

Explainable Ensemble Machine Learning Framework For Heart Disease Prediction

Mr. P. Ranjith¹, Krithik Ragul S², Moses Praveen R³, Vigneshwaran M⁴

^{1, 2, 3, 4}Dept of Computer Science and Business Systems

^{1, 2, 3, 4} K. Ramakrishnan College of Engineering, Trichy, India

Abstract- Heart disease is one of the leading causes of death worldwide, and early prediction plays an important role in reducing severe health risks. Traditional diagnostic methods such as ECG, blood tests, and imaging require expert interpretation, more time, and higher cost. This paper proposes an Efficient Heart Disease Prediction system using Machine Learning techniques to predict the possibility of heart disease based on clinical data such as age, blood pressure, cholesterol, chest pain type, ECG results, and heart rate. The system applies algorithms such as Logistic Regression, K-Nearest Neighbors, Support Vector Machine, Random Forest, and XGBoost. Among these models, Random Forest and XGBoost provide better accuracy and reliability. Explainable AI techniques such as SHAP and LIME are also used to make the prediction results transparent and understandable for doctors and patients.

Keywords: Heart Disease Prediction, Machine Learning, Random Forest, XGBoost, Logistic Regression, Explainable AI, SHAP, LIME, Healthcare Prediction.

I. INTRODUCTION

1.1 Background

Heart disease affects the heart and blood vessels and is mainly caused by factors such as unhealthy diet, stress, lack of physical activity, smoking, obesity, high blood pressure, and high cholesterol. Early detection is very important because it helps doctors provide proper treatment before the condition becomes serious.

Traditional diagnosis depends on medical tests such as ECG, blood tests, and imaging techniques. These methods are useful but require time, cost, and expert knowledge. Machine Learning helps in analyzing large medical datasets and identifying hidden patterns that may not be easily detected manually.

1.2 Need for Heart Disease Prediction

In many cases, heart disease symptoms are ignored during the early stage. Late diagnosis can lead to serious conditions such as heart attack, stroke, or other complications.

Therefore, an automated prediction system is needed to support doctors in early diagnosis.

Machine Learning models can process patient health details and predict whether a person has a risk of heart disease. This reduces diagnosis time and improves decision-making.

1.3 Scope of the System

The proposed system can be used in hospitals, clinics, diagnostic centers, and healthcare applications. It can assist doctors by providing fast and accurate prediction results based on patient clinical data. The system can also be integrated with web applications for easy access.

II. PROBLEM STATEMENT

Early detection of heart disease is difficult because traditional diagnosis depends heavily on expert doctors, medical equipment, and laboratory tests. These methods are time-consuming, costly, and may not be easily available in rural areas.

Existing AI-based systems also have limitations such as low accuracy, limited interpretability, and dependency on large datasets. Some models provide predictions but do not explain why a patient is classified as high-risk or low-risk. Therefore, there is a need for an efficient, accurate, and interpretable heart disease prediction system.

III. OBJECTIVES

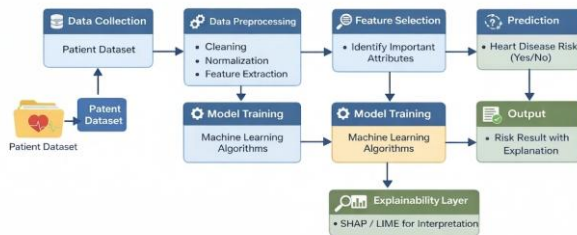
3.1 Main Objective

The main objective of this project is to develop an Efficient Heart Disease Prediction system using Machine Learning algorithms to predict heart disease risk accurately and quickly.

3.2 Specific Objectives

To collect and process patient clinical data; to apply Machine Learning algorithms for heart disease prediction; to compare Logistic Regression, KNN, SVM, Random Forest,

and XGBoost; to improve accuracy using ensemble and boosting methods; and to use Explainable AI techniques such as SHAP and LIME.



IV. LITERATURE SURVEY

Several research studies have used Machine Learning techniques for heart disease prediction. Traditional models such as Logistic Regression, KNN, Decision Tree, and SVM are commonly used for classification problems in healthcare. Recent studies show that ensemble models such as Random Forest and boosting algorithms such as XGBoost provide higher accuracy compared to basic models. Explainable AI methods such as SHAP and LIME are also gaining importance in healthcare because they help explain how a model makes a prediction.

V. PROPOSED SYSTEM

5.1 Overview

The proposed system predicts heart disease using patient clinical data such as age, blood pressure, cholesterol level, ECG result, chest pain type, maximum heart rate, and other medical attributes. The system uses multiple Machine Learning algorithms and compares their performance.

5.2 Working Principle

First, the patient data is collected and preprocessed. Missing values are handled, unnecessary data is removed, and important features are selected. Then, the dataset is divided into training and testing data. Finally, the model predicts whether the patient has a risk of heart disease or not.

VI. SYSTEM ARCHITECTURE

The system architecture consists of the user layer, data processing layer, machine learning layer, prediction layer, explainable AI layer, and output layer. The user enters patient details, the data processing layer cleans and prepares the data, the machine learning layer applies algorithms, and the output layer displays the final prediction with explanation.

VII. METHODOLOGY

The methodology includes data collection, data preprocessing, feature selection, model training, model testing, model comparison, prediction, and explanation. Different algorithms are trained and tested to identify the best-performing model for heart disease prediction.

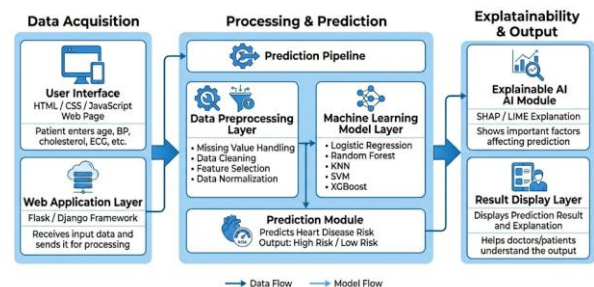
VIII. HARDWARE IMPLEMENTATION

The hardware requirements include a computer or laptop, minimum 4 GB RAM, Intel i3 processor or above, internet connection, and storage for dataset and project files. Since this is a software-based prediction system, no special medical hardware is required.

IX. SOFTWARE IMPLEMENTATION

The software requirements include Python, HTML, CSS, JavaScript, Flask or Django, Scikit-learn, Pandas, NumPy, Matplotlib, SHAP, LIME, Jupyter Notebook, and VS Code. The trained model can be connected to a web application where users enter patient details and receive prediction results.

SOFTWARE IMPLEMENTATION DIAGRAM



X. RESULTS AND DISCUSSION

Basic models such as Logistic Regression and KNN provide moderate accuracy of around 80–85%. SVM and Decision Tree provide better performance but may require tuning. Random Forest and XGBoost provide higher accuracy because they use ensemble and boosting techniques.

The proposed system improves accuracy from around 80–90% to nearly 95–98% using advanced ensemble models. Explainable AI techniques make the system more useful by explaining the reason behind each prediction.

XI. ADVANTAGES

The system provides early prediction of heart disease, reduces diagnosis time, improves accuracy, supports doctors in decision-making, reduces dependency on manual diagnosis, uses Explainable AI for transparent results, and can be implemented as a web-based healthcare application.

XII. APPLICATIONS

The system can be used in hospitals, clinics, diagnostic centers, healthcare web applications, remote healthcare systems, medical research, preventive healthcare monitoring, and rural healthcare support systems.

XIII. FUTURE ENHANCEMENTS

In the future, the system can be improved by integrating real-time patient monitoring devices such as smartwatches and IoT-based health sensors. Deep Learning models can also be used for improving prediction accuracy. A mobile application and cloud storage can be added for better accessibility and secure patient record management.

XIV. CONCLUSION

The Efficient Heart Disease Prediction system provides a smart and accurate method for predicting heart disease using Machine Learning algorithms. The system analyzes clinical data such as age, blood pressure, cholesterol, ECG results, and heart rate to predict heart disease risk. Advanced models such as Random Forest and XGBoost improve accuracy, while Explainable AI techniques such as SHAP and LIME make the prediction more transparent.

REFERENCES

- [1] UCI Machine Learning Repository, "Heart Disease Dataset."
- [2] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," IEEE Access, 2019.
- [3] A. K. Gárate-Escamila et al., "Classification models for heart disease prediction using machine learning techniques," Informatics in Medicine Unlocked, 2020.
- [4] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," ACM SIGKDD, 2016.
- [5] S. Lundberg and S. Lee, "A unified approach to interpreting model predictions," NeurIPS, 2017.
- [6] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," ACM SIGKDD, 2016.