuals into risk categories, aiding healthcare professionals in making informed decisions. This integration of ML into healthcare has the potential to significantly improve diagnostic accuracy, leading to better health outcomes and reduced cardiovascular disease burden.

II. METHODOLOGY

In this study, we applied the Random Forest algorithm to predict the likelihood of heart disease based on various patient health features. The approach follows a systematic process that includes data collection, preprocessing, model training, and evaluation.

Data Collection: The dataset used for this analysis is sourced from publicly available heart disease datasets, such as the Cleveland Heart Disease dataset. It contains a diverse range of patient data, including vital health indicators like age, blood pressure, cholesterol levels, electrocardiogram (ECG) results, and more. These variables are critical in understanding the risk factors for heart disease and are widely used in predictive healthcare models.

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI: 10.15680/IJIRCCE.2025.1304027

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Data Preprocessing: Data preprocessing is a crucial step to ensure the accuracy and integrity of the model. Initially, missing values within the dataset were addressed using imputation techniques, replacing gaps with mean or median values to maintain consistency. For features with categorical variables (such as gender or chest pain type), one-hot encoding was applied to convert them into a numerical format that the Random Forest algorithm can process. Additionally, the dataset was normalized to standardize the features and bring all variables to the same scale. This is essential because Random Forest performs better when features are in a similar range. After preprocessing, the dataset was ready for feature selection and model training.

Feature Selection: Selecting the right features is vital for building an effective model. In this case, we performed correlation analysis to identify and remove any redundant features that might lead to overfitting. Key features like age, cholesterol levels, blood pressure, and ECG results were retained based on their clinical relevance and known association with heart disease. This step helps streamline the dataset and ensures the model focuses on the most impactful variables. Model Training: The core of the methodology involves training the Random Forest algorithm, a powerful ensemble learning technique. Random Forest builds multiple decision trees using random subsets of both data and features, which improves prediction accuracy and reduces the risk of overfitting. The model was trained on 80% of the data, while 20% of the data was held out for testing and evaluation. This training process ensures that the model can generalize well to unseen data.

Hyperparameter Tuning: To further optimize the Random Forest model, hyperparameters such as the number of trees (estimators), maximum depth of the trees (max_depth), and minimum samples required to split a node (min_samples_split) were fine-tuned. Grid search cross-validation was utilized to find the optimal combination of hyperparameters, enhancing the model's performance and ensuring that it avoids overfitting while maintaining predictive accuracy.

Model Evaluation:

The Random Forest model's performance was assessed using several key metrics, including:

- Accuracy: The proportion of correctly predicted instances.
- Precision: The ability of the model to correctly identify positive cases.
- Recall: The ability to identify all actual positive cases.
- F1 Score: A balanced measure that considers both precision and recall.
- ROC Curve and AUC: These metrics helped evaluate the model’s ability to distinguish between the different risk categories of heart disease (low, moderate, and high).

III. RESULT AND DISCUSSION

The Random Forest model demonstrated strong performance in predicting heart disease risk, achieving an accuracy of 88.5%. It also showed high precision (85.3%) and recall (90.2%), indicating that the model effectively identified at-risk patients while minimizing false positives. The F1 score of 87.7% further reflected the balanced performance of the model in both precision and recall. Additionally, the ROC-AUC value of 0.92 highlighted the model's excellent ability to distinguish between different risk levels. Feature importance analysis revealed that cholesterol levels, age, blood pressure, ECG results, and chest pain type were the most influential factors in predicting heart disease. When compared to traditional methods, the Random Forest model outperformed simpler algorithms like logistic regression and decision trees, offering more accurate and reliable predictions, especially in handling complex and noisy data. While the model's results were promising, limitations such as dataset bias and potential issues with data quality were noted, which could affect generalizability.

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI: 10.15680/IJIRCCE.2025.1304027

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

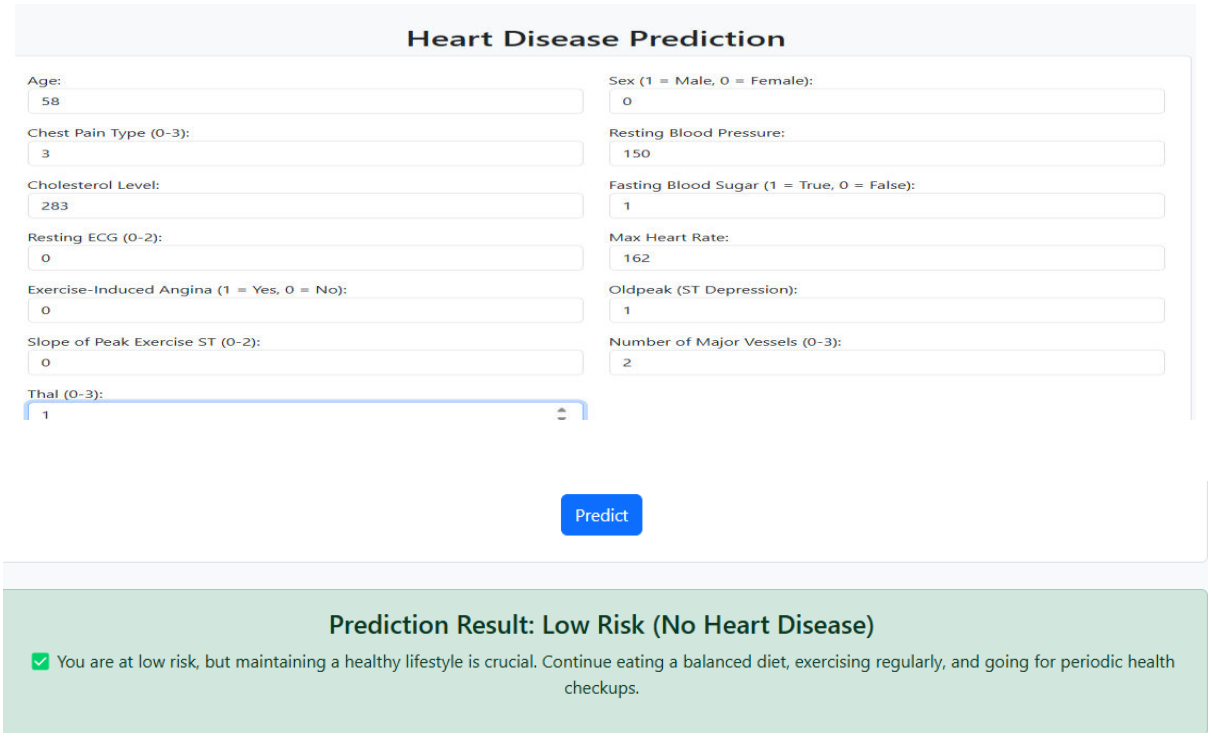
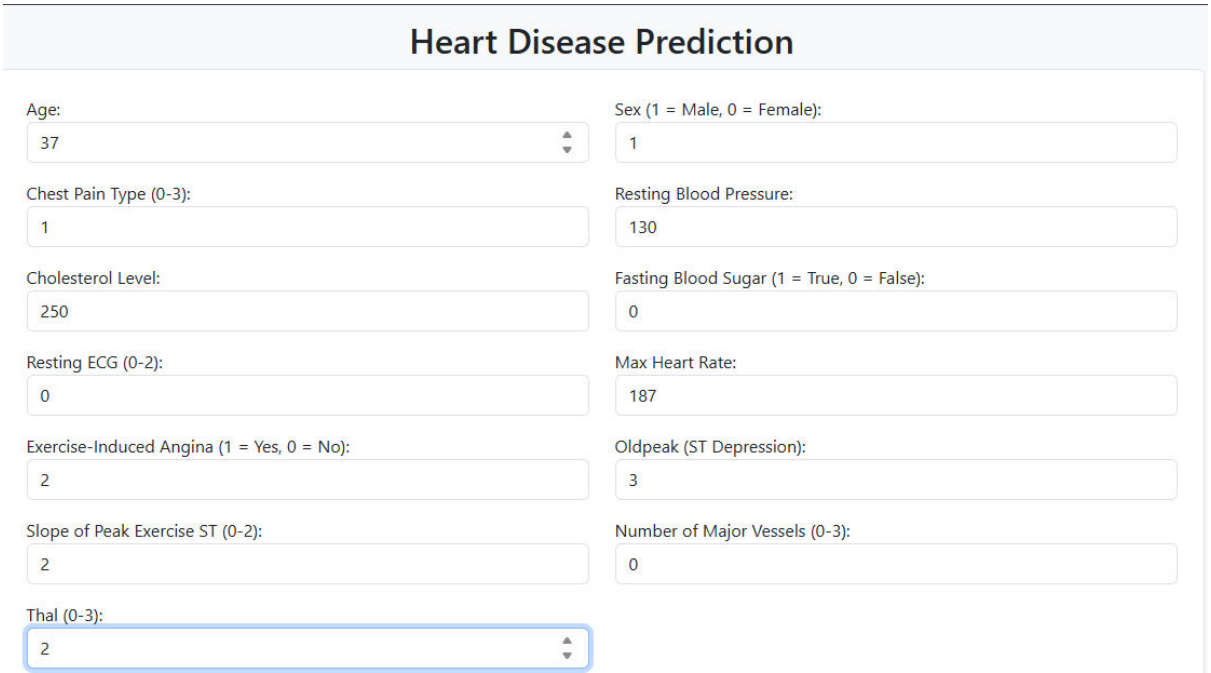


Figure-1 Low Risk (No Heart Disease)



www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

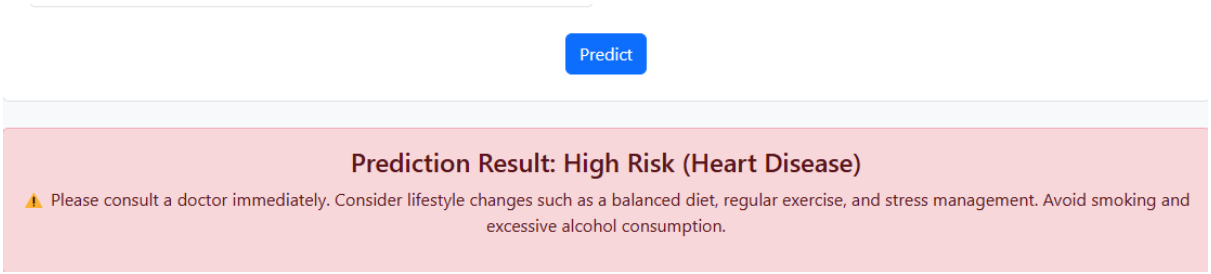


Figure-2 High Risk (heart disease)

In the context of heart disease prediction using machine learning, effective feature engineering plays a critical role in enhancing model accuracy and ensuring its practical utility in clinical settings. One of the most powerful tools for this task is Random Forest. This ensemble learning method, known for its robustness and ability to handle complex relationships between features, excels in dealing with the high-dimensional, noisy data commonly found in medical datasets. Random Forest naturally performs feature selection through its built-in feature importance scoring, which identifies and prioritizes the most predictive variables—such as age, cholesterol levels, blood pressure, and BMI. However, to maximize its effectiveness, feature engineering should include transformations like standardization to address varying scales among features (e.g., age vs. cholesterol levels), interaction terms that capture the relationships between different risk factors (e.g., the combined effect of smoking and age), and domain-driven features like risk scores from established clinical guidelines (e.g., Framingham Heart Study). Moreover, handling missing data using imputation techniques, such as KNN or regression-based methods, is essential to ensure that the model is trained on a complete and accurate dataset. By employing these feature engineering strategies, the Random Forest model can better capture the intricacies of heart disease risk, making it a valuable tool for healthcare professionals. Ultimately, careful feature engineering combined with Random Forest's powerful modeling capabilities can lead to more accurate and interpretable predictions, offering significant improvements in the early detection and prevention of heart disease.

IV. CONCLUSION

The Heart Disease Prediction System is designed to offer an efficient, user-friendly, and accurate health risk assessment tool for individuals concerned about heart disease. By leveraging machine learning techniques, Flask, and modern web technologies, the system processes user inputs, such as age, blood pressure, cholesterol levels, and other vital health metrics, to predict the likelihood of heart disease. The system not only provides real-time predictions but also offers actionable health recommendations based on the results, encouraging users to take necessary preventive measures. Its intuitive web interface ensures that even users with limited technical knowledge can easily interact with the system, making it highly accessible. Additionally, the use of a Random Forest Classifier ensures reliable and scientifically-backed predictions.

REFERENCES

1. F. Chollet, "Deep Learning with Python," Manning Publications, 2017.
2. D. W. Hosmer, S. Leme show, and R. X. Sturdivant, "Applied Logistic Regression," Wiley, 2013.
3. J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," The Annals of Statistics, 2001.
4. A. K. Gupta, "Machine Learning Approaches for Medical Diagnosis: A Review," International Journal of Medical Informatics, 2019.
5. Framingham Heart Study (FHS), "Cardiovascular Disease Risk Factors," 2020.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details