# Heart Disease Prediction using Machine Learning and Deep Learning

Tang Jia Hui[1], Lim Ka Li[2], Ooi Kah Choo[3]
Faculty of Information Science and Technolgy[1,2,3]
Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia

***Abstract*— Heart disease is one of the leading cause of death in humans, responsible for the death of millions worldwide. Many became a victim of this dreadful disease, although it is preventable because of their ignorance and unaware of being diagnosed with the disease. In the worst-case scenario, misdiagnosis of a deadly heart disease as other minor medical condition and the manual identification of heart disease are a delay for one to obtain a proper medical treatment for heart disease and at last resulted in death. An approach to overcome such tragedy is to apply machine learning (ML) and data mining in the diagnosis of heart disease which is a much faster and accurate technique. This research applied two ML algorithms - Naïve Bayes and Logistic Regression and a deep learning model - Multilayer Perceptron (MLP) to predict the presence and absence of heart disease based on a set of relevant heart disease determinator. The SelectKBest using ANOVA F-value feature selection technique was used to select the minimum number of determinator in predicting heart disease by each prediction model whilst maintaining a high prediction accuracy. Comparative performance analysis of the three models depicts that Naïve Bayes upon feature selection is the best predictive model based on the accuracy, F1 score, precision, recall and AUC score of 87.50%, 88.70%, 88%, 88% and 91.9% respectively, the order is then followed by the Logistic Regression and MLP model. In short, this justifies that Naïve Bayes model is the best prediction model among the three in predicting heart disease. The code implemented can be found in https://colab.research.google.com/drive/1x1jLOrhPu5CPmvWeaZDG47sKajPRAelk?usp=sharing**

***Keywords*— *Machine Learning; Heart Disease; Classification; Prediction; Logistic Regression; Naïve Bayes; Multilayer Perceptron***

## I. INTRODUCTION

Heart disease or also known as cardiovascular disease (CVD) generally refers to heart conditions that specifically affect the blood vessels making it either narrowed or blocked. It comprises a set of heart-related issues such as arrhythmia, a heart rhythm abnormality, cardiomyopathy, a condition that affect the heart muscle and many more. Nowadays, heart disease has become one of the most common diseases that is found in various people especially elderly due to a variety of contributing variables, such as unhealthy diet, physical inactivity, diabetes, high blood pressure and others. Typically, heart disease is a type of disease that necessitates lifelong treatment and close monitoring since it is incurable and irreversible. Nevertheless, various approach such as medications, treatments, and lifestyle modifications has shown to be able to alleviate the symptoms of the heart disease and improve the patient's life expectancy [1]. Furthermore, heart disease or CVD should not be taken lightly as it is claiming the lives of an estimated 17.9 million people per year based on the reports published by the World Health Organization (WHO) [2]. Therefore, it is very important to get a proper diagnosis in the early stage of the disease so that we can prevent any permanent damages to the heart. Any negligence such as delayed diagnosis, inappropriate treatment or misdiagnosed will worsen the condition and lead to devastating effects which are fatal.

Even so, with the evolution of the technology at the moment, there have been myriad of resources and information that are waiting to be exploited. With the implementation of technologies such as artificial intelligent (AI) and big data in the medical field, a proper prediction model can be made from scratch to help the health care providers such as doctors to diagnosis heart disease in patient more accurately and efficiently. This can be achieved through the application of machine learning algorithm on the data that was collected from the hospital. Machine learning (ML) and deep learning is a type of data analysis that automates the creation of analytical models. It is a field of artificial intelligence that is based on the premise that computers can learn from data, recognize patterns, and make judgments with little or no human input. Therefore, with the patient's medical history as input for the classifier, the model will then help in the predicting the condition of the patient. The introduction of AI has brought various changes in the field of medical as it has significantly help improve the accuracy and efficiency of diagnosis and treatment across various specializations [3]. This is important because heart disease is often being misdiagnose with other health related disease due to the similarity in symptoms. Based on a study that was done by a group of doctors, the rates of misdiagnosed heart failure ranged from 16.1% in hospitals to 68.5% when patients were referred to specialists by a general practitioner [4]. Misdiagnosis is something that everyone especially doctors would like to avoid because an incorrect or delayed diagnosis is the leading cause of serious injury among medical errors, accounting for one-third of malpractice cases that result in death or permanent disability [5]. Thus, major steps must be made to enhance the accuracy of diagnosis in order to reduce serious consequences from medical error.

In the past, there are also various studies related to prediction of heart disease with the application machine learning that have been carried while using the same dataset as this research. In particular, a research that focuses on the prediction of heart disease using a combination of machine learning and deep learning has been done by Bharti et al [6]. In this study, the researcher has emphasis the use of various machine learning

algorithm which including Logistic Regression, K-Nearest Neighbor (KNN) to focus on neighbor selection, Decision Tree, Random Forest, Support Vector Machine (SVM) to check and handle the data's high dimensionality and lastly the XGBoost classifier which is the combination of ensemble method and Decision Tree method. Meanwhile, the architecture that was used for the deep learning algorithm is the Artificial Neural Network (ANN). There are three approach that was used in the studies, for the first approach, the dataset is utilized directly for classification, for the second approach, data with feature selection is handled and there are also no outliers detected, finally for the third approach, the dataset was normalized while taking into the outliers and feature selection is removed. Based on the result, it has been shown that SVM (84.09%) have the highest accuracy for the first approach, Random Forest (88%) has the highest accuracy for the second approach while for the third approach, the deep learning algorithm, ANN have achieved an accuracy of 94.2% which is higher than any ML model. Moreover, other study also includes using feature selection techniques on classification algorithms to predict heart disease. Kaushalya, Gapar and Johar has conducted an experimental assessment of the performance of models generated using classification algorithms and suitable characteristics chosen using various feature selection procedures [7]. Ten feature selection technique and six classification approaches were used to analyze the results of exploratory analysis. The result produced shows that the combination of the backward feature selection technique and Decision Tree classifier has the highest accuracy at 88.52%. With that being said, it is safe to that either ML algorithms or deep learning will be very helpful in prediction of heart disease.

In conclusion, the main objective of this research is to develop three different machine learning and deep learning model which are able to classify heart disease and then compared them in order to get the best model with the highest accuracy. The ML classifier that we will be using includes Naïve Bayes, Logistic Regression while the deep learning model used is the Multilayer Perceptron. Other than that, the feature selection, ANOVA will also be applied on both the ML model. The dataset used in this research are obtained through Kaggle which originate under the index of heart disease datasets from the UCI Machine Learning Repository. It will consist of 918 observations with a total of 12 columns. Heart disease is a serious health issue that must be addresses immediately, in order to reduce the societal effect of these disease, the healthcare industry must reinforce the way they are dealt with, and this includes investing in artificial intelligent. Machine learning plays an important role in analysis medical data whereby it has so far been useful in assisting in the decision making and prediction of enormous amounts of data generated. So, with the development and evaluation of the predictive models, we hoped to find out the most outstanding one that could be used to diagnose cardiovascular disease more efficiently.

## II. LITERATURE REVIEW

With the advancement of medical science and machine learning, much work has been done to predict heart disease. In recent years, various experiments have been conducted, with a variety of machine learning algorithms to predict heart disease, as mentioned below.

Prasanta Kumar Sahoo and Pravalika Jeripothula conducted research on developing a reliable system that can work efficiently and accurately predicting the heart failure. This work was performed using five algorithms namely Support Vector Machine, Naïve Bayes, Logistic Regression, Decision Tree and K-Nearest Neighbors. The best result was determined as the Support Vector Machine which had the highest accuracy among the implemented algorithms [8]. Apurb Rajdhan and et. al., proposed work that use machine learning techniques such as Naïve Bayes, Decision Tree, Logistic Regression and Random Forest to predict the likelihood of heart disease and classify patient risk levels. This research was conducted to analyze and compare the performance of several machine learning algorithms. The best result was determined as the Random Forest algorithm which had the highest accuracy among the implemented algorithms [9]. Avinash Golande and et. al., investigated various machine learning techniques for heart disease classification problem.

The accuracy of the Decision Tree, K-Nearest Neighbors, and K-Means algorithms that can be used for classification is compared in this work. Research has proved Decision Tree has the better performance among the other machine learning algorithms, and the combination of few methodologies and adjusting the parameter can increase the performance of the model [10]. Nagaraj M Lutimath, et al., used Naive Bayes classification and Support Vector Machine to predict the prognosis of heart disease. Mean Absolute Error, Sum of Squared Error and Root Mean Squared Error were used as the evaluation metric to determine the performance of the machine learning algorithms. Support Vector Machine was found to be superior to Naïve Bayes in terms of accuracy [11].

Rajesh N et al. conducted study in order to develop a better approach and algorithm for heart disease prediction using a machine learning algorithm. The machine learning algorithm is analyzed on the dataset used based on risk factors. Naïve Bayes, decision trees and a combination of algorithms are employed to predict the heart disease. They found that decision tree produces more accurate result compared to Naïve Bayes with the increasing of dataset, whereas naïve bayes provide more accurate result compared to decision tree when the dataset is small [12]. Fahd Saleh Alotaibi has created a model for predicting heart failure that compares five machine learning algorithms. The accuracy of the Decision Tree, Logistic Regression, Random Forest, Naïve Bayes, and Support Vector Machines algorithms that can be used for classification is compared in this work. Among all of the algorithms, Decision Tree has emerged as the most accurate [13].

After reading previous studies, the basic goal behind our proposed system is to build a heart disease prediction model to classify the risk of having heart disease by using 2 different machine learning models and 1 deep learning model. We compared the accuracy, precision, recall and f-measure scores for the algorithms which are Naïve Bayes, Logistic Regression and Multi-Layer Perceptron to find the best classification algorithm for heart disease prediction.

*Table 1: Metadata and description of attributes and target variable in the heart disease dataset prior preprocessing*

| Column Name | Description / Meaning | Unique Values / Range | Data Level | Data Type |
|---|---|---|---|---|
| Age | Patient's age (year-old) | Minimum: 28; Maximum: 77 | Ratio | Int64 |
| Sex | Patient's gender | M: Male; F: Female | Nominal | Object |
| ChestPainType | Type of chest pain experienced by the patient | ASY: Asymptomatic; NAP: Non-angina pain; ATA: Atypical angina; TA: Typical angina | Nominal | Object |
| RestingBP | Patient's resting blood pressure (mmHg) | Minimum: 0; Maximum: 200 | Ratio | Int64 |
| Cholesterol | Patient's serum cholesterol (mm/dl) | Minimum: 0; Maximum: 603 | Ratio | Int64 |
| FastingBS | Patient's blood sugar level when on fasting (mg/dl) | 0: less than 120; 1: more than 120 | Ordinal | Int64 |
| RestingECG | Patient's resting electrocardiogram result | Normal: Normal ECG; ST: ST-T wave abnormality (ST-Slope/Elevation/Depression > 0.05mV); LVH: Probable left ventricular hypertrophy by Estes' criteria | Nominal | Object |
| MaxHR | Patient's maximum measured heart rate | Minimum value: 60; Maximum value: 202 | Interval | Int64 |
| ExerciseAngina | Patient's angina (chest pain induced by exercise) | Y: Yes; N: No | Nominal | Object |
| Oldpeak | The ST depression induced by exercise relative to rest. | Minimum value: -2.6; Maximum value: 6.2 | Interval | Float64 |
| ST_Slope | The slope of the peak exercise ST segment | Up: Upsloping, Flat: Flat, Down: Downsloping | Ordinal | Object |
| **HeartDisease**** | Presence of Heart Disease | 0: Normal/ No heart disease; 1: Has heart disease | Nominal | Int64 |

****target variable**

## III. RESEARCH METHODOLOGY

This research is segmented into 3 phases – Feasibility study, Data Preparation, Modelling and Evaluation where each of phase has its subprocesses as depicted in Diagram 1. A thorough analysis, visualization and understanding of heart disease dataset is performed, followed by the machine learning and deep learning methods to be applied in predicting the chance of having heart disease, namely Logistic Regression, Naïve Bayes and Multilayer Perceptron.

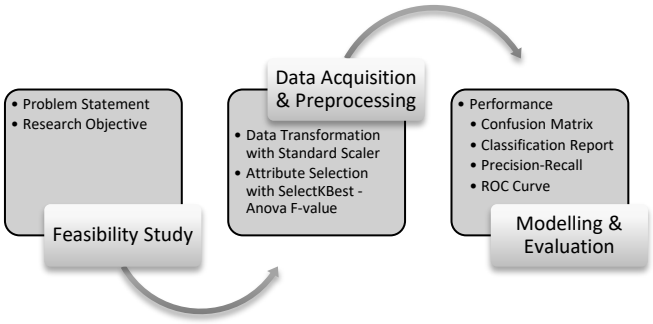



*Figure 1: Heart Disease Prediction Research Workflow*

### A. Data Preparation

*1) Data Acquisition:* The analysis is performed on a publicly available medical dataset for heart disease which combines 11 common clinical attributes from 5 different heart disease dataset collected from the healthcare databases of Cleveland, Hungarian, Switzerland, Long Beach VA and the Stalog Heart dataset. This heart disease dataset is by far the largest heart disease dataset available with crucial attributes that are significant in predicting the chance of having detrimental heart disease. The data in the heart disease dataset is organized into a total of 918 unduplicated medical records of patients with and without heart disease. Each record has 12 columns, made up of the common clinical attributes and a target column. Table 1 shows the 11 clinical attributes and the target attribute, along with their meaning and purpose in the research of heart disease, its level of data and data type.

*2) Data Preprocessing:* There is no missing data in the heart disease dataset and all attributes are in their correct data types. However, there are outliers in the column Resting BP, Cholesterol, MaxHR, and Old Peak. However, there are not addressed by imputing with median nor by eliminating records that has outliers with respect to the addressed columns as these extreme values highlight the abnormality in a patient's health and may have a high contribution in determining if one has a higher risk of having heart disease. Two data transformation were performed, namely conversion of categorical data to numerical variables via one hot encoding and data standardization with Standard Scaler. The conversion of categorical to numerical data via one hot encoding technique is performed on the column Sex, ChestPainType, FastingBS, RestingECG, ExerciseAngina and ST_Slope, causing each unique value in these column to be a column itself with values of 0's and 1's. This step is required because machine learning and deep learning models works only with numerical data or categorical variables that are encoded to numbers as that is the sole type of information they understood. Hence, such conversion enhances the quality and usability of data for model training. Next, data standardization was applied on the column Age, RestingBP, Cholesterol, MaxHR and Oldpeak because there is a significant difference in the scale of values in the aforementioned column. For instance, the range of Cholesterol is (0, 603) whereas it is (-2.6, 6.2) for Oldpeak. This may induce biases where attributes with a larger range of values plays a

stronger role in the heart disease prediction as compared to those with smaller range of values. Therefore, data standardization aims to overcome the bias problem by ensuring all continuous numerical values to be at a similar scale, as well as speed up the convergence of gradient descent when neural network such as Multilayer Perceptron is used as the heart disease predictive model.

The data preprocessing phase ends with feature selection prior model development and training as selecting only useful attributes in training not only resulted in a better heart disease prediction, but it reduces training time, complexity and interpretation. The approach used in selecting important features is a univariate feature selection technique that selects k top features according to the k highest scores obtained [14]. It then removes all irrelevant attributes, leaving only the k highest scoring features to be used in model training. Such selection is dependent on the univariate statistical tests in which each attribute or feature is compared to the target variable to identify if there exists a statistically significant correlation between them (ANOVA process) whilst ignoring other feature at that moment. At last, the attributes chosen for model training differ for all three predictive models due to their varying k value. However, this ensures all three predictive models developed achieve their maximum prediction performance.

Table 2 shows the attributes (upon data transformation and standardization) used by each predictive model developed to achieve maximum performance in heart disease prediction.

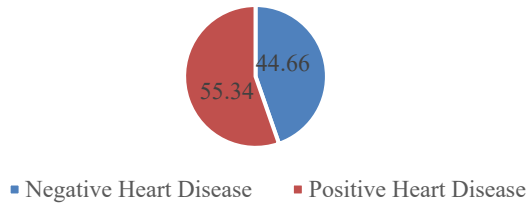
*Table 2: Attributes selected for model training with respect to each machine learning or deep learning model*

| Predictive Model | k | Attributes |
|---|---|---|
| Multilayer Perceptron | 20 | Age, RestingBP, Cholesterol, MaxHR, Oldpeak, sex_F, sex_M, chestpain_ASY, chestpain_ATA, chestpain_NAP, chestpain_TA, restecg_LVH, restecg_Normal, restecg_ST,, exerciseagina_N, exerciseagina_Y, stslope_Down, stslope_Flat, stslope_Up, FastingBS |
| Logistic Regression | 9 | MaxHR, Oldpeak, sex_F, chestpain_ASY, chestpain_ATA, exerciseagina_N, exerciseagina_Y, stslope_Flat, stslope_Up |
| Naïve Bayes | 11 | MaxHR, Oldpeak, sex_F, sex_M, chestpain_ASY, chestpain_ATA, exerciseagina_N, exerciseagina_Y, stslope_Flat, stslope_Up, FastingBS |

## B. *Descriptive Analysis*

Figure 2 reports on the statistical information of the total number of heart disease records that the patient is diagnosed with and without heart disease. The conclusion drawn is that the majority of the heart disease record are patients with heart disease (55.34%) and those without heart disease occupy 44.66% of the dataset. As the number of records of patients with and without heart disease are almost equal, the dataset is claimed to be a balanced dataset and the risk of introducing bias in the heart disease prediction made by the predictive model is low because the predictive model learns the data pattern of a patient with and without heart disease equally.

Figure 2: Total number of medical heart disease records with respect to 2 prediction classes

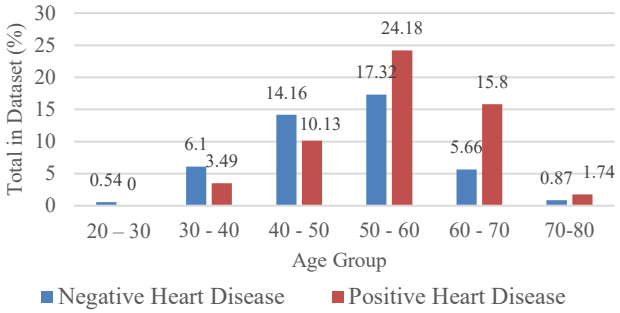



Next, table 3 depicts the statistical analysis on each attribute as an approach to gain insights on the heart disease medical records, along with support graphical charts to visualize the role and correlation of each attribute with the final prediction of a patient to have heart disease or otherwise. Additional reference on the correlation between attributes are depicted in the correlation heatmap.


*Figure 3: Correlation heatmap between quantitative variables*

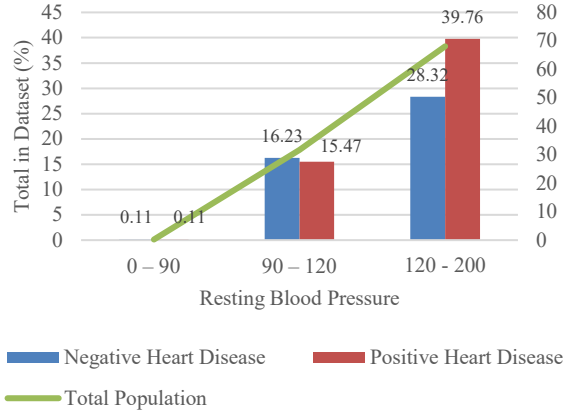
The age attribute is binned into 6 classes and as the age passed to be 50-year-old and above, the risk of having heart disease increases as compared to those with age lower than 50-year-old. This shows a direct proportional correlation of attribute age with the chance of having heart disease (*Refer Figure 4*).

Figure 4: Percentage of patients with and without heart disease based on age group



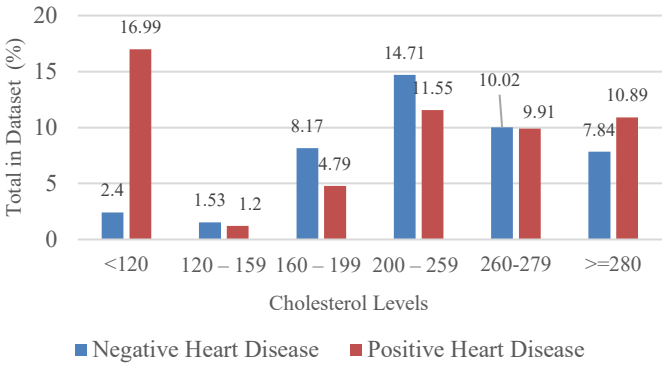
Majority of the patients have a high resting blood pressure (RestingBP), which is approximately 3 times the number of patients with normal or moderately high resting blood pressure. Patient with high RestingBP has a higher chance of being diagnosed with heart disease whereas those with moderately high RestingBP has a lower chance (*Refer Figure 5*) as justified in [15].

Figure 5: Percentage of patients with and without heart disease based on resting blood pressure




The cholesterol attribute shows a unique statistical pattern where the majority of the patients with very low cholesterol level have heart disease but those with normal to higher cholesterol has lower chance of having heart disease (*Refer Figure 6)*. Such occurrence may due to the build-up of plaques (atherosclerosis) in the coronary arteries that reduces blood and nutrients supply to cardiac muscles despite having a normal cholesterol level [16]. This eventually leads to coronary heart disease, angina, carotid artery disease and more [17].

Figure 6: Percentage of patients with and without heart disease based on resting blood pressure




The majority of patients have maximum heart rate between 100 to 150 and those with maximum heart rate lower or at this range will have a higher chance of heart disease as compared to those with maximum heart rate more than 150 (*Refer Figure 7*). Such occurrence is associated with the age factor or there is a negative correlation between the two attributes (*Refer* Figure 8-9). Older individuals have a lower maximum heart rate as aging depresses the heart pacemaker's spontaneous electrical activity, responsible in the contraction and relaxation of cardiac muscles [18].

Figure 7: Percentage of patient with and without heart disease based on Maximum Heart Rate

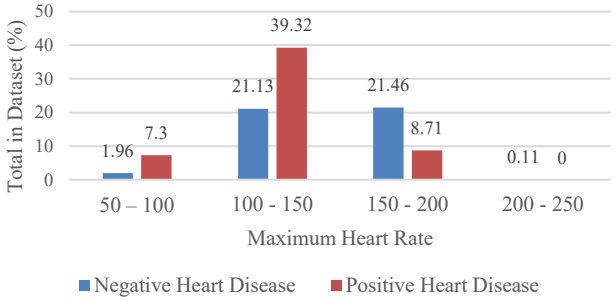



Figure 8: Percentage of patient with and without heart disease and has maximum heart rate less than 150

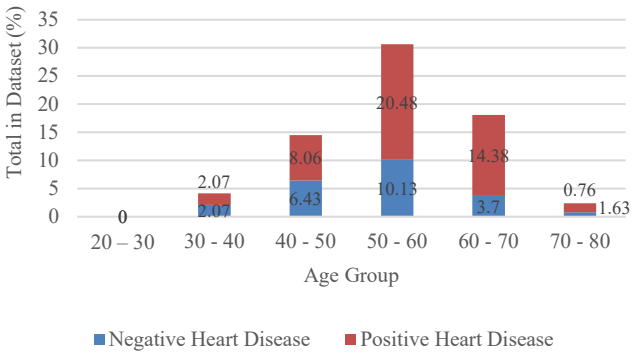
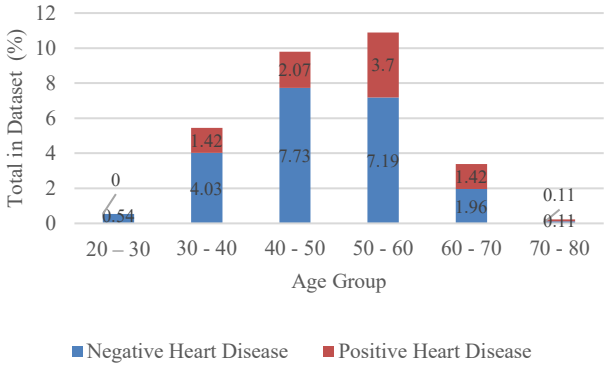



Figure 9: Percentage of patient with and without heart disease and has maximum heart rate more than 150

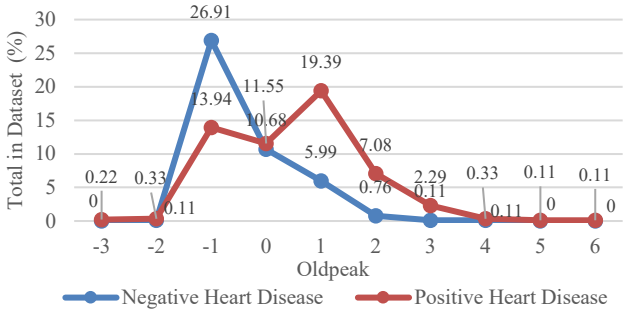



From the correlation heatmap (*Refer Figure 3)*, old peak has the highest positive correlation with Heart Disease, hence explains it is the attribute chose by all three predictive models in predicting heart disease. Most of the patients have old peak at the range of (-1.0, 0.0], occupying a grand total of 40.85% from the dataset. However, 65.87% of the 40.85% do not have heart disease. In contrast, 76.39% patients from the old peak of (1.0, 2.0] (or 19.39% of the dataset) accounts for the highest percentage of patients diagnosed with heart disease. Nevertheless, the general pattern of the oldpeak data depicted that any patients with oldpeak not within the range of (-1.0, 0.0] has a higher chance of having heart disease (*Refer Figure 10)*.

Table 3: Statistical analysis of all attributes in the heart disease dataset

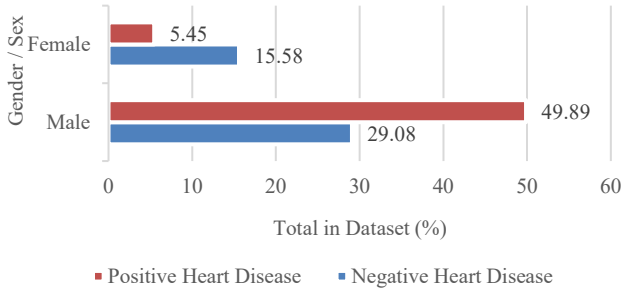
| Attribute | Range/Bins | Heart Disease (Count) | | Heart Disease by Range (%) | | Heart Disease by Population (%) | | Total Sample per Range | Total Sample per Population (%) |
|---|---|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 0 | 1 | 0 | 1 | | |
| Age | 20 – 30 | 5 | 0 | 100 | 0 | 0.54 | 0 | 5 | 0.54 |
| | 30 - 40 | 56 | 32 | 63.64 | 36.36 | 6.1 | 3.49 | 88 | 9.59 |
| | 40 - 50 | 130 | 93 | 58.3 | 41.7 | 14.16 | 10.13 | 223 | 24.29 |
| | 50 - 60 | 159 | 222 | 41.73 | 58.27 | 17.32 | 24.18 | 381 | 41.5 |
| | 60 - 70 | 52 | 145 | 26.4 | 73.6 | 5.66 | 15.8 | 197 | 21.46 |
| | 70 - 80 | 8 | 16 | 33.33 | 66.67 | 0.87 | 1.74 | 24 | 2.61 |
| RestingBP | 0 – 90 | 1 | 1 | 50 | 50 | 0.11 | 0.11 | 2 | 0.22 |
| | 90 – 120 | 149 | 142 | 51.2 | 48.8 | 16.23 | 15.47 | 291 | 31.7 |
| | 120 - 200 | 260 | 365 | 41.6 | 58.4 | 28.32 | 39.76 | 625 | 68.08 |
| Cholesterol | <120 | 22 | 156 | 12.36 | 87.64 | 2.4 | 16.99 | 178 | 19.39 |
| | 120 – 159 | 14 | 11 | 56 | 44 | 1.53 | 1.2 | 25 | 2.72 |
| | 160 – 199 | 75 | 44 | 63.03 | 36.97 | 8.17 | 4.79 | 119 | 12.96 |
| | 200 – 259 | 135 | 106 | 56.02 | 43.98 | 14.71 | 11.55 | 241 | 26.25 |
| | 260 – 279 | 92 | 91 | 50.27 | 49.73 | 10.02 | 9.91 | 183 | 19.93 |
| | > =280 | 72 | 100 | 41.86 | 58.14 | 7.84 | 10.89 | 172 | 18.74 |
| MaxHR | 50 – 100 | 18 | 67 | 21.18 | 78.82 | 1.96 | 7.3 | 85 | 9.26 |
| | 100 - 150 | 194 | 361 | 34.95 | 65.05 | 21.13 | 39.32 | 555 | 60.46 |
| | 150 - 200 | 197 | 80 | 71.12 | 28.88 | 21.46 | 8.71 | 277 | 30.17 |
| | 200 - 250 | 1 | 0 | 100 | 0 | 0.11 | 0 | 1 | 0.11 |
| Oldpeak | (-3.0, -2.0] | 0 | 2 | 0 | 100 | 0 | 0.22 | 2 | 0.22 |
| | (-2.0, -1.0] | 1 | 3 | 25 | 75 | 0.11 | 0.33 | 4 | 0.44 |
| | (-1.0, 0.0] | 247 | 128 | 65.87 | 34.13 | 26.91 | 13.94 | 375 | 40.85 |
| | (0.0, 1.0] | 98 | 106 | 48.04 | 51.96 | 10.68 | 11.55 | 204 | 22.22 |
| | (1.0, 2.0] | 55 | 178 | 23.61 | 76.39 | 5.99 | 19.39 | 233 | 25.38 |
| | (2.0, 3.0] | 7 | 65 | 9.72 | 90.28 | 0.76 | 7.08 | 72 | 7.84 |
| | (3.0, 4.0] | 1 | 21 | 4.55 | 95.45 | 0.11 | 2.29 | 22 | 2.4 |
| | (4.0, 5.0] | 1 | 3 | 25 | 75 | 0.11 | 0.33 | 4 | 0.44 |
| | (5.0, 6.0] | 0 | 1 | 0 | 100 | 0 | 0.11 | 1 | 0.11 |
| | (6.0, 7.0] | 0 | 1 | 0 | 100 | 0 | 0.11 | 1 | 0.11 |
| Sex | M | 267 | 458 | 36.83 | 63.17 | 29.08 | 49.89 | 725 | 78.98 |
| | F | 143 | 50 | 74.09 | 25.91 | 15.58 | 5.45 | 193 | 21.02 |
| FastingBP | 0 | 366 | 338 | 51.99 | 48.01 | 39.87 | 36.82 | 704 | 76.69 |
| | 1 | 44 | 170 | 20.56 | 79.44 | 4.79 | 18.52 | 214 | 23.31 |
| ChestPainType | ASY | 104 | 392 | 20.97 | 79.03 | 11.33 | 42.7 | 496 | 54.03 |
| | ATA | 149 | 24 | 86.13 | 13.87 | 16.23 | 2.61 | 173 | 18.85 |
| | NAP | 131 | 72 | 64.53 | 35.47 | 14.27 | 7.84 | 203 | 22.11 |
| | TA | 26 | 20 | 56.52 | 43.48 | 2.83 | 2.18 | 46 | 5.01 |
| RestingECG | LVH | 82 | 106 | 43.62 | 56.38 | 8.93 | 11.55 | 188 | 20.48 |
| | Normal | 267 | 285 | 48.37 | 51.63 | 29.08 | 31.05 | 552 | 60.13 |
| | ST | 61 | 117 | 34.27 | 65.73 | 6.64 | 12.75 | 178 | 19.39 |
| ExerciseAngina | N | 355 | 192 | 64.9 | 35.1 | 38.67 | 20.92 | 547 | 59.59 |
| | Y | 55 | 316 | 14.82 | 85.18 | 5.99 | 34.42 | 371 | 40.41 |
| ST_Slope | Down | 14 | 49 | 22.22 | 77.78 | 1.53 | 5.34 | 63 | 6.86 |
| | Flat | 79 | 381 | 17.17 | 82.83 | 8.61 | 41.5 | 460 | 50.11 |
| | Up | 317 | 78 | 80.25 | 19.75 | 34.53 | 8.5 | 395 | 43.03 |

Figure 10: Percentage of patient with and without heart disease based on old peak



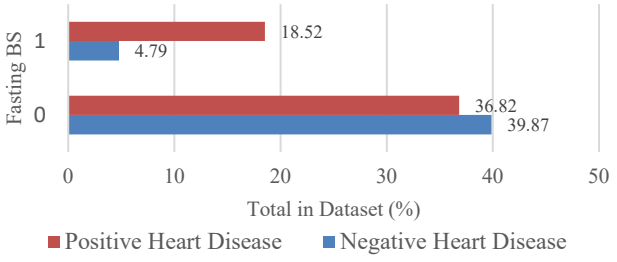

Majority of the medical heart disease records are male patient that accounts for 78.98% from the grand total whereas it is only 21.02% for females. However, males have a 63.17% chance to be diagnosed with heart disease (49.89% from the entire dataset) which is higher than females that possess a risk of 25.91% to have heart disease (5.45% from the entire dataset). Therefore, this concludes that females generally have a lower risk of heart disease as compared to males (*Refer Figure 11)*.

Figure 11: Percentage of patient with and without heart disease based on gender



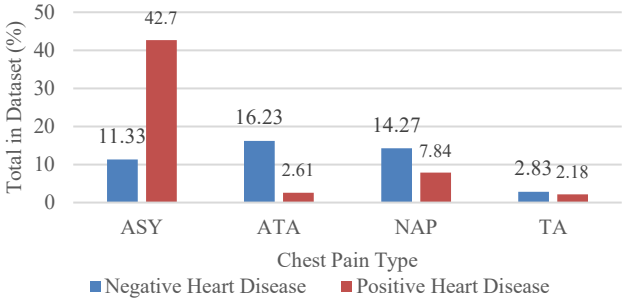

The majority of the patient has fasting blood sugar less than 120mg/dl (0), which is 76.69% of the grand total and most of them do not having heart disease (39.87%) while only 36.82% has heart disease. However, those who have fasting blood sugar higher than 120mg/dl will have a higher risk of having heart disease, which is 18.52% out of the 23.31% (Refer Figure 12).

Figure 12: Percentage of patient with and without heart disease based on fasting blood sugar



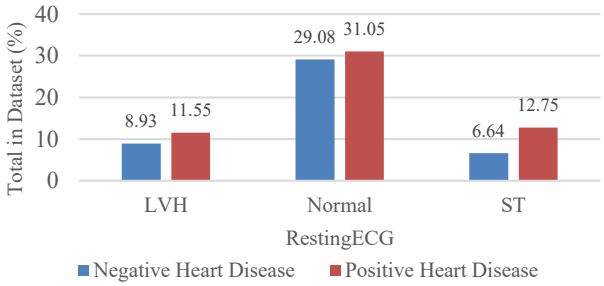

The common type of chest pain experienced by patients is asymptomatic whereby it comes without additional symptoms of having a heart disease. From the total percentage of asymptomatic cases, 79.03% of them (42.7% from the total in dataset) have heart disease while 20.97% does not have heart disease. The rest of the types of chest pain - non-angina pain, atypical angina, typical angina is uncommonly experienced by patients. Even though there are patients who have experienced the 3 aforementioned chest pains, but they do not have heart disease (*Refer Figure 13)*.

Figure 13: Percentage of patient with and without heart disease based on chest pain type



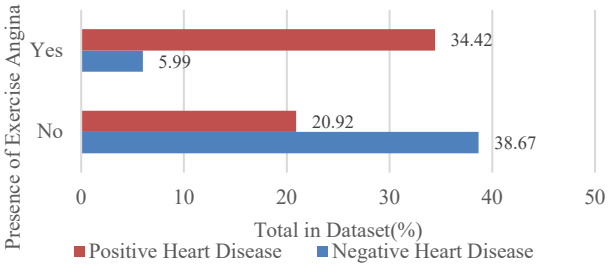

Most patients have a normal resting ECG reading (60.13%) while LVH and ST account for 20.48% and 19.39% of the grand total respectively. Regardless of which type of ECG result, the chance of having heart disease is almost similar as shown in Figure 14.

Figure 14: Percentage of patient with and without heart disease based on RestingECG




Most of the patients do not have exercise angina (59.59%) while only 40.41% of them have exercise angina. Those without exercise angina has a 35.1% (from the 59.59%) to have heart disease and 64.9% (from the 59.59%) to not have heart disease. Unfortunately, those with exercise angina has a higher chance to have heart disease, which is 34.42% from the grand total or 85.18% from the 40.41%. They only have a 14.82% chance from the 40.41% to not have heart disease (*Refer Figure 15)*.

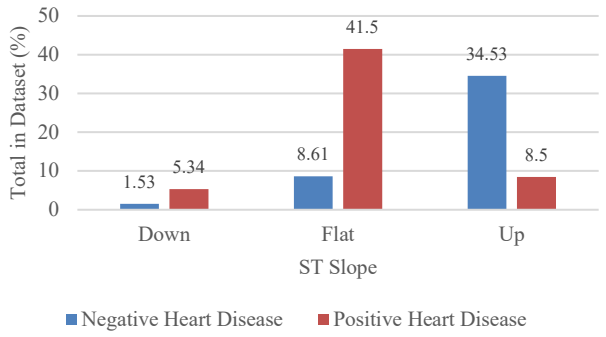
Figure 15: Percentage of patient with and without heart disease based on the presence of exercise angina




The majority of the patient has Flat ST slope (50.11%), followed by Up (43.03%) and Down (6.86%) slope. From the

total in flat ST slope, the risk of having heart disease is 82.83% (41.5% from the grand total of records) while having only 17.17% (8.61% from the grand total of records) to be free from heart disease. Even though 43.03% of patients experienced an Up ST slope, but most of them do not have heart disease, whereby 82.83% out of the 43.03% are free from heart disease. In contrast, although a small portion of patients have Down ST slope, but they will have a higher chance to be diagnosed with heart disease, which is 77.78% out of the 6.86% (*Refer Figure 16)*.

Figure 16: Percentage of patient with and without heart disease based on the ST Slope




## C. *Modelling & Evaluation*

*1) Model Construction and Testing:* Three prediction model developed - Multilayer Perceptron, Logistic Regression and Naïve Bayes will be trained with the training dataset and evaluated with a test dataset obtained from stratified shuffle split cross validator under random state 42 for reproducible results and with 10 folds. The ratio of split is 4:1. However, an additional spliting is required on the training dataset in order to create a development set used in the training of Multilayer Perceptron. This is to validate the model's generalization in learning to predict heart disease and provides insights if the model is overfitting, underfitting or generalize well.

*2) Learning Algoritms Concept and Specifications*

*a) Multilayer Perceptron, MLP:* It is a class of neural network with multiple layers, such as an input layer, one or more hidden layers and an output layer. These layers consists of nodes with learnable weights and biases that will be constantly updated during backpropagation after each feedforward step in order to improvise its classification or prediction performance. Feedforward step involves the nodes in learning to classify before producing the final output, which is either "have heart disease" represented by 1, and "do not have heart disease" represented by 0. Once MLP outputs its prediction, it will be compared with the ground truth (the true label corresponding to a single medical heart disease record) before computing the loss or error in their prediction. Lastly, the amount of updates required to be made on the weights and biases of the nodes through backpropagation is performed to minimize the total loss in its prediction. With sufficient training and proper optimization, a generalized MLP-based prediction model with decent heart disease prediction performance is obtained. Figure 17 depicts an the implemented MLP architecture.

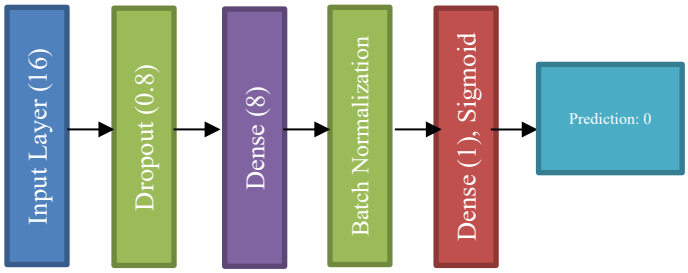



*Figure 17: The implemented MLP architecture that consists the basic layers, namely the Input layer, Hidden layer or Dense(8) and Output layer or Dense(1). Two additional layers, which are Dropout and Batch Normalization were added for regularization to prevent model overfitting. The final prediction shall be either 0 (no heart disease) or 1 (has heart disease).*

The MLP model developed uses sigmoid activation function (*Refer Figure 18)*in the output layer as it is tailored to be used in binary classification problem such as heart disease prediction. When the predicted output exceeded a specific threshold (normally 0.5), the predicted output is 1 (has heart disease) while any value lower than 0.5 produces output value 0 (no heart disease).

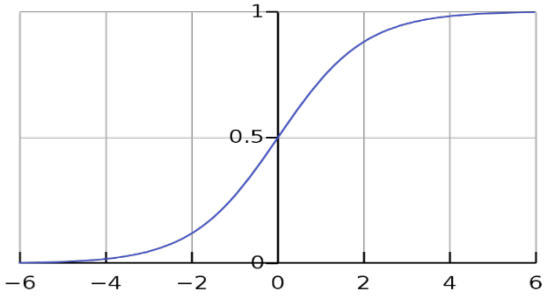



*Figure 18: Sigmoid Activation Function*

The concept of early stopping is another regularization technique used to avoid the overfitting of MLP model, hence securing the model's generalizability in heart disease prediction. The optimizer used in MLP model training is AdamW optimizer with learning weight and weight decay rate of 0.00025 to decouple the optimization of both learning rate and weight decay. Last but not least, the entire training process will run for 500 epochs with a batch size of 256.

*b) Logistic Regression:* It is a simple machine learninig algoritm commonly used in binary classification or prediction. It uses categorical or/and quantitative independent variables in learning before producing a categorical output. In the heart disease prediction, the expected output is "have heart disease" represented by 1, and "do not have heart disease" represented by 0. The concept of producing the predicted output in logistic regression is the same as the mechanism as explained in the sigmoid activation function.

*c) Naïve Bayes*: It is a machine learning classification algorithm based on Bayesian theorem, whereby the probability of having heart disease is calculated, given the occurrence or

presence of a set of specific characteristics or attributes. As the dataset consists of both categorical and continuous variables, a mixed Naïve Bayes is applied. The mixed Naïve Bayes is a combination of Gaussian Naïve Bayes algorithm that will be trained on continous variable and Categorical Naïve Bayes algorithm to be trained with categorical variables. Then, the prediction probabilities from the two models will be used to train the final Naïve Bayes model.

Last but not least, each of the three models will be trained on different numbers of attributes as stated in Table 2. The attributes shown in Table 2 were obtained via a univariate statistical test (Select K Best with ANOVA F-score) as described. Through the iteration of k=1 to k=20, the final k value that resulted in the highest performance is the finalized version of the respective predictive model.

*3) Model Evaluation:* The performance of each of the three prediction model was evaluated by its accuracy, precision, recall, F1 score measure. Additional visualization of results, such as confusion matrix, classification report, AUC-ROC score will be reported as well. Besides, evaluation on the multilayer perceptron model performance and generalizability is observed by comparing the convergence of training and validation loss, as well as the increment pattern of training and validation accuracy. A final testing accuracy of the multilayer perceptron model will be reported as well.

## IV. Results & Discussion

As shown in Table 4 below, the accuracy that was achieved by implementing the three different ML model is 85.33% for Multilayer Perceptron, 85.87% for Logistic Regression (with and without ANOVA), 85.87% for Naïve Bayes without feature selection and 88% for Naïve Bayes with ANOVA.

Table 4: Comparative Analysis of Performance for ML and Deep Learning Models

| Classifier | Accuracy (%) | Precision | Recall | F1-score |
|---|---|---|---|---|
| Multilayer Perceptron (MLP) + Feature Selection (ANOVA) | 85.33 | 87.88 | 85.29 | 86.57 |
| Logistic Regression | 85.87 | 86.00 | 86.00 | 87.00 |
| Logistic Regression + Feature Selection (ANOVA) | 85.87 | 87.00 | 88.00 | 88.00 |
| Naïve Bayes | 85.87 | 86.00 | 86.00 | 87.00 |
| Naïve Bayes + Feature Selection (ANOVA) | 87.50 | 88.00 | 88.00 | 88.70 |

In addition to that, the score for the ROC AUC curve and the Precision Recall graph is also shown in Table 5 below. Based on the table, it is indicated that the Naïve Bayes with Feature Selection model has the highest AUC score among the other ML classifiser. Since AUC is the measure of ability of a classifier to distinguish between classes, so this means that the Naïve Bayes with Feature Selection perform the best at differentiating between those with and without heart disease. On the contrary, between the Logistic Regression with and without the feature selection, the classifier without Feature Selection appear to have a higher ROC AUC (Logistic) score.

Table 5: Score for the ROC and Precision Recall Curve

| Classifier | AUC | ROC AUC (No Skill) | ROC AUC (Logistic) |
|---|---|---|---|
| Logistic Regression | 0.911 | 0.500 | 0.934 |
| Logistic Regression + Feature Selection (ANOVA) | 0.918 | 0.500 | 0.928 |
| Naïve Bayes | 0.911 | - | - |
| Naïve Bayes + Feature Selection (ANOVA) | 0.919 | - | - |

In general, the ROC curve displays the trade-off between sensitivity and specificity. The true positive rate (TPR) is plotted against the false positive rate to form a ROC curve (FPR). Therefore, classifier with curve that are closer to the top-left corner will typically indicate a better performance. For instance, if we were to compare the curve in Figure 19 and 20, we can see that the curve in Figure 19 is nearer to the top-left corner. As a result, this will indirectly justify the score in Table 5 whereby the Logistic Regression classifier perform better without the feature selection, additionally this mean more people will be diagnosed correctly with this model.

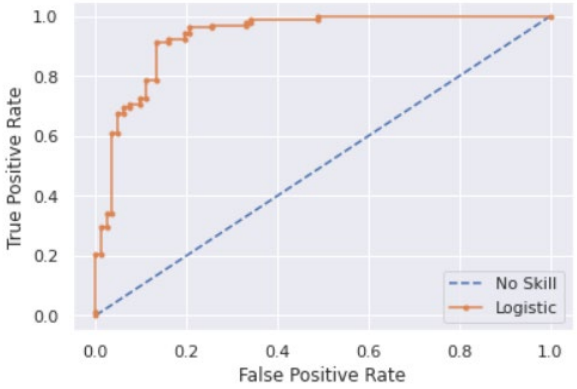



*Figure 19: ROC AUC Curve for Logistic Regression*

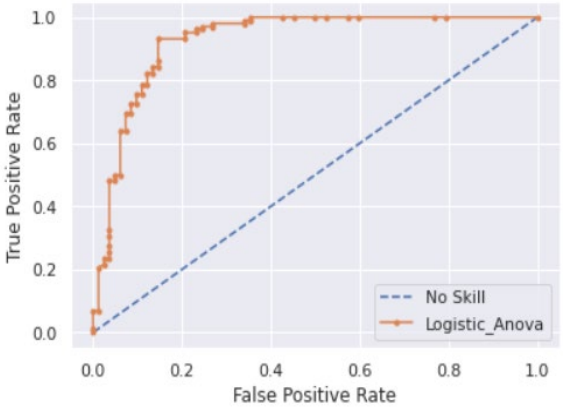



*Figure 20: ROC AUC Curve for Logistic Regression + Feature Selection (ANOVA method)*

Based on the result that is produced, we can rank the Naïve Bayes with Feature Selection as first, MLP as third and the rest of the model as second. The main reason for the performance

of MLP model being poorer than the other ML model is due to the fact that the sample size of the dataset is not large enough for a deep learning algorithm. Generally, deep learning require a large amount of data in order to perform better than other techniques. The neural network used in a standard deep learning models normally have a lot of nodes and layers, so this means that there will be a lot of parameters to be estimated. Therefore, if a deep learning model were to learn with a small data, it will likely to be overfit and thus producing disappointing result. Other than that, the absence of feature selection being applied in the MLP model might also be one of the factor in its ranking.

Meanwhile. If we were to compare the different between the ML model with feature selection and the ML model without it, it is shown that the implementation of Feature Selection, using SelectKBest with ANOVA F-value of the model has enhance the performance of the model. Although in term of Logistic Regression, the accuracy of the model remained constant with or without the feature selection, however the same accuracy was able to be obtained in the Logistic Regression with Feature Selection model even though the feature was reduced to only 9. On the other hand, for Naïve Bayes, the accuracy of the model improve slightly from 85.85% to 87.5% when the feature was reduced to 11. In this case, ANOVA has help improve the performance of the model by identifying the features that are independent of the target variable and removed them, thus, with only the best feature being remained, the performance will improve.

Besides that, even though feature selection is applied to both ML classifier, Naïve Bayes still perform better than Logistic Regression. This is mainly because Naïve Bayes is relatively immune to outliers while Logistic Regression is easily affected by outliers. As shown in the Figure 21 below, we can see that the presence of the outlier in the dataset is abundant. In general, Naïve Bayes assumes that all features are conditionally independent of each other while Logistic regression divides feature space in a linear fashion, and usually works well even when certain variables are correlated. This means that it has a simpler hypothesis function as compared to Logistic Regression. Since the interaction are not modelled, so some of the information in the data will be ignored. As a result, the Naïve Bayes produced will have a high bias model by nature and does not overfit easily. On the contrary, Logistic Regression is a higher variance model which seeks to model all the data, including the noise.

Overall, the model that perform the best in the prediction of heart disease based on the medical history is the Naïve Bayes with Feature Selection classifier which has an accuracy of 87.5%. Naïve Bayes is a machine learning algorithm that solve multi-class prediction problem using the Bayes Theorem. All predictors are assumed to be independent. To put it another way, this classifier considers that the presence of one feature in a class has no bearing on the presence of another. So, in this case, this classifier will be suitable for this dataset as it was initially created by combining different datasets that are already available independently. Furthermore, if the premise of feature independence remains true, Naïve Bayes can also outperform other models while using far less training data. As compared to other classifier, it has also performed better in term of accuracy, precision, recall and F1-score. In a nutshell, the result that was produced by the Naïve Bayes model is on par with its expectation, with the application of feature selection on the model and the compatibility of the model with the dataset, it has been proven that Naïve Bayes shown the most promising result among the three models.

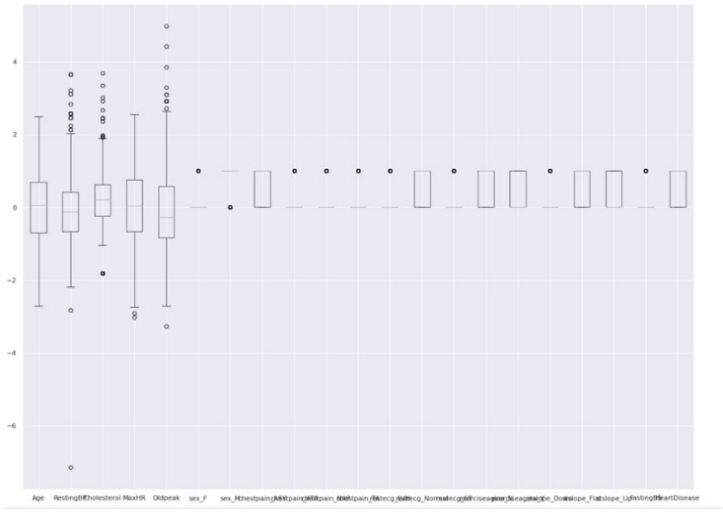


*Figure 21: Distribution of data using Boxplot*

## V. CONCLUSION

Heart failure is a long -term disorder in which the heart muscle is unable to pump enough blood to meet the body's needs for blood and oxygen. With the increasing number of deaths caused by heart failure, it is increasingly important to design systems that can predict heart failure effectively and accurately. A good data-driven system for heart failure prediction can improve the overall research and prevention process, allowing more people to have a healthy life.

The goal of the research is to discover the most effective machine learning methods for detecting heart failure. Using the Kaggle Heart Failure Prediction Dataset, this study analyzed the accuracy scores of the Naïve Bayes algorithm, Logistic Regression and Multi -Layer Perceptron to predict heart failure. According to the findings of this study, the Naïve Bayes algorithm with feature selection - SelectKBest with ANOVA F-value is the most effective method for predicting heart failure, with an accuracy score of 87.5%. Our findings support the premise that machine learning methods perform well in predicting heart failure because of their ability to exploit all available data and their intricate linkages.

The limitation of this research is the lack of demographic information in predicting heart failure. This information is crucial as factors such as safe housing, nutritious diets, and behavioral risks such as smoking, illegal drug use, or alcohol misuse may play a strong influence in a patient's susceptibility to heart diseases. Therefore, the suggested enhancement and future work on heart disease prediction is to specializes the prediction of heart disease based on demographic information by having a comprehensive dataset with the required

information. This shall help in tailoring the prediction better to one's ethnic, habits, geographical locations and more. Besides, the predictability of the models developed can be further enhanced to provide a more accurate prediction in order to assist health professionals in diagnosing a patient.

## ACKNOWLEDGMENT

The authors would like to thank Dr. Nor Samsiah Sani for her advice, patience and guidance throughout the work of this research.

### *A. Authors and Affiliation*

*Tang Jia Hui is a computer science undergraduate from the National University of Malaysia (UKM), specializes in data science studies. Her contribution to the research focuses on research methodology and coding of this study of heart disease prediction.*
*Contact: a176297@siswa.ukm.edu.my*

*Lim Ka Li is a computer science undergraduate from the National University of Malaysia (UKM), specializes in data science studies. Her contribution to the research focuses on the abstract, introduction, results and discussion of this study of heart disease prediction.*
*Contact: a176496@siswa.ukm.edu.my*

*Ooi Kah Choo is a computer science undergraduate from the National University of Malaysia (UKM), specializes in data science studies. Her contribution to the research focuses on literature review and conclusion of this study of heart disease prediction.*
*Contact: a176225@siswa.ukm.edu.my*

## REFERENCES

[1] Li Y, Pan AN, Wang DD, Liu X, Dhana K, Franco OH, Kaptoge S, Di Angelantonio E, Stampfer M, Willett WC, et al. Impact of healthy lifestyle factors on life expectancies in the US population. Circulation. 2018; 138:345–355.

[2] Riegel B, Moser DK, Buck HG, Dickson VV, Dunbar SB, Lee CS, Lennie TA, Lindenfeld J, Mitchell JE, Treat-Jacobson DJ, et al.; on behalf of the American Heart Association Council on Cardiovascular and Stroke Nursing; Council on Peripheral Vascular Disease; and Council on Quality of Care and Outcomes Research. Self-care for the prevention and management of cardiovascular disease: a scientific statement for healthcare professionals from the American Heart Association Am Heart Assoc. 2017; 6:e006997. doi: 10.1161/JAHA.117.006997

[3] Ahuja A. S. (2019). The impact of artificial intelligence in medicine on the future role of the physician. PeerJ, 7, e7702. https://doi.org/10.7717/peerj.7702

[4] Chun Wai Wong, Jacopo Tafuro, Ziyad Azam, Duwarakan Satchithananda, Simon Duckett, Diane Barker, Ashish Patwala, Fozia Z. Ahmed, Christian Mallen, Chun Shing Kwok, 2021, Misdiagnosis of Heart Failure: A Systematic Review of the Literature, Journal of Cardiac Failure, Volume 27, Issue 9, Pages 925-933, ISSN 1071-9164, https://doi.org/10.1016/j.cardfail.2021.05.014.

[5] Newman-Toker, D., Schaffer, A., Yu-Moe, C., Nassery, N., Saber Tehrani, A., Clemens, G., Wang, Z., Zhu, Y., Fanai, M. and Siegal, D. (2019) Serious misdiagnosis-related harms in malpractice claims: The "Big Three" – vascular events, infections, and cancers. Diagnosis, Vol. 6 (Issue 3), pp. 227-240. https://doi.org/10.1515/dx-2019-0019

[6] Rohit Bharti, Aditya Khamparia, Mohammad Shabaz, Gaurav Dhiman, Sagar Pande, Parneet Singh, 2021, "Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning", Computational Intelligence and Neuroscience, vol. 2021, Article ID 8387680, 11 pages. https://doi.org/10.1155/2021/8387680

[7] Kaushalya Dissanayake, Md Gapar Md Johar, 2021, "Comparative Study on Heart Disease Prediction Using Feature Selection Techniques on Classification Algorithms", Applied Computational Intelligence and Soft Computing, vol. 2021, Article ID 5581806, 17 pages. https://doi.org/10.1155/2021/5581806

[8] P. K. a. J. P. Sahoo, "Heart Failure Prediction Using Machine," *Journal of Machine Learning Research,* 15 December 2020.

[9] A. A. ,. M. S. ,. D. R. D. P. G. Apurb Rajdhan, Heart Disease Prediction using Machine Learning, vol. 9, International Journal of Engineering Research & Technology (IJERT), April 2020.

[10] P. K. T. Avinash Golande, Heart Disease Prediction Using Effective Machine, vol. 8, International Journal of Recent Technology and Engineering, 2019, pp. 944-950.

[11] C. C. S. P. Nagaraj M Lutimath, Prediction Of Heart Disease using Machine, vol. 8, International journal Of Recent Technology and Engineering, 2019, pp. 474-477.

[12] T. M. S. H. H. K. Rajesh Nichenametla, Prediction of Heart Disease Using Machine Learning Algorithms, 2018.

[13] F. Saled, Implementation of Machine Learning Model to Predict Heart Failure Disease, vol. 10, International Journal of Advanced Computer Science and Applications (IJACSA), 20119.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. T. B. Michel, O. Grisel, M. Blondel, P. W. R. Prettenhofer, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, Scikit-learn: Machine Learning in Python, vol. 12, Journal of Machine Learning Research, 2011, pp. 2825-2830.

[15] Z. M. Dongfeng, W. M. Weijing and L. M. Fang, "Association between resting heart rate and coronary artery disease, stroke, sudden death and noncardiovascular diseases: a meta-analysis," *CMAJ: Canadian Medical Association Journal,* vol. 188, no. 15, pp. E384-E392, 2016.

[16] American College of Cardiology, "Half of Patients with Ideal Cholesterol Have Underlying Heart Risks," CardioSmart American College of Cardiology, 14 December 2017. [Online]. Available: https://www.cardiosmart.org/news/2017/12/half-of-patients-with-ideal-cholesterol-have-underlying-heart-risks. [Accessed 6 January 2022].

[17] American Heart Association, "Atherosclerosis," American Heart Association, 6 November 2020. [Online]. Available: https://www.heart.org/en/health-topics/cholesterol/about-cholesterol/atherosclerosis. [Accessed 6 January 2022].

[18] University of Colorado Denver, "Why does maximum heart rate drop with age?," ScienceDaily, 14 October 2013. [Online]. Available: https://www.sciencedaily.com/releases/2013/10/131014155744.htm. [Accessed 6 January 2022].