Prediction of Heart Disease with Machine Learning

Jianxing Wu, Jieming Yu
University of Waterloo, System Design Engineering

Introduction

Heart diseases, especially Cardiovascular diseases, are currently the leading cause of mortality worldwide. They share several key clinical features, including sleep disturbances, irregular heartbeat, and swollen legs with other various diseases, which make accurate diagnosis challenging. Machine learning models trained on enormous medical records can provide instant preliminary diagnosis to assist modern clinical practices.

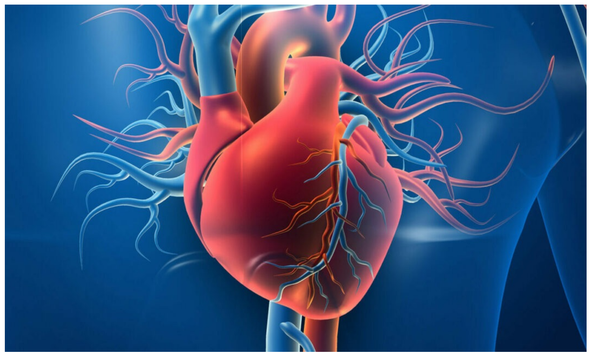


Figure 1: Heart Disease (illustration)
Source: https://www.hri.org.nz/health/learn/cardiovascular-disease

key words: heart disease, machine learning

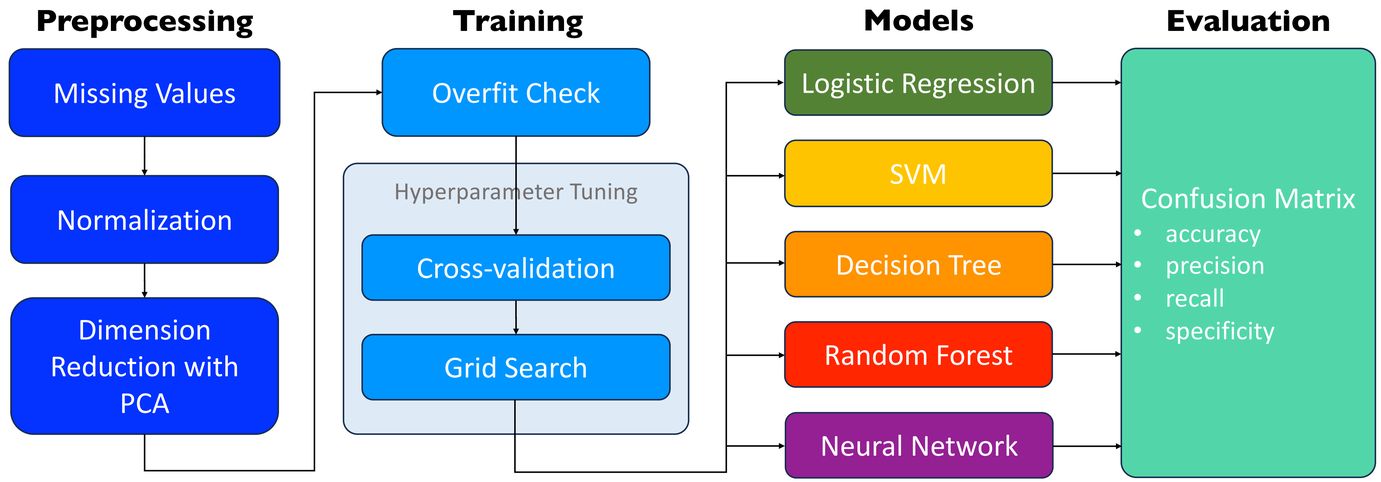
Summary

- We trained and tested five machine learning models on the Cleveland Heart Disease Dataset with normalized and dimensionally reduced features.
- SVM was the most effective model.
- Larger datasets with recent patient information could be used to train models with better practicality.

Dataset: Cleveland Heart Disease Dataset from UC Irving, Machine Learning Repository
Link: https://archive.ics.uci.edu/dataset/45/heart+disease

Methods and Techniques

A variety of machine learning models have been employed to predict the presence of heart disease based on the UCI Cleveland Heart Disease Dataset. A schematic representation of the general process is presented below.



NOTE: Grid search for each model is varied in terms of hyperparameter type and number.

Comparison and Analysis

Five models were trained and evaluated on our dataset. We compared their performance on testing accuracy, precision, recall, specificity and their cost of hyperparameter tuning time. Results are shown in the table below.

Table 1: Model Comparison (scores are given as percentages)

	Tuning Time	Validation	Train	Test Accuracy	Test Precision	Test Recall	Test Specificity
Logistic Regression	0.16s	81.93	81.94	82.89	81.58	83.78	82.05
SVM	2.80s	83.25	80.62	85.53	86.11	83.78	87.18
Decision Tree	6.67s	82.36	86.78	82.89	90.00	72.97	92.31
Random Forest	196.82s	84.56	88.11	82.89	85.29	78.37	87.18
Neural Network	3 hours	85.02	83.70	80.26	84.38	72.97	87.18

SVM was the most effective model, achieving the highest accuracy (85.53%) and recall (83.78%) with a relatively short tuning time (2.80s). Higher performance on linear models suggested a potential linear relationship between the features and presence of heart disease.