Heart Disease Detection Using Machine Learning

By: Prawinkumar

Date: May 2025

Abstract

Heart disease is one of the primary causes of death globally. The early and accurate detection of heart disease is crucial to improving the survival rates of patients. In this project, machine learning algorithms were implemented to create a predictive model for heart disease detection. Specifically, Random Forest and K-Nearest Neighbors (KNN) algorithms were used. The Random Forest algorithm achieved an impressive accuracy of 98%, while KNN reached 95%, both surpassing the traditional model's accuracy of 83%. This result demonstrates the efficiency of ensemble and instance-based learning methods in healthcare predictive analytics.

Introduction

Heart disease is a growing health concern worldwide, responsible for a significant number of premature deaths. Accurate and timely diagnosis can help reduce the risk of severe health complications. Machine learning (ML) algorithms are being increasingly adopted in the medical field to support and enhance diagnostic systems. Python, a powerful programming language with vast libraries such as Scikit-learn, Pandas, and Matplotlib, provides robust tools for developing such systems. In this work, we explore the use of Random Forest and KNN to build an effective heart disease prediction system.

Problem Statement

The existing systems for heart disease detection have limited accuracy, with many models achieving around 83%. These models often suffer from overfitting, underfitting, or poor generalization. The goal of this project is to increase the predictive accuracy and reliability of heart disease diagnosis by using more advanced

machine learning algorithms - Random Forest and KNN - and evaluating their effectiveness on publicly available medical data.

Methodology

The methodology consists of several key steps:

1. Data Collection: The dataset used includes various patient health attributes such as age, sex, blood pressure, cholesterol levels, and ECG results. The data was sourced from the UCI Machine Learning Repository.
2. Data Preprocessing: Missing values were handled, and categorical variables were encoded. The dataset was standardized to normalize the range of features.
3. Model Training: The dataset was split into training and testing sets in an 80/20 ratio. Both Random Forest and KNN classifiers were trained using Scikit-learn.
4. Evaluation: Models were evaluated using metrics such as accuracy, confusion matrix, and cross-validation scores.

Existing System

Existing models for heart disease prediction often rely on logistic regression or decision trees. These models have shown limited accuracy, typically around 83%. While simple and interpretable, these methods may not capture complex interactions among features, which can lead to missed diagnoses or false positives. The goal is to move beyond such limitations using more advanced ML models.

Proposed System

The proposed system incorporates two powerful machine learning algorithms:

- Random Forest: An ensemble learning method that builds multiple decision trees and combines their outputs to improve accuracy and prevent overfitting. It handles both classification and regression tasks effectively.
- K-Nearest Neighbors (KNN): A simple, instance-based learning algorithm that classifies new data points based on the majority class among their nearest neighbors. It is particularly effective with well-distributed data.

Together, these models achieved accuracies of 98% (Random Forest) and 95% (KNN), representing a significant improvement.

Results & Comparison

A comparative study was conducted among various models. The results are summarized below:

Algorithm	Accuracy
Existing Model	83%
K-Nearest Neighbors	95%
Random Forest	98%

The Random Forest model demonstrated the highest accuracy, indicating that ensemble learning is more effective for this type of predictive task.

Conclusion

In this project, we have demonstrated that machine learning algorithms, particularly Random Forest and KNN, can significantly enhance the accuracy of heart disease detection. The models trained on a structured dataset achieved accuracy levels of 98% and 95%, respectively. This work suggests that advanced ML

techniques can be a valuable asset in medical diagnostics, potentially saving lives through early detection.

References

1. UCI Machine Learning Repository

2. Scikit-learn Documentation

3. Healthcare Analytics, Elsevier, 2022

4. Python Libraries: Pandas, Numpy, Matplotlib, Seaborn

5. Research paper by Victor Chang et al. on Heart Disease Detection