CHAPTER 3
COMPARING MACHINE LEARNING FOR HEART DISEASE PREDICTION

Aiman Safwan Sulaiman, Mohd Faaizie Darmawan, Mohd Zamri Osman, and Ahmad Firdaus Zainal Abidin

3.1 INTRODUCTION

Heart disease remains a major global health challenge, causing high rates of illness and death worldwide. It has been identified as one of the most lethal diseases affecting people of all ages globally (Baviskar et al., 2023). Various data mining and neural network techniques have been used to evaluate the severity of heart conditions. Despite progress in medical research and technology, heart disease continues to be a leading health issue globally, resulting in significant morbidity and mortality. Improving patient outcomes and reducing healthcare burdens are crucial, achievable through early detection and prompt intervention. Methods such as genetic algorithm (GA) (Jabbar et al., 2013), artificial neural network (ANN) (Durairaj & Revathi, 2015; Gavhane et al., 2018), decision trees (DT) (Maheswari & Pitchai, 2019; Ozcan & Peker, 2023), and K-nearest neighbour algorithm (KNN) (Jabbar et al., 2013; Rahmat et al., 2021) have been employed for this purpose. Congestive heart failure is the most prevalent type of heart disease and is considered the least severe. In human anatomy, veins are responsible for carrying blood

back to the heart. Additional factors contributing to heart disease include malfunctioning cardiac valves, which can result in heart failure. Congestive heart failure, stroke, dilated cardiomyopathy and angina pectoris are among the conditions most associated with heart disease. Therefore, monitoring biomarkers for heart disease and consulting with medical experts is necessary (Sarra et al., 2022).

Heart disease remains the leading cause of death worldwide, as reported by the World Health Organisation (WHO). In 2019, approximately 17.9 million people died from heart-related conditions, accounting for 32% of all global fatalities. Early detection and prognosis are crucial for extending patient lifespan. However, a major challenge in this area is the shortage of resources, including the limited availability of doctors and radiologists in developing or low-income countries. This results in many diagnoses occurring at advanced stages of the disease, creating a significant bottleneck (Gupta et al., 2019).

Researchers have increasingly turned to classification models powered by machine learning (ML) and data analysis models to forecast the likelihood of heart disease. According to previous research, combining artificial intelligence (AI) models, patient records, and clinical decision support with various ML classification and feature selection approaches can help predict multiple diseases, according to previous research (Satish et al., 2019). This study's classifier models are support vector machine (SVM), random forest (RF), and DT. The SVM model is a supervised learning algorithm used for classification and regression analysis. Due to its ability to manage both linear and nonlinear relationships, SVM is a viable option. With heart disease prediction frequently involving intricate interactions between patient attributes, SVM can identify the optimal hyperplane for classifying the data. In cardiac disease, SVM is a valuable model for capturing complex relationships and making accurate predictions.

Random forest model (RF) (Breiman, 2001) is an ensemble learning model that employs a bagging strategy; bootstrap (with replacement sampling) is used to generate the sample set selection of each classifier, the unselected data is used as the test set of the classifier, and a random

extraction mechanism is added to the feature selection. RF is selected because it excels at managing complex data relationships and interactions. RF can effectively capture nonlinear patterns and provide accurate heart disease predictions with a dataset containing various patient attributes and diagnostic results. Combining multiple decision trees can improve prediction accuracy and help prevent overfitting.

Decision trees (DT) model extracts the most relevant features while ignoring the irrelevant ones. DT are included in the study to gain interpretability and understand the importance of individual features in predicting heart disease. By analysing the decision-making process of DT, it is possible to gain insight into which patient characteristics are crucial in determining the presence of cardiac disease. By incorporating these classifier models, the study seeks to capitalise on their respective strengths to accurately predict heart disease in individuals.

3.2 DATA COLLECTION AND PROCESSING

The heart disease dataset was sourced from the open.ml website, which offers a variety of free datasets. It includes approximately 270 data samples and 13 features, as listed in Table 3.1, covering both data inputs and targets used for model development. In ML, training and testing datasets are vital for evaluating performance, detecting overfitting, tuning hyperparameters, and selecting the best model.

The dataset is split into two main sections: one for training the model and the other for testing its performance. Approximately 80% of the data, totalling 216 samples, is allocated for training, while the remaining 20%, or 54 samples, is used for testing. To ensure representative samples in both parts, the *train_test_split* function from the Scikit-learn library is employed. This function randomly shuffles and divides the data based on the specified ratio. As a result of this randomisation, both the training and testing sets accurately mirror the overall data distribution. This method enables effective model training and ensures a fair, unbiased evaluation of its performance.