CHAPTER 1
INTRODUCTION

1.1. Background

Referring to the information from World Health Organization (WHO), heart disease has become the leading cause of millions of deaths globally [1]. Heart disease may show up in individuals due to behavioral risk factors such as high blood pressure, diabetes, and many others [1]. Many different risk factors can impact and complicate heart disease diagnosis and result in an inaccurate or delayed diagnosis. Sometimes, experts encounter trouble diagnosing heart disease and require accurate tools and machines that can consider all the different risk factors and give an accurate result in a short time [2].

With the need for accurate results in a short time with different risk factors, unwanted bias, and data error, the need for machine learning methods will be necessary. Machine learning can provide high accuracy and a quick process in classification. Getting high accuracy in prediction is essential as it can lead to proper protection, especially in health.

In earlier research, the researchers solved the problem by comparing the algorithms, combining several algorithms, and using ensemble methods in the algorithms the researchers used. Some researchers achieve <80% accuracy which is not good enough when classifying heart disease. While some others researchers achieve >90% accuracy. With this various accuracy gained, there is a gap/limitation where almost all the researchers used the same algorithm in their research. So, other algorithms that have better performance than algorithms that have been researched are needed.

In this research, Extreme Gradient Boosting (XGBoost) algorithm which belongs to the ensemble method is proposed to predict Heart Disease. This algorithm is built on top of Lasso Regression and Ridge Regression that can avoid overfitting to happen. There are several researchers concluded in a paper that feature selection may be used in detecting heart disease as it can improve the accuracy of the model and remove redundant and irrelevant features [3]–[5]. So, feature selection also needs to be evaluated if it can boost the algorithm accuracy more in diagnosing heart disease.

1.2. Problem Formulation

In particular, this study will examine two main research questions:

1. Will Extreme Gradient Boosting Algorithm perform better than Adaptive Boosting Algorithm (AdaBoost) in detecting heart disease?
2. Does feature selection affect the performance of Extreme Gradient Boosting Algorithm in detecting heart disease?

1.3. Scope

This study focuses on finding out how well the Extreme Gradient Boosting Algorithm can perform better than Adaptive Boosting Algorithm in detecting heart disease and finding out whether feature selection can affect XGBoost perfomance in detecting heart disease using a limited dataset (approx. 303 data from Heart Disease Cleveland Dataset).

1.4. Objective

This study aims to know whether the Extreme Gradient Boosting Algorithm performs better in detecting heart disease or not and to know whether feature selection can affect the performance of XGBoost algorithm in detecting heart disease.