SURVIVAL VERSUS NON-SURVIVAL PREDICTION AFTER ACUTE CORONARY SYNDROME IN MALAYSIAN POPULATION USING MACHINE LEARNING TECHNIQUE

NANYONGA AZIIDA

FACULTY OF SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR

2019

University of Malaya

SURVIVAL VERSUS NON-SURVIVAL PREDICTION AFTER ACUTE CORONARY SYNDROME IN MALAYSIAN POPULATION USING MACHINE LEARNING TECHNIQUE

NANYONGA AZIIDA

DISSERTATION SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE

INSTITUTE OF BIOLOGICAL SCIENCES
FACULTY OF SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR

2019

University of Malaya

UNIVERSITY OF MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: **NANYONGA AZIIDA**

Matric No: **SMA170030**

Name of Degree: **MASTER OF SCIENCE**

TITLE OF THESIS: **SURVIVAL VERSUS NON-SURVIVAL PREDICTION AFTER ACUTE CORONARY SYNDROME IN MALAYSIAN POPULATION USING MACHINE LEARNING TECHNIQUES.**

Field of Study:

BIOINFORMATICS

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date:

Subscribed and solemnly declared before,

Witness's Signature Date:

Name:

Designation:

University of Malaya

SURVIVAL VERSUS NON-SURVIVAL PREDICTION AFTER ACUTE CORONARY SYNDROME IN MALAYSIAN POPULATION USING MACHINE LEARNING TECHNIQUE

ABSTRACT

Prediction, identification, understanding and visualization of relationship between factors affecting mortality in ACS patients using feature selection and ML algorithms. Feature selection, classification and pattern recognition methods have been used in this research. From a group of 1480 patients drawn from the Acute Coronary Syndrome Malaysian registry, 302 people satisfied the inclusion criteria, and 54 variables were duly considered. Combinations of feature selection and classification algorithms were used for mortality prediction post ACS. Self-Organizing Feature (SOM) was used to visualize and identify the relationship and pattern between factors affecting mortality after ACS. Prediction models' performance criteria was measured using area under the curve (AUC) ranged from 0.62 to 0.795. The best model (RF) executed using 5 predictors (Age, TG, creatinine, Troponin and TC). Most model's performance plateaued using five predictors. The best performing model was compared with TIMI using an additional dataset that resulted in the ML model outperforming TIMI score (AUC 0.75 vs 0.60). Machine learning techniques for prediction and visualization of mortality related to ACS is presented in this study. The selected algorithms effectively show increase in prediction performance with decreasing features. Combination of ML prediction and visualization capabilities indicate effectiveness in predicting outcomes for clinical cardiology settings.

Keywords: Cardiovascular disease; Classification; Acute coronary syndrome; machine learning; feature selection

University of Malaya

RAMALAN ANTARA ORANG YANG TERSELAMAT DENGAN TIDAK TERSELAMAT SELEPAS SINDROM KORONARI AKUT PADA JUMLAH PENDUDUK MALAYSIA MENGGUNAKAN TEKNIK PEMBELAJARAN MESIN

ABSTRAK

Ramalan, pengenalan, pemahaman dan visualisasi hubungan antara faktor yang mempengaruhi kematian pesakit ACS menggunakan pemilihan ciri dan algoritma ML.Pemilihan ciri, klasifikasi dan kaedah pengenalan corak telah digunakan dalam kajian ini. Daripada kohort 1480 pesakit dari Sindrom Coronary Akut Malaysia, 302 pesakit memenuhi kriteria pemasukan dan 54 pembolehubah telah dipertimbangkan. Gabungan pemilihan ciri dan klasifikasi digunakan untuk pos ramalan mortaliti ACS. Ciri Penyusunan Sendiri (SOM) digunakan untuk memvisualisasikan dan mengenal pasti hubungan dan corak antara faktor-faktor yang mempengaruhi kematian selepas ACS. Kriteria prestasi model ramalan diukur dengan menggunakan kurva ciri operasi penerima (AUC) berkisar antara 0.62 hingga 0.795. Model terbaik (RF) dilakukan menggunakan 5 prediktor (Umur, TG, kreatinin, Troponin dan TC). Kebanyakan prestasi model diukur menggunakan Lima prediktor. Kami membentangkan pendekatan pembelajaran mesin untuk ramalan dan visualisasi mortaliti yang berkaitan dengan ACS. Algoritma yang dipilih menunjukkan peningkatan prestasi ramalan dengan mengurangkan bilangan pembolehubah. Gabungan ramalan ML dan keupayaan visualisasi boleh digunakan untuk ramalan hasil untuk tetapan kardiologi klinikal

Kata Kunci: Penyakit; Kardiovaskular; Akut Koronari Sindrom; Pembelajaraan Mesin; Pemilihan Ciri.

University of Malaya

ACKNOWLEDGEMENTS

In the name of Allah, the Most Merciful and the Most Gracious, I give praise and thanks to Him for giving me the strength to complete this research. This successful achievement could not have been without blessing from Allah Almighty and patience.

I feel immense pleasure in taking the opportunity to thank those who helped me in completing this thesis work. I am thankful to my supervisor Dr. Sorayya Malek for her immense support and patience; this thesis would have never been completed without her support. I am also very grateful to Dr. Sazzli for his valuable comments and directions, which added value to my work.

I gratefully acknowledge the funding received towards my masters from the Islamic development bank (IsDB) scholarship. To the IsDB group I greatly appreciate the support received through. I am also very grateful to all those at the IsDB office, especially Dr. Nazar Elhilali for his encouragement, supervisory role and his valuable input and others who were always so helpful and provided me with their assistance throughout my study.

To my two mothers; my mother, Zainabu and Mrs. Anna Bwetunge - the two most amazing women I have ever known. Thank you for inspiring me to ask many questions. Your love and kindness will always guide me. To my friend Hassan your just like a brother to me thanks for all your help. To my sister Habibah, Baby T and Aminah thanks.

To my father, Shaban, special and very grateful thanks must go to you. I pray to Allah and thank him, for cultivating in me the spirit of appreciation of the high value of education. Thank you for your prayers and for raising me to become a woman of high quality and value. Your words of encouragement and push for tenacity ring in my ears.

I doubt that I will ever be able to convey my appreciation fully to any of you.

University of Malaya

TABLE OF CONTENTS

University of Malaya

University of Malaya

University of Malaya

University of Malaya

LIST OF FIGURES

University of Malaya

LIST OF TABLES

University of Malaya

LIST OF SYMBOLS AND ABBREVIATIONS

Γ	:	Gamma
ACS	:	Acute coronary Syndrome
AMI	:	Acute Myocardial Infarction
AUC	:	Area under the Curve
ANN	:	Artificial Neural Network
CV	:	Cross-Validation
CD	:	Cluster Dendrogram
CVD	:	Cardiovascular Disease
CAD		Coronary Artery Disease
DT	:	Decision Tree
EN	:	Elastic Net
ECG	:	Electrocardiography
FN	:	False Negative
FP	:	False Positive
GA	:	Genetic Algorithm
GRACE	:	Global Registry of Acute Coronary Events
HEART	:	History, Electrocardiogram, Age, Risk factors, Troponin
HD	:	Heart Disease
IG	:	Information Gain
LR	:	Logistic Regression
LVQ	:	Learning Vector Quantization
MTYR	:	Number of selected variables
ML		Machine Learning
NTREE	:	Number of trees

University of Malaya

NSTEMI	:	Non-ST Segment Elevation Myocardial Infarction
RFE	:	Recursive Feature Elimination
RF	:	Random Forest
SVM	:	Support Vector Machine
STE-ACS	:	ST-Elevation Acute Coronary Syndrome
STEMI	:	ST Segment Elevation Myocardial Infarction
SBS	:	Sequential Backward Selection
TIMI	:	Thrombolysis in Myocardial Infarction
UA	:	Unstable Angina
WHO	:	World Health Organization

University of Malaya

LIST OF APPENDICES

University of Malaya

CHAPTER 1: INTRODUCTION

1.1 Background of the Study

Acute coronary syndromes (ACS) are clinical symptoms that are consistent with acute myocardial ischemia that comprises of clinical cases like ST elevation myocardial infarction (STEMI), non-ST elevation myocardial infarction (NSTEMI), and unstable angina (UA) (Hamilton et al. 2013; Kumar & Cannon, 2009). The supply of blood to the heart muscle cells is accomplished through two coronary arteries. Once these are blocked, the heart suffers from ischemia and if this obstruction is prolonged, heart cells die a state known as myocardial infarction (MI) (Dohare et al., 2018)

ACS occurs when the supply of blood is blocked or insufficient causing damage to the heart muscles. The rupture of an atherosclerotic plaque causing an incomplete or complete coronary artery blockage commonly affects it.

ACS is among the leading cause of mortality worldwide and in USA; about 1.36 million hospitalizations are presented with ACS alone (Castro-Dominguez et al., 2018; Kumar & Cannon, 2009). In Malaysia, 20-25% of all deaths in public hospital are attributed to coronary artery disease (CAD) (Hoo et al., 2016).

Various studies have been conducted across the world to get an insight of the risk and severity of ACS using conventional statistical approaches such as Thrombolysis in Myocardial Infarction (TIMI) score and the Global Registry of Acute Cardiac Events (GRACE) scores. However, these approaches have limitations, as there are very rigid. The limitation of these scores is possible loss of information due to fixed expectations on data performance and requirement to preselect features during the development stage (Shouval et al., 2017). It is important to recognize the most significant features affecting mortality rate in ACS patients in order to achieve a reliable and effective clinical

University of Malaya

diagnosis. This is also important in development of medical decision support tools linked with clinical and laboratory measures in order to decrease the mortality rate and monetary costs related with ACS. Mortality prediction related to ACS involves multiple features or variables where non-linear modeling methods or machine learning (ML) methods have the necessary flexibility to construct classifiers with good predictive performance. Compared with statistical approach, ML models are not pre-determined instead ML models are determined by underlying relationship, interactions and patterns of the data that allows discovery of additional knowledge. ML methods decrease the extent of human involvement necessary in fitting predictive models. ML methods comprise of automatic feature selection that allows manipulation of large numbers of predictors and does not require underlying assumptions regarding the relationship between input features and output (Chen & Ishwaran, 2012). Achieving highest performance accuracy and selecting smallest or optimum numbers of features are essential in optimizing ML classification algorithms performance. Hence, feature selection plays an important role in ML methods development. Feature selection methods can be categorized into embedded, filter, and wrapper methods, subjected to the ML classification algorithms used (Saeys et al., 2007; Salappa et al., 2007; Guyon & Elisseeff, 2003). The classification algorithm uses the filter method to rank features based on indices such as correlation coefficient and the selected features. The filter method is considered as a standalone feature selection method irrespective of the classification algorithm used. Wrapper method is an addition of filter method using data mining algorithms for variable ranking such as recursive feature elimination (RFE), Sequential backward selection (SBS) and forward feature selection. The embedded method is a combination of both filter and wrapper with variables generation is built into the model construction. Well establish example for wrapper and embedded method is Random Forest (RF), Elastic Net (EN) and decision trees (DT) (Chandrashekar et al., 2014; Saeys et al., 2007).

University of Malaya

Previous studies on application of ML methods on coronary diseases comprises ACS risk prediction using Random Forest (RF), Elastic Net (EN) and ridge regression for feature selection and risk classification by VanHouten et al. (2014). Genetic algorithm (GA) was used for feature selection by Amma (2012) and Nikam et al. (2017) to reduced number of attributes involved in the prediction of heart disease using artificial neural network (ANN) classifier. Mokeddem et al. (2013, 2014) applied GA for feature selection with Naïve Bayes (NB) classifier for CAD classification. The author then compared with other methods that are; support vector machine (SVM), Decision Tree (DT) and multiple layer perceptron (MLP). Salari et al. (2013) used k-nearest neighbor (k-NN) to remove redundant features to increase accuracy of classifying ACS subtypes using ML algorithms such as radial basis functions (RBF), k-NN, MLP, NB, iterative dichotomiser-3 (ID3), and Baggin-ID3, to identify the existence or absence of heart disease. Sonawane and Patil (2014) applied Learning vector quantization (LVQ). LVQ is made up of two layers, a competitive layer for feature selection and a linear layer for classification.

ML application for ACS mortality study comprises feature selection algorithm using filter and wrapper approach with SVM, ANN, RF and EN as the classifiers (Steele et al., 2018; Collazo et al., 2016). ML methods such as NB, DT, Logistic Regression (LR) and RF were used for feature selection and prediction of mortality 30 days after MI. ML methods outperformed conventional methods such TIMI and GRACE (Shouval et al., 2017). SVM, RF, LR and DT were used to predict 2 years' mortality after MI (Wallert et al., 2017). RF and SVM methods demonstrated high predictive performance in mortality studies compared with other classification algorithm even when presented with larger number of variables. RF is also robust to transformation of variables eliminating the need for variable transformation or normalization and is able to accommodate nonlinearities and relationship between predictive variables compared to other ML algorithms (Wiens & Shenoy, 2017; Ross et al., 2016; Schmid et al., 2016; Ishwaran et al., 2008).

University of Malaya

Discovery of relationship between variables is important besides variable selection. Kohonen Self-Organizing Map (SOM) is an unsupervised ANN which performs a topology maintaining prediction at the same time from instance vectors to a consistent 2D grid (Kohonen, 2001). It involves an iterative process based on the cluster analysis method that allows discovery of relationship and pattern in data set that leads to additional knowledge discovery via visualization of SOM maps. SOM method has been applied in maps that monitor the progress of trends and the extent of the degree of injury in dysphasia and disordered speech analysis (Tuckova, 2013). SOM was also applied to analyze the association of factors affecting lower limb pediatric fracture healing time (Malek et al., 2018).

None of the above studies has explicitly focused on survival prediction for ACS patients based on the Malaysian population and no literature was reported on application of SOM to understand relationship between factors that affects mortality.

TIMI and Framingham risk score (FRS) are the most commonly used risk scores in Malaysia for predicting ACS. However, these two methods have their limitations. On the one hand, evidence shows that TIMI is not suitable for prediction of coronary heart disease in adults above 75 years (Feder et al., 2015). In other words, there is a prevalence of poor prediction of mortality after ACS that would enhance the efficiency of allocating limited clinician resources.

On the contrary, FRS is suitable for adults although it inadequately predicts cardiac risk in young people and it could not predict future total cardiovascular events like risk for stroke, transient ischemic attack and heart failure (Lee et al., 2010).

Finally, the existing ML models for prediction of ACS were not based on Malaysian population. Since Malaysia is not an exceptional of these diseases, it is important to find

University of Malaya

out the relevant model and appropriate methodology to predict the ACS in Malaysia. Therefore, this research aims at implementing ML algorithm based on Malaysian population. This is aiming at predicting survival versus no-survival after ACS on Malaysian population.

1.2 Problem Statement

Very little research efforts have been conducted to implement any ML model that can be used to predict mortality after ACS based on the Malaysian population. Hence, it is important for health care practitioners in Malaysia to identify which patient requires intensive attention and care and efficiently allocate the limited clinician resources. Therefore, the research problems can be summarized thus:

- To be able to predict mortality after ACS would enhance your efficiency of allocating the limited clinician resources available.

1.3 Research questions

RQ1: What are the major predictors of ACS mortality among the Malaysian population?

RQ2: Feasibility of ML techniques to predict mortality after ACS?

RQ3: What is the performance of different ML techniques in predicting survival and non-survival after ACS?

1.4 Aims and Objectives

1. To investigate the major predictors of ACS mortality among the Malaysian population using ML.
2. To compare and implement models for predicting mortality after ACS using ML techniques based on Malaysian population.

University of Malaya

3. To visualize and discover relationship between various factors that affects mortality among ACS patients.

1.5 Scope of the study

This study was carried out on a subset of Malaysian population. It covers a model development on ACS mortality prediction. Ten different methods were used for feature selection; RF, RFE, Boruta, Cluster Dendrogram (CD), GA, EN, LR, LVQ, DT and SVM. Methods such as RFE, Boruta, CD, GA and LVQ were used only for feature selection and later combined with RF and SVM for classification. RF, SVM, LR, EN, DT were used for both feature selection and classification. RF and SVM models were later compared with each other to determine which among the two can highly predict mortality after ACS. SOM was also used in this study to visualize and determine the relationship between the variables selected based on the best model.

1.6 Contribution of the study

This can be explained in two aspects. First, it proposed and implemented an algorithm for predicting survival versus no-survival of patients after ACS using ML techniques. Since the algorithm was based on Malaysian population, it gave a clear insight of the predictors of ACS among Malaysian population.

Secondly, it identified the best and most efficient ML algorithm deployed in the prediction of survival versus non-survival patients after ACS which enables health care practitioners to easily identify a patient who needs immediate care and attention.

1.7 Thesis structure

The thesis organization is as follows:

Chapter 2: Gives detail understanding of ACS. Furthermore, it discusses the different types of ACS, conventional methods and ML techniques that were previously used in

University of Malaya

previous studies about Coronary artery disease with ML, feature selection methods, mortality prediction, risk scores and many others.

Chapter 3: Defines and discusses ML development processes. It explains different ML techniques and feature selection methods for solving mortality related problems.

Chapter 4: This chapter presents major predictors of mortality after ACS where feature selection and machine learning results were presented and discussed hence answering RQ1 and RQ 2.

Chapter 5: This chapter describes different ML and feature selection methods that were used in this study and presents the performance of different ML techniques in predicting survival and non-survival after ACS. The chapter ends by giving a brief over view of the study and the future work hence answering RQ3.

Chapter 6: this chapter presents the conclusive remarks about the overall study.

University of Malaya

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

This chapter provides detailed formal assessment on the relevant literature for gaining an insight into the related work done in the mortality prediction after ACS using ML techniques. The review is broadly classified into five sections; ACS, mortality prediction, risk scores, ML and feature selection.

2.2 Description Of ACS

ACS is simply a subset of coronary heart disease (CHD) ranging from STEMI to NSTEMI. ACS occurs when part of a muscular tissue of the heart is blocked from receiving blood. The segment of the heart muscle dies if there is no supply of oxygen-rich in blood that is required for its survival (Christenson et al., 2013).

Kumar and Cannon, (2009) explained ACS as a general term for series of illnesses or disorders that rapidly affects the coronary artery blood flow. Sometimes blood may be sufficient when flowing but may be inadequate in case there is a need of higher blood flow for-example during exercise and this is simply referred to as stable angina, which is not part of ACS (Thygesen et al., 2012). There are three main types of ACS, namely; unstable angina (UN), non-ST segment elevation myocardial infarction (NSTEMI), and ST segment elevation myocardial infarction (NSTEMI). Figure 1 shows ACS types and how it can be determined.

University of Malaya

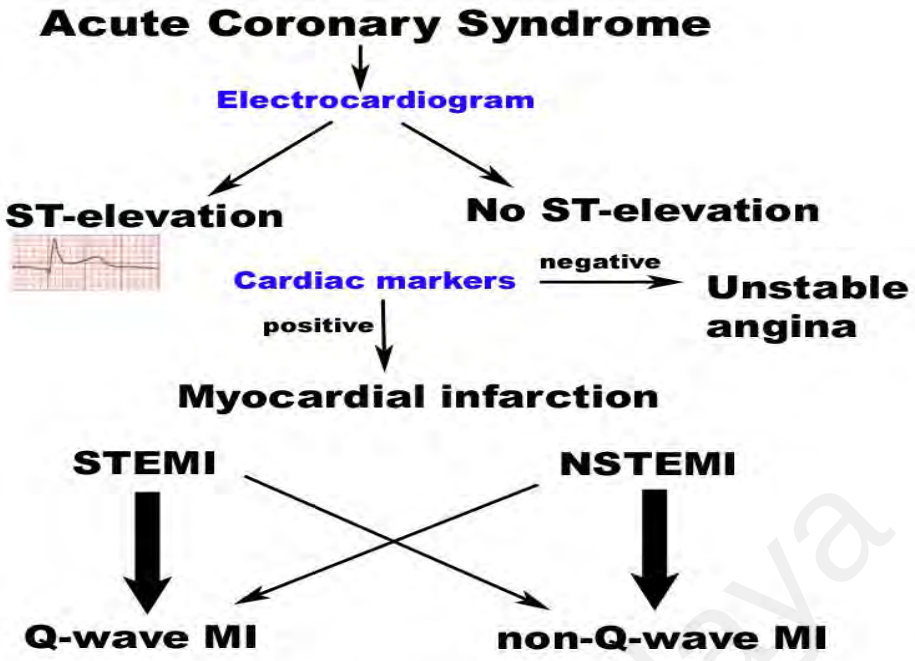

Figure 2.1: Illustration on ACS: Image Adopted from ACS scheme.

STEMI is a dangerous kind of heart attack where one of the heart's major arteries is blocked. NSTEMI is a type of a heart attack that typically less damage to your heart. Grech et al. (2003a, 2003b) mentioned that difference of STEMI and NSTEMI patients as "the absence of ST elevation on the presenting ECG". UA has no clear description, but it is known as a medical condition concerning both stable angina and MI. UA is any kind of persistent chest pain than the patient's normal signs of angina that take place when during resting, with little exercise or can't be controlled by medications. Angina pectoris typically occurs in the chest sub sternal part and this can move to other parts of the body such as left arm (Pollack et al., 2008).

UA is a condition in which your heart does not get enough blood flow and oxygen. Altman et al., (2008), identified angina pectoris as the main sign for patients with CVD. When the angina is less predictable or occurs during rest, the angina is called unstable angina pectoris (UAP). Many patients with UAP progress to myocardial infarction without intervention to open up the coronary artery. In UA, there is no elevation of

University of Malaya

biomarkers compared to non-STEMI, where there is an elevation of biomarkers (Thygesen et al., 2012)

MI is simply the medical term for heart attack. Meier et al., (2009) described MI as the terminology used when myocardial necrosis signs are present in medical settings with regular ischemia. This MI commonly known as heart attack is a myocardial cell death due to prolonged ischemia and is the main reason of death and ill health in the whole world (Mendis, 2010, 2011). Figure 2.2 shows two blood vessels that supplies blood to the heart and these are; the left and right coronary (labelled LCA and RCA). A myocardial infarction (2) has occurred with blockage of a branch of the left coronary artery (1).

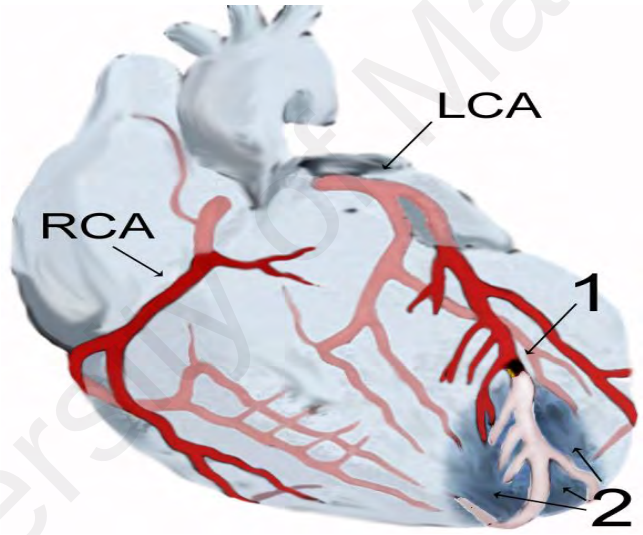

Figure 2.2: Major blood vessels for the blood stream to the heart (After Thygesen et al., 2012).

The MI classifications are highlighted in Table 2:1 below. Differences among MI types are based on the condition of the coronary arteries adopted from (Thygesen et al., 2012; Chapman et al., 2016; Collinson et al., 2015).

University of Malaya

Table 2.1: Describes different MI TYPE.

Classifications of Myocardial Infarction. Summaries the MI Type from (chapman, et al, 2016)	Description
Type 1: spontaneous myocardial infarction	This is related to atherosclerotic plaque rupture, ulceration, fissuring, erosion, or dissection with resulting intraluminal thrombus in one or more of the coronary arteries leading to decreased myocardial blood flow or distal platelet emboli with ensuing myocyte necrosis.
Type 2: MI secondary to an ischemic imbalance	Myocardial injury with necrosis where a condition other than coronary artery disease contributes to an imbalance between myocardial oxygen supply and demand.
Type 3: MI resulting in death when biomarker values are unavailable	Cardiac death with symptoms suggestive of myocardial ischemia and presumed new ischemic ECG changes or new left bundle branch block
Type 4a: MI related to percutaneous coronary intervention Type 4b: MI associated with stent thrombosis	Myocardial injury or infarction associated with mechanical revascularization procedures such as percutaneous coronary intervention or coronary artery bypass grafting (CABG) surgery or myocardial infarction associated with stent thrombosis detected by coronary angiography or autopsy in the setting of myocardial ischemia.
Type 5: MI related to coronary artery bypass grafting	Elevation of cardiac troponin values may be detected following these procedures, since various insults may occur that can lead to myocardial injury with necrosis.

University of Mosul

2.2.1 Common Predictors affecting mortality after ACS

The primary risk factors associated with the development of ACS are hyperlipidemia, diabetes mellitus, hypertension, and use of tobacco, male gender, older age, obesity, race and family history (Hajar, 2017; Ahmed et al., 2017). However, previous research also indicates that more women die due to NSTE-ACS than men do. Furthermore, number of mortalities in women has continued to rise regardless of the existence of timely primary

percutaneous coronary intervention (PCI) (Mansoor et al., 2017). Some of the common predictors affecting mortality in CVD patients includes but not limited to the following.

Stroke. When the movement of blood to the brain is constrained, ischemic stroke occurs. Kenney et al. (2012) characterized stroke as a disease of the cerebral arteries. Stroke may also be the outcome hemorrhage in the brain caused by artery blockage. Ischemic stroke in most cases is the effect of thrombosis, which may cause brain tissue damage hence a risk factor of mortality to ACS patients (Asadi et al., 2014).

Hypertension. The hypertension indicates that the heart has to pump harder to circulate the same amount of blood due to the increased resistance in the arteries. Overtime, the heart muscle becomes strained and enlarged, and the arteries become less elastic and damaged which puts a patient with ACS (Kenney et al., 2012).

Diabetes mellitus (DM) is a condition categorized by rising up of blood sugar due to lack of insulin production or resistance. Insulin is a hormone unrestricted by the pancreas used to regulate carbohydrate metabolism. DM is known to be the major risk of CVD (Zaccardi et al., 2015). High levels of sugar or glucose in the blood lead to damage of the arterial walls contributing to atherogenesis (Kenny et al., 2016).

Smoking is known to increase heart attack risks as it increases inciting response of the body hence a contribution to calcification of the artery of the wall. Assessing the history of smoke can also help to determine the levels at which a patient is at risk of heart disease (Lloyd-Jones et al., 2010).

Some signs of ACS risk are not rehabilitated, and these are commonly known as non-modifiable factors such as age and gender. Dagostino et al. (2008) stated majority of the people who die of heart diseases are above 65 years and male are at a high risk of CVD death. Hozawa et al. (2007) reported race as one of the risk factors of heart attack.

University of Malaya

Meanwhile Wilson et al. (1998) identified gender, age, smoke, TC, HDL, as the major risk factors for CHD. The Table below summarizes some of the common predictors that are reported in literature.

Table 2.2: Summary of the Common Predictors Mortality in ACS patients.

Authors (Ref #)	**Application**	**Instances**	**Variables included**	**Risk factors / variables selected**
Wang et al., 2018	Risk Factors Associated with Major Cardiovascular Events 1 Year After Acute Myocardial Infarction	4227 p atients	age, education, prior AMI, prior fibrillation, hypertension, angina, ejection fraction (EF), renal dysfunction, heart rate, SBP, white blood cell count (WBC), FBS	Age, EF, WBC, fibrillation, prior angina, and heart rate
Ahmed et al., 2017	Prevalence and Risk Factors for Acute Coronary Syndrome Among Sudanese Individuals with Diabetes: A Population-Based Study	496 respondents	HbA1c, cholesterol and triglycerides levels, age, gender, smoking, alcohol, DM duration, BMI, HDL, LDL, hypertension	hypertension, older age and increase in duration of DM.
Adhikari et al., 2018	Clinical profile of patients presenting with acute myocardial infarction	132 patients	age, gender, tobacco, smoking, hypertension: BP under medication, diabetes, FBS, dyslipidaemia, HDL, triglycerides, TC total cholesterol, alcohol, chest pain, shortness of breath, syncope, vomiting e.t.c	chest pain, shortness of breath, vomiting, Tobacco, smoking, hypertension and diabetes
Mirza et al., 2018	Risk factors for acute coronary syndrome in patients below the age of 40 years	100 patients	DM, Obesity Hypertension, Smoking, Family history of ACS, WBC count, age, gender, BMI, Lymphocyte count	Obesity, smoking ,hypertension, diabetes mellitus and family history.
Alhassan et al., 2017	Risk Factors Associated with Acute Coronary Syndrome in Northern Saudi Arabia	156 patients	Age, nationality, gender, Hypertension, Ischemic Heart Disease (IHD), Smoking, Diabetes Miletus (DM), and Dyslipidaemia	Hypertension, Ischemic Heart Disease (IHD), Smoking, Diabetes Miletus (DM), and Dyslipidaemia
Bęćkowski et al., 2018	Risk factors predisposing to acute coronary syndromes in young women ≤45 years of age.	1941 women patients	hypertension, obesity, hypercholesterolemia, di abetes mellitus, and cigarette smoking, family history of CAD, kidney disease, lung disease, ischemic stroke, peripheral arterial disease, age, body mass index (BMI) , FBS, smoking, Creatinine, SBP, DBP, TC	Age, diabetes, smoking, obesity, hypertension, hypercholesterolemia , history of stroke

University of Malaya

Table 2.2, Continued.

Authors (Ref #)	Application	Instances	Variables included	Risk factors / variables selected
Hodzic et al., 2018	Seasonal Incidence of Acute Coronary Syndrome and Its Features	250 patients	Age, gender, hypertension, hyperlipidaemia, diabetes mellitus, positive family history, tobacco smoking, Employment status, Troponin I, Killip III, IV, and fatal outcomes of ACS.	hypertension, hyperlipidemia, diabetes mellitus, positive family history, smoking, Troponin I
Vedanthan et al., 2014	Global Perspective on Acute Coronary Syndrome.		ASA, ACE, B-BLOCKER, statins, Cholesterol and Recurrent Events, obesity, and diabetes mellitus. Body mass index, AGE, income bracket.	IHD, age, poor people
Ricci et al., 2017	Acute Coronary Syndrome: The Risk to Young Women	14931 patients	aged ≤45 years, PCI, hypercholesterolemia, hypertension, and diabetes mellitus, smoking status, family history of CAD, and BMI; clinical history of ischemic heart disease, stroke, SBP and heart rate, chronic kidney disease, aspirin, clopidogrel, heparins, b blockers, and ACE	aged ≤45 years, smokers, men, diabetes mellitus, hypercholesterolemia, and hypertension
Kayani et al.,2018	Improving Outcomes After Myocardial Infarction in the US Population	13079 respondents	blood cholesterol, blood pressure, blood glucose, diet, physical activity, smoking, and body mass index,	FBS, TC, smoking, BP
Haneef et al., 2010	Risk Factors Among Patients with Acute Coronary Syndrome in Rural Kerala	130 patients	Obesity, Dyslipidemia, Alcohol, Smoking, Hypertension	Obesity, Dyslipidemia, Alcohol, Smoking, Hypertension

University of Malaya

2.3 Mortality prediction

The risk of mortality (ROM) estimates the likelihood of death of a patient and provides a medical classification of patient mortality. Several studies assessing the general health of a person has based on the survey rating provided by individuals as a response questionnaire. The score is useful in finding a rough estimate of the individuals who are not in a healthy condition and are seeking for medical assistance. DeSalvo et al. (2006), found that there is a statistically significant relationship between general self-rated health and high risk of mortality. Individual with poor general self-rated health had higher mortality risk as compared to the person with self-rated health as excellent.

Health planners for health as well as makers of policies are trying to find out a feasible method to identify the most vulnerable person with highest health requirements. ML algorithms can be used to improve the health given to a patient through the identification of groups who are more susceptible to mortality risk. The collection of such data may help in offering a beneficial tool in health and care planning sector and allocation of resources to those who require immediate attention

2.3.1 Mortality prediction techniques

Prediction of future health status can be significant in the medical domain as it can contribute to early detection of a disease, effective treatment, prevention and identification of high-risk patients (Hoogendoorn et al., 2016). ML and convention methods commonly known as risk scores are used in predicting mortality in various cases and mostly ACS. The health-related information of an individual stored in Electronic Medical Records can be used to generate accurate predictions for the occurrence of health issues. Predictive data mining has received increasing interest as an instrument for researchers across various fields. ML offers new methodological and technical solutions for the analysis of medical data and the construction of prediction models. Examples of these techniques include RF and SVM. These techniques are based on algorithms which

University of Malaya

operate by building a model from example inputs to make data-driven predictions or decisions, rather than following strictly static program instructions as used in traditional classification modelling. Further details are presented in the next sections of this chapter.

In Malaysia, conventional methods are currently used in prediction of patients' mortality. The conventional methods commonly known as risk scores, which are currently used in Malaysia, is discussed more in detail in the next section of this study.

2.4 Risk scores

Risk scores have been used in identifying patients with ACS. These risk scores were developed based on expert opinion to include variables that were thought to be more significant according to the expert for example cardiologist. There many risk scores used worldwide, the most commonly used in Malaysian population are; TIMI, PURSUIT, GRACE, HEART, FRISC, SCORE and Reynolds as explained in more details below.

2.4.1 TIMI risk score

TIMI (thrombolysis in Myocardial Infarction) is the risk score used for NSTEMI and UA. It was designed to clinically predict mortality or major complications over 14 days; multivariable LR with SBS was used to build a mathematical representation, the risk score was designed to contain only the seven clinical variables with significant effects on outcome, each of which contributing a maximum of one point to the overall seven-point score. The variables that were included in the TIMI score were age> 65years, risk factors for CAD (at least three), significant priory coronary stenosis, ST deviation on ECG, severe angina symptoms, the use of aspirin in the past 7 days, and elevated serum cardiac markers. Some studies suggested that TIMI should be modified from the existing one to include newer existing biomarkers and to permit a broad definition of ischemic changes on ECG (Hess et al., 2010; Body et al., 2009).

University of Malaya

2.4.2 PURSUIT Score

The PURSUIT score (2000) was developed in a multinational randomized clinical trial (Platelet glycoprotein IIb/IIIa in unstable angina: Receptor Suppression using Integrilin (eptifibatide) therapy. The score was derived via multiple LR with backwards stepwise selection, but unlike the TIMI score, allowed for graded responses for the different clinical variables. The variables which were included in this score were; age, sex, heart failure symptoms, heart rate, SBP, the presence for rales on examination, and ECG on ST-depression. The PURSUIT score is well- known in guiding triage or treatment decisions in the emergency department (Boersma et al., 2000).

2.4.3 GRACE Score

The Global Registry of Acute Coronary Events (GRACE) score was published in 2003 (Granger et al., 2003). The GRACE score is derived from patients in a registry, where no experimental treatment was explored. However, patients in this registry were required to have received a final diagnosis of ACS, and patients were included in the registry only if they had ECG alterations signifying ACS, sequential rise in cardiac enzymes, or documented CAD. Included variables were the Killip class of heart failure, SBP, heart rate, age, creatinine, and existence or absence of cardiac arrest at the time of admission, ST-segment abnormality, and high cardiac enzyme levels. While the original GRACE model was developed for predicting in-hospital mortality to predict mortality and myocardial infarction over longer durations following (Gray et al., 2011; Fox et al., 2006).

2.4.4 HEART

The HEART risk score was developed mainly for patients who present with chest pain at the emergency department. This was developed to predict ACS by European society of cardiology in order to improve health and reduce risks in patients with cardiovascular problems (Ma et al., 2016; Fesmire et al., 2012). The acronym HEART was developed with the first letter of each of its predictors (Six et al., 2008). The HEART score is

University of Malaya

composed of five variables and these include; History, ECG, Age, Risk factors and Troponin. The structure of HEART score was mainly based on decision making clinical factor according to expertise opinion.

2.4.5 FRISC

In his study, Lagerqvist (2005) based on FRICS (Fast Revascularisation in Instability in Coronary disease) score to select patients for an early invasive treatment in unstable coronary artery disease. This risk score was composed of age greater than 70years, patients with diabetes, male, with the history of MI and troponin on the admission.

2.4.6 The Reynolds Risk Score

The Reynolds Risk Score was developed to improve prediction of CVD risk in women and a model for men was later developed (Ridker et al., 2008). The score uses similar features as FRS in addition to family history with age of 60. It was developed to work on non-diabetic patients with the age between 45 and 80 to predict any future heart problems.

2.4.7 SCORE

The SCORE (Systematic Coronary Risk Evaluation) development was mainly based on 12 European cohort studies. It focused on these risks; gender, age, SBP, smoking and cholesterols (Conroy et al., 2003).

Table 2.3 summarizes the above-mentioned conventional risk scores used for heart risk prediction.

University of Malaya

Table 2.3: Summary of Common Conventional Risk Scores for HD prediction.

Author (Ref#)	Risk score	Variables used
Conroy et al., 2003	Score	Gender, age, SBP, smoke, TC and HDL
Ridker et al., 2008	Reynolds	Age, SBP, cholesterol levels, family history and smoke
Lagerqvist (2005)	FRISC	History of MI, troponin, age, patient with diabetes, gender
six et al., 2008	HEART	History, ECG, Age, Risk factors and Troponin
Boersma et al., 2000	PURSUIT	age, sex, heartfailure,heartrate,SBP,presenceof rales,and ECG
Antman et al., 2000	TIMI	Age, CAD risk factors, cardiac marker, severe angina, and ASA
Grangeret al.,2003	GRACE	Age, heart rate. SBP, creatinine, cardiac arrest, killip class, ECG.
Rodondi et al., 2012	FRSI	Age, gender, smoker, SBP, TC, HDL, blood pressure being treated with medicine

2.5 Overview on Machine Learning

According to Paluszek and Thomas (2016), ML allows computers to decide basing on experiences, reaction and actions. ML has been successfully used in many fields of medicine, bioinformatics, biology, business and many others. ML offers advantages over statistical methods used for predictions i.e. easing the process of knowledge acquisition from a system or reducing the time consumption (Kesavaraj et al., 2013).

Kononenko (2007), states that the quality of ML classification algorithms depends on the selection of the classifier and concluded that combinations of classifiers are more reliable in a diagnostic system problem instead of single classifier. In addition, classification performance is highly impacted by data pre-processing and tuning of algorithms (Kesavaraj et al., 2013).

ML models for predicting mortality after ACS are developed to predict the benefits of cardiac surgery in the event of ACS. Various studies have indicated that ML, though relatively a new approach is by far a better approach for predicting mortality after cardiac surgery than conventional risk scores (Allyn et al., 2017).

University of Malaya

Tapas et al. (2017) proposed ensemble classifiers based on RF for prediction of cardiac arrest. Their system showed high accuracy compared to other ML algorithms.

Having proper data for ML algorithms is very important for training and testing ML algorithms. A number of ML techniques have been deployed in developing and validating prediction models for ACS that include among others LR and RF (Mansoor et al., 2017).

None of the studies reviewed explicitly focused on the predicting survival and non-survival after ACS based on the Malaysian population. The following sub-sections gives an overview of classifiers used in this study which include DT, RF**,** SVM, LR, EN and LVQ that are supervised learning and SOM which is unsupervised.

2.5.1 ML algorithms

ML algorithms are categorized as supervised and unsupervised learning. Both types of ML algorithms have been deployed in this study.

2.5.1.1 *Supervised learning*

Supervised learning is defined as when data with corresponding correct outputs is provided during training for predicting the future unknown outputs of a given instance. The common algorithms are; LR, SVM, K-NN, ANN, NB and DT (Chandralekha & Shenbagavadivu, 2018).

Supervised ML models have been used to build predictive models for medical diagnosis (Maroco et al., 2011). A classifier is a function that given an instance assigns it to one of the predefined classes. In this study, classification algorithms such as DT, RF, SVM, LR and EN are used and explained briefly as follows:

(a) Decision tree (DT)

DT is a graphic representation of obtained knowledge in the form of a tree or flow chart, where each non-leaf node denotes a test on an attribute, and each branch indicates

University of Malaya

an output of the test (Hachesu et al., 2013; Sundaram et al., 2012; Jenhani et al., 2008). A classifier starts with testing the values of features one by one while considering only the important ones. It later divides the data and tests the results into separate classifications basing on the selected features (Jiang & Shekhar, 2017; Du et al., 2011; Dong et al., 2009; Li et al., 2006).

The most popular implementations of DT algorithm are C4.5 where a feature is selected by the algorithm at the best split of the samples according to the normalization (Quinlan 1993, 1986). C4.5 constructs an ensemble tree through stage-wise development of many decision trees or corresponding rule-sets, emphasizing misclassified cases in previously developed trees. Let (m) denote the classified cases. For the growth of one tree based on these cases(T_m), the algorithm first decides the predictor and predictor cut-off value that provides the optimal single-split. This decision is based on entropy(I_E), which is defined

$$I_E(f) = -\sum_{i=1}^{m} f_i \log_2 f_i \qquad (1)$$

Here (Eq. 2) f_i denotes the probability of each case being chosen for the split. The greatest reduction in entropy before and after this split is the greatest increase in information gain, since information gain = entropy (before split) – entropy (weighted sum after split).

Other implementations of DT algorithm include Information gain (IG) and Gini Index (GI). IG is the expected reduction in entropy caused by partitioning the examples according to this attribute. The Gini Index is a measure of node purity due to the very small values of the index when the node observations are predominantly from a single class (James et al., 2013; Breiman et al., 1984). DT was chosen for this study using information gain criteria to select variables as well as for prediction.

University of Malaya

Once all training observations are sorted and assigned to a terminal node or region, the large initial tree T has to be grown and is ready to be pruned to identify the best subtree out of T (James et al., 2013). Pruning involves deleting a branch and all its descendants from a tree, leaving only the branch's root node. Pruning is how the tree methodology deals with the concept of bias-variance trade-off. Potentially, a tree could be grown to the point that every node was pure with a misclassification error of zero.

The predictive accuracy of DT increases as more features are added, the number of features is limited for optimum performance i.e. adding features to DT beyond a particular number can significantly lower the performance of the entire prediction model (Özçift, 2011). Randomized ensembles aggregate a combination of tree predictors based on random, independently sampled vectors through similar supply (Breiman, 2001).

(b) Random forest (RF)

RF (Brieman, 2001) defined it as an ensemble classifier with the combined tree predictors where an additional randomness is added to each tree (Liaw & Wiener, 2002). The difference between RF compared to other trees is that RF chooses predictors at random from the whole set of predictors (Genuer et al., 2010). This is symbolized by *mtry* and the best split is determined using Gini index node of impurity that is calculated from the subset of predictors. The value of Gini index is 0 and 1. 0 indicates that all predictors at the node are of the same class history (Khalilia et al., 2011). For the error rate to be reduced, at each node, the value for mtry should be *mtry* = p1/2 classification or *mtry* = p/3 for regression. No pruning step is needed in RF hence the trees generated are the maximal (Datla, 2015).

Test set error estimate is obtained from growing a tree from a bootstrap data (Verikas et al., 2011) which then be used to estimate the variable importance and these two are the useful byproducts of RF. One of the byproducts of RF is the variable importance. The

University of Malaya

four measures of the variable importance are raw importance score for class 0, raw importance score for class 1, a decrease in accuracy and the Gini index. Increase in the error rate is expected from the permutation of variable importance thus leading to high permutation value (Genuer et al., 2010). The calculations are carried out as each of trees in the forest is being grown. Therefore, RF was used in this study for feature selection and model development.

(c) Support vector machine (SVM)

SVM can be used to model and predict responses in linear and non-linear data dealing with high-dimensional data such as gene expression (scholkopf et al., 2018; Ben-Hur et al., 2008; Karatzoglou et al., 2006). SVM technique for classification goal is to use vector of explanatory variables to estimate the optimal decision boundary that best separates the class labels (Cortes & Vapnik, 1995; Clarke et al., 2009). SVM uses optimization parameters in case of grid search which is known as large margin classifier. In the simple binary cases, the two classes separate linearly and the boundary between the two classes is called the hyperplane. Kernelization of the SVM classifier enables the actual learning to take place in the feature space. The kernel function returns the inner product between the images of two data points in feature space (Karatzoglou et al., 2006). This referred to in literature as the "kernel trick" (Scholkopf, 2018).

SVMs kernel methods are constructed to use a kernel for a particular problem that could be applied directly to the data without the need for a feature extraction process. This is particularly important in problems where a lot of structure of the data is lost by the feature extraction process (Suykens, 2001; Bao et al., 2007).

Some widely used kernels in SVM are: polynomial, Radial Basic Function (RBF) and Linear (Rai, 2011).

University of Malaya

Linear kernel $\kappa(\chi_\iota, \chi_j) = 1 + \chi_\iota^\tau \chi_j$ is a simple kernel function based on the penalty parameter C, since parameter C controls the trade-off between frequencies of error c and complexity of decision rule but it is not suitable for large datasets (Cortes & Vapnik, 1995).

Polynomial kernel $\kappa(\chi_\iota, \chi_j) = (1 + \chi_\iota^\tau \chi_j)^p$ also known as global kernel, is non-stochastic kernel estimate with two parameters i.e. C and polynomial degree p. Each data from the set x_i has an influence on the kernel point of the test value χ_j, irrespective of it's the actual distance from χ_ι . It gives good classification accuracy with minimum number of support vectors and low classification error.

Radial basis function $\kappa(\chi_\iota, \chi_j) = exp(-\gamma \|\chi_\iota, \chi_j\|^2)$ also known as local kernel, is equivalent to transforming the data into an infinite dimensional Hilbert space .Thus, it can easily solve the non-linear classification problem. RBF gives similar result as polynomial with minimum training error but for some cases, the number of support vector and classification error increases (Álvarez et al., 2018).

SVM has to be fine-tuned depending on the type of data it will be used for. Some decisions that have to be made are: how to pre-process the data, what kernel to use (linearly separable, linearly non-separable or non-linear).

This study used linear, RBF and polynomial kernels for both feature selection and model development.

(d) Genetic Algorithm (GA)

GA is an adaptive heuristic search algorithm used in finding optimal parameters for real-world problems and are widely used in random search problems within a defined search space when the algorithm is well tailored to specific problem with appropriate fitness function and search operators (Holland, 1992). Figure 2.3 illustrates GA algorithm

University of Malaya

process that describes basic process of fitness evaluation, natural selection and cross over and mutation. GA was utilized in this study for parameter optimization. The choice of parameter settings for GA is experimentally determined as follows (Tay et al., 2014 & 2013).

a) Population size: maximum generation, natural selection and stochastic universal sampling.

b) Crossover type: discrete recombination, crossover probability, mutation rate: 1/P, where P is the number of parameters.

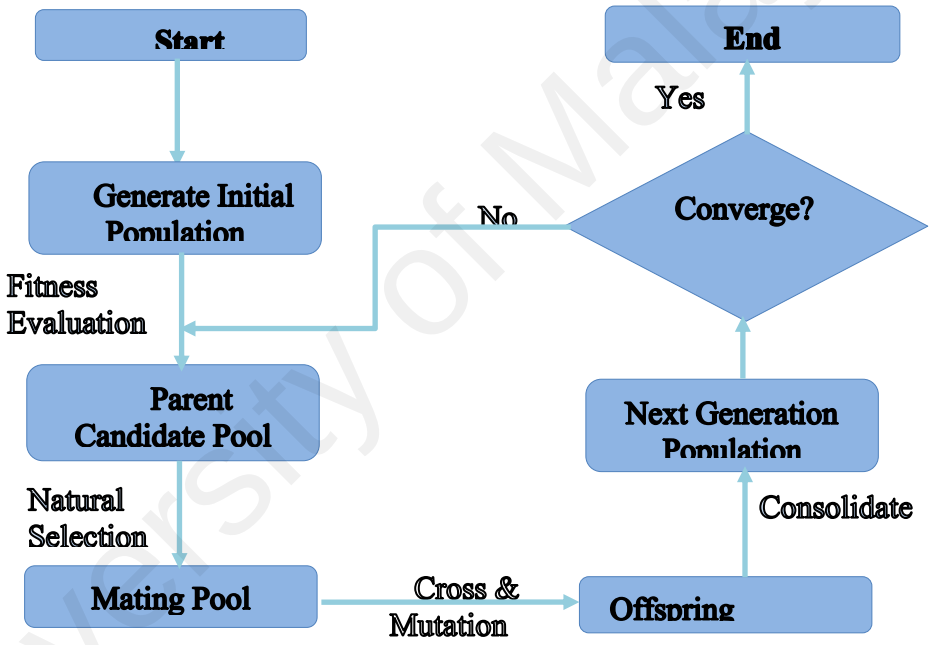

Figure 2.3: Genetic algorithm process.

(e) Elastic Net (EN)

EN (Zou & Hastie, 2005) is a popular regularization and variable selection method that merges the useful properties of ridge regression and lasso. It can handle multi-collinearity and it possesses variable selection property. EN is designed to combine these two measures as the EN penalty P. The entire family of P_α creates a useful compromise between ridge and lasso regression (Friedman et al., 2010).

University of Malaya

Formula for calculating EN: (2)

$$\min_{(\beta_0,\beta)\in R^{p+1}} \left[\frac{1}{2N}\sum_{i=1}^{N}(y_i - x_i^T\beta)^2 + \lambda P_\alpha(\beta)\right]$$

Where β_0 and β are the regression coefficients and $P_\alpha(\beta)$ is that all variables are treated as being independent of each other. (3)

$$P_\alpha(\beta) = \sum_{j=1}^{p}\left[\frac{1}{2}(1-\alpha)\beta_j^2 + |\beta_j|\right] \; with \; j = 1, \ldots\ldots, p$$

P_α is the EN penalty and α can be used to get a compromise between the ridge regression penalty (α = 0) and the lasso regression penalty (α = 1). If you choose α = 1-£ for some small £ > 0, then the EN results in lasso regression but removes degeneracies caused by extreme correlations (Friedman et al., 2010). EN optimizes the coefficients until the change of the coefficients is smaller than a predetermined toleration value. Choosing a small toleration value causes the algorithm to take longer to find the best values for the coefficients. For fitting EN model, cv. glmnet () function is recommended. EN has been used in the current study for both feature selection and model development. The function cv. Glmnet () has also been used in the current study for model fitting as recommended by Friedman et al. (2010).

(f) Learning Vector Quantization (LVQ)

LVQ is related to SOM but with the difference that the LVQ is a supervised learning algorithm and the SOM is an unsupervised algorithm (Nova et al., 2015), The LVQ is a classification model where the classification of given vector is equivalent to find the class label of the nearest prototype of vector. The prototypes are the neurons learned with the LVQ in the learning phase. At the starting point neurons are initialized in a random way from the training set and such type of classification is equivalent which base on the prototypes constructed by the LVQ learning (Grbovic & Vucetic 2009; Pedreira, et al., 2006). LVQ is a powerful classifier for high dimensional input data. A major advantage

University of Malaya

of LVQ is its simplicity: the system does not require expertise or explicit methods for normalization and feature selection. Moreover, it is possible to revise the feature selection with latest data samples. This is particularly important for applications where the relevance of certain features might change during operation of the system (Thakare & Patil, 2014; Saulnier et al., 2011). LVQ has been used in the current study for feature selection. LVQ is simple and does not require explicit method for feature selection as already stated above hence making it suitable to use in the current study.

(g) Logistic regression (LR)

LR is a statistical classification model that can be applied in the situations where the outcome is categorical. It has become a standard method of analysis in the situation where outcome variable is discrete taking two or more possible values (Park, 2013 and Al-Ghamdi, 2002). The outcome variable is dichotomous as it can take only two values such as yes/no, 0/1; such LR models are called as Binary LR Model and they are known to be multinomial if the outcome takes more than two values (Agresti, 2002, 2014). Binary LR is a prognostic model that is fitted where there is a dichotomous or binary dependent variable like in this instance where the researcher is interested in whether the patient survived after ACS or not. LR is the most popular technique that is used for modeling categorical dependent variables and that it does not require rigorous assumptions to be met (Kleinbaum et al., 2014, 2008; Al-Ghamdi, 2002).

Assuming a Bernoulli distribution of the dependent outcome (y) that is conditional on a set of input predictors $(x_1, \dots, x_k)$ we can write $y \mid x_1, \dots, x_k \sim \text{Bernoulli}(p)$. LR (Cox, 1958) then estimates the binary response probability (Eq. 1) through the function

$$\log\left[\frac{pr(y=1|x)}{(1-pr(y=1|x)}\right] = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \qquad (4)$$

University of Malaya

Where (β_0) is the intercept and ($\beta_1, \ldots, \beta_k$) are the estimated coefficients. We used LR to generate individual predictive probabilities between 0 and 1 using a cut-off at 0.5 for binary classification. LR lacks tuning parameters, which sets it apart from the other models.

LR has been used in the current study for both feature selection and prediction where the odds ratio was the probability that a patient may survive divided by the probability that the patient may not survive after ACS.

2.5.1.2 *Unsupervised learning*

Unsupervised learning is the family of ML algorithms and is mainly used in pattern detection and descriptive modeling. However, there are no output categories or labels based on which the algorithm can try to model relationships (Kohonen, 1995). These algorithms try to use techniques on the input data to mine for rules, detect patterns, summarize and group the data points, which help in deriving meaningful insights and describe the data better to the user as the data is un-labeled. SOM is as one of the common unsupervised learning algorithm used is SOM.

(a) Self-Organizing Map (SOM)

Kohonen's SOM is unsupervised mathematical model of topological mapping. SOMs learn on their own through unsupervised competitive learning, where it attempts and maps the weight to fit in the dataset. Topology relationship among inputs is conserved once plotted to SOM that is suitable for representing complex data. SOMs provide a way of representing multidimensional data in a much lower dimensional space into one or two dimensions (kohonen, 2001). SOM consists of two main Kohonen layer. Input layer of neurons in SOM are connected to the Kohonen layer. Input layer is presented and linked to all neurons which their connection is established in weight which vary for every iteration adaptively. Small value of weights is designated randomly to the input vector

University of Malaya

which later the space among the input and the summed weights are calculated in each of the neurons (Chaudhary et al., 2014).

This algorithm with the additional property preserves the topological mapping from input space to output space making it a great tool for visualization of high dimensional data in a lower dimension. The quality of learning of SOM is determined by the initial conditions: initial weight of the map, the neighborhood function, the learning rate, sequence of training vector and number of iterations (Pal & Pal, 1993). SOM was used in the current study to visualize and identify the relationship between the best predictors chosen by the best model.

2.6 Feature Selection

The probability of having features with inappropriate, redundant, and noisy characteristics increases when data dimensionality increases (Chang et al., 2014). Reducing number of attributes to a convenient amount but keeping the usefulness of the study is mostly used. Selection of features involves reducing on the existing variables to a minimal set which can produce better results, reduces training time and increases the accurateness of results making the overall results more understandable and appropriate (Kumbhar & Mali, 2016; Guyon et al., 2003).

Feature reduction process generates variables that improve model performance. The feature selection process can be identified in four different stages and these are; generation, evaluation, stopping criterion, and validation (Liu and Yu, 2005). Subset generation generates a sub set of features based on specific searching strategy, which can be used in evaluation metrics. When the criteria occur, the variable search stops and features selected are validated in determining its importance in predicting a certain criterion. Figure 2.4 illustrates the feature selection process as explained by Liu and Yu (2005).

University of Malaya

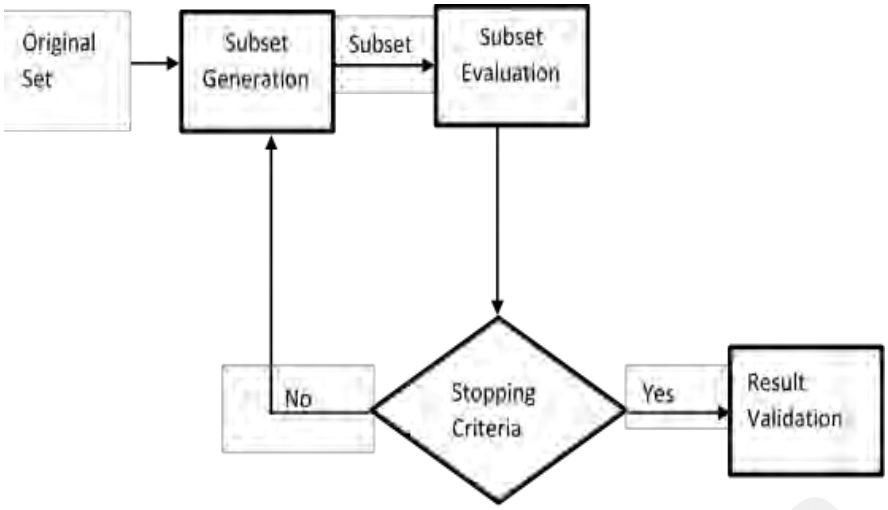

Figure 2.4: Feature Selection Procedure.

Feature selection is categorised into filter, wrapper and embedded methods (Jovic et al., 2015) and these have been used in the current study as explained in detail below;

2.6.1 Filter

Filter methods ranks variables one by one basing on some univariate search where variables with high rank are selected. Common univariate filter methods examples include NB, Information Gain and Euclidean Distance (George et al., 2011). Information gain is a filter feature selection technique, which ranks variables based on the scored significance of individually specific variable. It is univariate method since the relationship among variables is considered. The variables are selected by ranking them. Each input feature is calculated separately to get an insight on how it can be related to the targeted class and by doing this, the scored results is used in sorting the variables into a descending order. A predictive model is constructed using all features with high score without including low score variables (Blachnik et al., 2009).

In cluster dendrogram (CD) feature selection technique, the correlation between features is calculated in terms of Euclidean distance (Agarwal et al., 2010). This method produces a tree like diagram to illustrate the arrangement of clusters called dendrogram. The tree structure in dendrogram is not a single set of clusters but is a multilevel hierarchy,

University of Malaya

where clusters at one level are joined as clusters at the next level. At each stage, the algorithm joins the two clusters that are closer together and uses the distance between clusters. In a dendrogram, the height of the lines indicates the distance between the objects that are connected. Unlike partitioning method, this method gradually merges objects or divides a cluster (Galili, 2015; Prokashgoswami & Mahanta 2013; Lu & Liang, 2008 and Zhang et al., 2017).

Hall et al. (2003) benchmarked filter method with a wrapper method. They concluded that filter methods' performance depends on the dataset used and these methods were fast and increased the efficiency of the algorithm classified.

Filter method does not combine the final learning algorithm in its stages compared to wrapper and the selected features can at the same time be used in other algorithms for more investigations (Ladha & Deepa 2011). On the other hand, Saeys et al. (2007), mentioned the drawbacks of filter method that it poorly interacts with classifiers algorithms when used in the long run, and they added in their study that since the utmost filter methods are univariate in nature, these methods may not put much concern on values of other variables. Filter methods drawbacks are redundancy of the selected features and it ignores the association among variables.

Filter methods using Euclidean Distance was implemented in this study by CD algorithm for feature selection (Galili, 2015; George et al., 2011; Blachnik et al., 2009).

2.6.2 Wrapper

Wrapper Method: this involves carrying out a query where the certain classifier locates a set of variables where prediction models can perform optimally. Feature selection takes into account the contribution to the performance of a given type of classifier. Using a set of defined rules, the wrapper methods class performs a step by step selection of variables

University of Malaya

involving forward/backward criterion. Although the wrapper methods are slow, they are suitable for final model building compared to other methods. Common wrapper methods include LASSO (Tibshirani, 1996), EN (Zou & Hastie, 2005), RF, Sequential Forward Selection and Sequential Backward Selection.

Forward Selection is an iterating approach where a model begins with no features or empty set and then in each iteration, features are added to the model which leads to the enhancement in the performance of the model. This is carried on until no further enhancement in performance of the model can be achieved by the addition of new features. At each iteration the predictor that gives the highest improvement to the model is added (Gareth 2013; Kabir et al., 2010).

Sequential Backward Selection (SBS) is an alternative to sequential forward selection. This SBS starts with full set of variables and removes the least significant variable at each iteration to improve the model performance. The process repeats until no further improvement is obtained for the removal of features (Gareth 2013).

Doak (1992), compared both backward and forward selection methods where backward feature selection method outperformed forward feature selection method with the best. And hence backward feature selection method has been used in this study to select the best predictors for ACS. Backward searching strategies was used in selecting most significant variables for predicting mortality in patients with AC-STEMI (Stebbins et al., 2010).

The Boruta algorithm is a wrapper method that was designed to identify all variables that are associated within a classification framework (Miron et al., 2010). The Boruta approach was used by Guo et al. (2014) and Saulnier et al. (2011) to select significant variables and to analyze microbiome data.

University of Malaya

This method compares the usefulness of the real predictor variables with those of random so-called shadow variables using statistical testing and several runs of RF. The values of those shadow variables are generated by permuting the original values across observations and therefore destroying the relationship with the outcome and the variable importance values are collected. For each real variable, a statistical test is performed comparing its importance with the maximum value of all the shadow variables. Variables with significantly larger or smaller importance values are declared as important or unimportant, respectively. All the unwanted and shadow variables are discarded, and the procedures are again repeated until all variables are successfully classified, (Kursa, 2014).

Recursive Feature elimination (RFE) is a wrapper method that aims at finding a minimal and best performing set of variables, which leads to a good prediction model (Dı´az et al., 2006). It repeatedly creates model and keeps a side the best or the worst performing features at each iteration. It then constructs the next model with remaining features until all the features are exhausted. It then ranks the features based on the order of their elimination (Dietrich et al., 2016; Gregorutti et al., 2013; Habermann et al., 2009; Fusaro et al., 2009). RFE has been used in this study to select important predictors for predicting mortality after ACS.

Wrapper methods such as RFE, BORUTA, GA and sequential backward selection (SBS) have been used in the current study for feature selection.

2.6.3 Embedded Method

Embedded feature selection method is set to a particular algorithm where best feature searching criteria is constructed in classifiers. The method automatically selects the significant variables leaving the unimportant ones out. The approach is among the procedures that is used to train classifiers for example DT (Witten et al., 2011).

University of Malaya

Embedded methods advantages are; including the interactions to prediction models and low intensive computations when comparing with wrapper method. DT select relevant features using top-down, hierarchical partitioning schemes, where the output in a model that uses only a subset of features that appear in the nodes of the tree such as CART (Breiman et al., 1984) and C4.5 (Quinlan, 1993).

Belle et al. (2009) embeds features in predicting survivors of breast cancer where they changed functional losses in order to best differentiate variables. This was because less important variables have smaller differences as compared to most important ones.

Embedded methods such as the EN and LVQ have been used in this study for both feature selection and prediction.

2.7 Performance Measure

ML model performance measurements are required to determine the efficiency of classification algorithms. Common performance measure used for ML classification are and are adopted in this study are; accuracy, sensitivity, specificity and Area under Receiver Operating Characteristic Curve (AUC) and these can be reported in metric format table commonly known as confusion matrix (Fawcett, 2006).

2.7.1 Confusion Matrix

In a standard binary classification problem, the classifiers label all items as either positive or negative where confusion matrix summarizes the outcome of the algorithms into a matrix format (Chawla, 2005).

a) Accuracy

Accuracy is the easiest performance measure. This is where correct labels are divided against all classifications through calculating the complete efficiency of the algorithm.

University of Malaya

For binary classifications, accuracy can be described as the proportion of true results between the total populations. It indicates the nearness of anticipated values to the true or theoretic values. Formula for accuracy is denoted as follows;

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

Where TP = True Positive, TN = True Negative, FP= False Positive and FN = False Negative. TP is a value representing the count of patients at high risk and are classified correctly while FN is a value representing the count of patients at high risk and are predicted as patients at low risk.

FP is a value representing the count of patients at low risk and are classified as patients at high risk while FN is a value representing the count of patients at low risk and are classified correctly (Sakr et al., 2017 & 2018).

b) Sensitivity and Specificity

When the type of misclassification is vital in the classification problem, sensitivity and specificity is a better performance measure compared to accuracy (Powers, 2011). Sensitivity also known as recall or True Positive Rate (TPR) is the ratio of true positive prediction over the number of positive instances in the whole dataset or a classification performance measure defined as the proportion of correctly classified positives (Sakr et al., 2017 & 2018). The formula for sensitivity is as given below:

$$Sensitivity = \frac{TP}{TP + FN} \quad (6)$$

University of Malaya

The specificity or True Negative Rate (TNR) is the ratio of true negative predictions over the number of negative instances in the entire data set as explained in the formula below.

$$Specificity = \frac{TN}{TN + FP} \quad (7)$$

These values can be further analyzed using a Receiver Operating Characteristic Curve (ROC) where the sensitivity is plotted against 1- specificity (Fawcett, 2006).

c) Area under Receiver Operating Characteristic Curve (AUROC)

Huang and Ling (2005) reported that Accuracy is not a suitable performance measurement for cases where classes are heavily unbalanced. They concluded that AUROC evaluation, as a performance measure is a better option instead of accuracy when comparing learning algorithms applied to real-world data sets. AUROC is a statistically consistent and more selective performance measure than accuracy. They also showed that by using the AUROC evaluation for measurement, a real-world concern could be easier optimized. Hence, AUROC was the most important criteria used in the current study to determine the performance of the models.

2.8 Data Preprocessing

Data preprocessing is an important phase in analyzing data and ML. Data preprocessing involves transforming data into desirable state (Kononenko & Kukar 2007). Data preprocessing involves;

a) Data cleansing, a process of identifying and removing missing values, outliers, inconsistencies and noise.

University of Malaya

b) Data integration is a process that involves using multiple databases, data cubes and files, data transformation which involves normalization (scaling attribute values to fall within a given range).

c) Data reduction, which reduces the number of attributes, attributes values and tuples (i.e. sampling, clustering, histogram, and PCA and attribute selection).

Many ML models require preprocessed data before entering the model. Centering and scaling are forms of pre-processing numerical data where centering subtracts the mean of a variable from each data point for new features to have a zero mean. For the case of scaling, it multiplies each data point by a constant in order to alter range of the data. These transformations help improve the interpretability of parameter estimates when there is interaction in the model. Reducing data dimensions is a form of transforming data. This reduces features in a way of bringing together small or less important variables and remains with the more important ones.

Principal components analysis (PCA) is the most popular and commonly used technique for the problem of dimensionality reduction (Jollife 2010., Abdi et al., 2010). The goal of this method is to convert a larger set of correlated variables into a smaller set of uncorrelated or orthogonal variables that is named principal components. All the principal components are linear functions of the original variables, and the principal component formula is as follows.

$$PC_{j=a_{jI}}X_I + a_{j2}X_{j2} + \cdots + a_{j\rho}X_\rho \qquad (8)$$

Where p is the total number of original variables, and the coefficient for each variable is called component weight or loading. Smaller coefficient means that the corresponding variable makes less contribution to the principal component. During the principal

University of Malaya

component analysis, the first component PC1 accounts for the most variability in the original dataset of all the new principal components (Bowden et al., 1997). The subsequent component PCj is a different linear combination that represents the most remaining variability, under the restriction that it is uncorrelated or orthogonal to all previous components (Kong et al., 2017; Sanguansat, 2012). PCA has been used in the current study to analyze the relationship between the survival and non-survival.

2.8.1 Consistency and Over fitting

Overfitting occurs when the model is selected for best describing the training data but fails to fit new data (Chicco, 2017). To encounter over fitting, data generally are split into three parts: training set that is used for parameter estimation, a validation set that is used to select the model and a test set is used to generalize and evaluate the ability of the model selected. The comparisons for the performance of feature selection algorithms with their errors on validation set should be avoided. Such comparisons are problematic, since the model selection are trained on the validation set, which might reduce the generalization ability of the model and in turn widen the confidence interval on the generalization error. Searching for optimal parameter regularization (e.g. penalization on some norm of the parameters) and greedy criterion (e.g. early stopping in search) should be included in model learning to cope with overfitting (Tušar et al., 2017; Pham et al., 2008 and Webb, 2017).

Validation methods of model to avoid over fitting include substitution method, retention method (called holdout method) and cross-validation (CV) method. There are leave-one-out CV (LOO-CV), leave-more-out CV (LMO-CV) and *k*-fold CV. Leave-p-Out Cross Validation approach leaves more data points out of training data. Leave-one-out is the degenerate case of K-Fold Cross Validation, where K is chosen as the total number of examples. Leave-one-out cross-validation (LOOCV) is a particular case of leave-p-out

University of Malaya

cross-validation with p = 1. LPO cross-validation requires training and validating the model C_p^m times. K-Fold Cross validation is advantageous in a way that all the examples in the dataset are used for both training and testing. To generate a train set and validation set pair, one of the K parts out as the validation set is kept and the remaining K − 1 parts to produce the training set are combined. If each time one instance for the validation set is left out, this technique is called leave-one-out. Since there is a need of training K times, this technique can usually be applied only on small datasets. There is also evidence that K-fold cross-validation can give better results than leave-one-out. In this type of cross-validation, data is divided randomly into two equal parts. Another cross-validation technique is bootstrap. The basic idea is to randomly draw datasets with replacement from the training set where each sample contains the same number of instances as the training set. This is done K times producing K bootstrap datasets.

The table below shows the Cross-Validation approaches used in previous studies relating to mortality prediction to avoid over fitting. Cross-validation approach is also adopted in this study with k = 10.

University of Malaya

Table 2.4: Summary of Previous Literature Using Cross-Validation in Mortality studies.

Author	k-fold	No. of times	Performance criteria	ML model	Instances
Motwani et al., 2016	10-fold cross-validation	10 repeats	AUC	Logit Boost model, information gain	10030 patients
Wallert et al., 2017	7-fold cross-validation	3 repeats	AUC	SVM with RBK, Boosted C5.0, Logistic Regression, Random Forest, BORUTA	51,943 patients
Steele1 et al., 2018	3- folds	3 repeats	AUC	Cox models, random forests and elastic net	80,000 patients
Shouval et al., 2017	10-fold cross-validation	10 repeats	AUC	RF, LR, AdaBoost, ADTree, pruning rules-based classification tree (PART) and Naïve bayes	2782 patients
Sakr, et al., 2017	10-fold cross-validation	10 repeats	AUC	DT, SVM, ANN, RF, Naïve Bayesian Classifier (BC), Bayesian Network (BN), KNN	34,212 patients
Hoogendoorn et al., 2016	5-fold cross-validation	5 repeats	AUC	Cox model, LR, KNN	26,647 patients
Mohamadlou et al., 2018	3-fold cross validation	3 repeats	AUC	20 models including Gradient boosting & LR	111,459 patients
Kuo et al., 2018	10- fold cross validation	10 repeats	AUC	LR, SVM, DT	7,252 patients
Goldstein et al., 2016	10- fold cross validation	10 repeats	C-statistics	LR, RF, ANN, DT, LASSO, Ridge, KNN	1,944 patients

University of Malaya

2.9 Application of ML in coronary artery disease related study

ML method has been applied to various applications on heart diseases. Khalilia et al. (2011) proposed a method, which used Healthcare Cost and Utilization Project (HCUP) database to foretell a person's chances of getting a heart disease based on his medical record. They used 126 features and RF classifiers to foretell heart disease risk. Their study also assessed the performance of four algorithms including SVM, bagging, boosting and RF in predicting the chance of eight diseases. RF ensemble had the best performance over the other three ensembles with an average AUC of approximately 89.05%.

Özçift (2011) proposed a resampling approach for improving diagnosis of cardiac arrhythmia based on RF ensemble classifier. Outcome of the study showed that random sampling approach in RF ensemble classification technique is highly efficient.

Myers et al. (2014) carried out a research in predicting cardiovascular (CV) death in patients with heart failure (HF) using ANN. They used and tested ANN with a single hidden layer containing a varying number of hidden neurons and their study concluded that ANN could be used in mortality prediction.

Colak et al. (2008) produced and tested eight different ANN models from 237 patients who had been referred to the cardiology department for the purpose of CAD prediction. Seventeen predictor variables describing demographics, lifestyle and biochemical information were included in the models. Among eight networks used, the best performance was obtained with a model showing an accuracy of 92%, sensitivity of 96%, and specificity of 89%.

Green et al. (2006) compared ANN and multiple LR to predict ACS from 634 patients presenting in an emergency department with chest pain. Only 38 variables that were immediately available at patient presentation were used, including ECG data and clinical

University of Malaya

data. For each approach, the authors produced several models based on the variables used and construction method. When all 38 variables were used, ANN with the best performance had an AUC of 0.791 while the LR model had an AUC of 0.757. Nonetheless, when the variables were limited to 16 ECG data only, the network with best performance showed an increased AUC of 0.802, but the AUC of LR model decreased to only 0.705, indicating the presence nonlinearities in the ECG data that the LR model could not capture.

Ambarasi et al. (2010) attempted to predict the existence of heart illnesses. Thirteen demographic and medical variables were originally used in the prediction of the heart diseases. The researchers utilized GA to determine the variables that contribute more to the diagnosis of heart disease, such that the number of tests needs to be taken by patients was reduced, resulting in the selection of 6 variables. The investigators tested DT (J48), NB, and classification via clustering. Observations showed that DT outperformed the other two classifiers after incorporating the variable subset selection but took a longer time to build the model. The accuracy of three classifiers was 99.2%, 96.5%, and 88.3% for DT, NB, and classification via clustering, respectively. The results also showed that NB performed consistently both, before and after the reduction of variables with the same model construction time.

Palaniappan & Awang (2008) developed a prototype Intelligent Heart Disease Prediction System (IHDPS) using three classification techniques, namely, DT, NB and ANNs. A total of 909 cases with 15 demographical and medical variables were achieved from an U.S. based heart disease database. The database was equally split to training set and test set in a random manner. After complementing three learning schemes, NB appeared to be the most effective classifier as it had the highest accuracy (86.53%) for patients with heart disease, followed by ANNs (86.12%) and DT (85.68%) without much difference. DT was most effective for predicting patients without heart disease (89%)

University of Malaya

compared to ANN and NB. Researchers further concluded that all three models could be used to provide decision support to doctors for diagnosing patients and discovering medical factors associated with heart disease.

ML algorithms such as RF, SVM, DT, LR and many others have been used in previous studies for prediction in heart diseases and these are summarized in Table 2.5.

Table 2.5: Summary of Previous Studies on different ML Methods in HD Predictions.

Authors (Ref #)	Application	Methods	Input Variables	Result
Helwan et al., 2017	One-Year Survival Prediction of MI	BPNN, RBFN	11 parameters	RBFN, Accuracy96. 8%
Fox et al, 2006	Predicting risk of death after (6month) acutecoronary syndrome	C-statistics, GRACE risk score	Age, hypertension, PCI, pulse, CHF, SBP, st-depression, medical history etc.	AUC 80%
Masetic & Subasi, 2016.	Congestive heart failure detection using RF	C5.0, SVM, ANN, KNN, RF, and CART.	14 feature variables	Accuracy 99.9%
Khalilia et al., 2011	Predicting disease risks using random forest	RF,svm,bagging and boosting	126 variables including; age, admission month, parient'info, diagnosis.	AUC 89.05%
Özçift, 2011	Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis	RF (mtry), 10-fold cv,	14 classes, Pvc, ischemic changes, old anterior MI etc	Accuracy 90.0%
Bhatia et al., 2008	SVM Based Decision Support System for HeartDisease Classification	RBF kernel, SVM	age, sex, chest pain type, RBS, cholesterol, fasting blood sugar, max heart rate, exercise-induced angina, old peak, slope & no. of vessels colored	Accuracy 90.57%
Asadi et al.,2014	Machine Learning for Outcome Prediction of Acute Ischemic Stroke Post Intra-Arterial Therapy.	SPSS, MATLAB, ANN, SVM, RapidMiner, classical statistics	diabetes, hypertension, hypercholesterolemia, atrial fibrillation, history of ischemic heart and previous cerebral stroke or transient ischaemic attack	root mean squared error 2.064 (SD: 60.408) and AUC = 60%
Sen, 2017	Predicting and Diagnosing of Heart Disease Using Machine Learning	WEKA, DT, KNN, ANN, SVM, Naïve base Classifier	pressure, Abnormal blood lipids, smoking, Obesity, Physical inactivity, Diabetes, Age, Gender, Family history	Accuracy 83% - SVM
Mastoi et al., 2018	Coronary heart disease	SVM, NN, KNN, and RF classifiers, GMM	Age, gender, hypertension, RBS, chest pain type, SBP, DBP, height, weight.	Accuracy (svm)99%

University of Malaya

2.9.1 Application of ML in ACS mortality related study

The current study focuses on survival vs non-survival after ACS event. This section discusses previous literature related to mortality study using ML approach. Collazo et al. (2016) compared the performance of ANN and SVM models. These models were developed using predictors selected by wrapper method associated with filter on Euclidean distance. These models were to distinguish mortality of patients who were hospitalized with ACS. They used a sample of 264 individuals (17 deaths and 247 survivals) and 28 variables. Their computational results showed that SVM performed better than ANN. The most significant variables were age and Creatinine.

Motwani et al. (2016) predicted all-cause mortality in patients with suspected coronary artery disease. The cohort consisted of 10030 patients and 69 variables. ML models was compared with other mortality prediction methods that is Framingham risk score (FRS), segment stenosis score (SSS), segment involvement score (SIS), modified Duke Index (DI). ML outperformed all other methods in predicting all-cause mortality (ACM) with the AUC (ML: 0.79 vs. FRS: 0.61, SSS: 0.64, SIS: 0.64, DI: 0.62). The authors also performed feature selection for ACM and age ranked the highest among all other features. Hence their study shows that ML is better than other mortality prediction methods for all causes of mortality where ACS is considered as one of the major causes since it was based on clinical dataset.

A study by Steele et al. (2018) compared ML with traditional approaches on electronic health records. The cohort of 80,000 patients with 586 variables was used. The 27 variables were selected based on expert opinion. The results for variables selected by experts were compared with those selected by ML to predict all causes of mortality. EN outperformed Cox models and RF model with an AUC of EN 0.80. ML outperformed traditional methods and the variables selected by experts performed very poor than those selected by ML. All ML methods selected age as the most important predictor in

University of Malaya

predicting of all-cause mortality. They concluded that ML when applied on electric health records dataset could be useful in building models for clinical practices and identifying predictive predictors for all causes of mortality.

Shouval et al. (2017) predicted mortality of STEMI patients after 30 days. They compared different ML algorithms with conventional risk scores. In their study, the best model performed similarly to GRACE (0.87 SD 0.06) and outperformed TIMI score (0.82 SD 0.06). A total number of 2782 of patients with 54 variables were used in their study. The performance used in their study was AUC and RF out performed ADTree, ADAboost, LR and Naïve Bayes.

Wallert et al. (2017) compared four different ML algorithms in predicting 2-year survival vs. non-survival after first MI. They used data from Sweden population consisting of 51943 patients with 39 predictors that were collected during the year 2006-2011. The four ML methods were SVM with RBF kernel, boosted C5.0, LR and RF. SVM outperformed all other three ML algorithms with an AUC AUROC = 0.845, PPV = 0.280, NPV = 0.966. All ML performed well with a slight difference of 0.01. These algorithms performed feature selection and age was considered to be the most significant predictor among all the 39 variables used in the study. They concluded that ML model can be used in differentiating new patients and more predictors should be used to further predict survival versus non-survival after first MI. Table 2.6 below summarizes literature review on ACS mortality prediction.

University of Malaya

Table 2.6: Summary of Previous Studies on Mortality Prediction using ML.

Authors (Ref #)	Application	Methods	Instances	Input Variables	Result
Wallert et al., 2017	Predicting two-year survival versus non-survival after first myocardial infarction using machine learning and Swedish national register data	SVM with RBK, Boosted C5.0, Logistic Regression, Random Forest, BORUTA	51,943 patients	39predictors including; age, discharge statins, weight, heartrate, sex, discharge other antiplatelet etc, Smoking, Diabetes, Hypertension, stroke, ACE, A2 blockers, Beta blockers, Statins	SVM: AUROC = 0.845, PPV = 0.280, NPV = 0.966
Steele1 et al., 2018	Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease	Cox models, random forests and elastic net	80,000 patients	586 variables including; hypertension, diabetes, smoking, lipid profile, creatinine, HDL, TC, CKD, stroke, HbA1c, eGFR, ACE etc.	EN: AUC = 0.80
Motwani et al., 2016	Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis	Logit Boost model, information gain, segment stenosis score (SSS), segment involvement score (SIS), modified Duke index (DI) and Framingham risk score (FRS	10030 patients	69 variables including; age, sex, gender, BMI, CAD, DM, ejection fraction, HDL, HTN, LAD, LCX, LDL, RCA etc	ML: 0.79, FRS: 0.61, SSS: 0.64, SIS: 0.64, DI: 0.62
Collazo1 et al., 2016	A comparative study between artificial neural network and support vector machine for acute coronary syndrome prognosis	Euclidean distance,	411 patients	28 variables; age, creatinine and systemic arterial hypertension, gender,	
Shouval et al., 2017	Machine learning for prediction of 30-day mortality after ST elevation myocardial infraction: An Acute Coronary Syndrome Israeli	RF, LR, AdaBoost, AD Tree, pruning rules-based classification tree (PART) and Naïve Bayes	2782 patients	54 variables; creatinine, Kilip class on admission, blood pressure, glucose level, and age.	RF: AUC = 0.91

University of Malaya

Summary

ML and feature selection methods are presented in this chapter and this includes; Support Vector Machines (SVM), Random Forest (RF), Decision Tree (DT), Elastic Net (EN), logistic regression (LR), and learning vector quantization (LVQ). SOM was also introduced. Mortality prediction, conventional methods (risk scores) methods are also illustrated.

University of Malaya

CHAPTER 3: METHODOLOGY

Chapter three describes different approaches and materials used and is detailed as follows; study data, data preprocessing, model tuning, implementation and tools used for this study.

3.1 Study data

The national cardiovascular disease database (NCVD), registered patients who were admitted from 2014 to 2016 to all Coronary Care Units (CCU) in UiTM Sg Buloh Hospital for symptoms of ACS. Data for the study was obtained from the NCVD-ACS registry.

The data obtained was data of patients who were diagnosed with ACS (that is; unstable angina, ST segment elevation myocardial infarction (STEMI) and non-ST segment elevation myocardial infarction (NSTEMI)). Using a standardized case report form, data was collected from the time the patient with ACS was admitted to the hospital until discharge from hospital, between 2014 and 2016. A unique national identification number was assigned to each patient to avoid duplication. Follow-up was done 30 days and 90 days after hospital discharge via phone call or when the patient came to the clinic for a review. The patients' baseline characteristics and clinical presentation, in-hospital treatment, procedural details and clinical outcome were recorded.

The patients' information like ethnicity was determined based on self-report and the ethnicity stated on their national identity cards. STEMI was defined as persistent ST segment elevation $\geq$ 1 mm in two contiguous electrocardiographic leads, or the presence of a new left bundle branch block in the setting of positive cardiac markers. NSTEMI was defined as the occurrence of acute myocardial infarction in the setting of positive cardiac markers, with or without accompanying electrocardiographic changes other than ST-

University of Malaya

segment elevation (Adhikari & Baral, 2018). Unstable angina was defined as symptoms that were judged to be consistent with acute cardiac ischemia within 24 hours of hospital presentation with serial cardiac markers negative for myocardial infarction.

The overall dataset consists of 1480 patients with 77 features in the registry, 302 patients fulfilled inclusion criteria. Total numbers of 55 input variables were used in this study based on cardiologist recommendation and variables that reported less than 5 percent missing values. The final dataset for this study was 302 datasets with 55 input variables. UITM medical center's institutional review board approved the study data used. Table 3.1 provides explanation on characteristic of variables used in this study.

Table 3.1: Describes the Features of the dataset used in this study.

Variables Used	**Categories**	**Explanation of Variables**
Sociodemographic characteristics		
Age (yrs)	Year = 25 – 82	No. of years a person has lived
Gender	1 = male, 2 = female	A person's sex as male, female
Ethnicity	1= Malay, 2= Chinese, 3 = Indian ,4= others	Shared culture practices; Malay, Chinese, Indian, others
CVD diagnosis and severity		
PCI_type (Percutaneous coronary intervention)	0=no,1=primary, 2=rescue,3=pharmacoinvasive <72hrs,4=stage >72hrs	PCI is a non-surgical procedure used to treat narrowing (stenosis) of the arteries of the heart found in CAD
Combined_dm_acs_subtypes	0="no dm", 1="dm+unstable angina", 2="dm+NSTEMI", 3="dm+STEMI"	Patients with ACS and diabetics
Acs_subtype	1 = unstable angina, 2= NSTEM, 3 = STEMI, 4=OTHERS	Different types of ACS
Thrombolysis	1= yes, 2 = no	A treatment to dissolve dangerous clots in blood vessels.

University of Malaya

Table 3.1, Continued.

Variables Used	Categories	Explanation of Variables
Stk_successful	0= not applicable, 1= SUCCESFUL, 2=UNSUCCESSFUL	Thrombolysis successful
eGFR(mL/min)		This test measures the level of creatinine to tell how well the kidneys function.
Treatment-modality	0=medical therapy, 1=PCI, 2=CABG	A method used to treat a patient for a particular condition
LAD_stent; left anterior descending artery (LAD)	0=no stents,1=1 stent,2=2 stents,3=3 stents	LAD stands for left anterior descending artery. It is a coronary artery, which the name is given to arteries that supply the heart muscle with blood.
LCx_stent	0=no stents,1=1stent,2=2 stents,3=3stents	left circumflex coronary artery
RCA_stent = right coronary artery (RCA)	0=no, stents,1=1 stent,2=2 stents,3=3 stents	right coronary artery
LM_coros = Left Main (LM) coros	0=normal / <50%,1=>50% stenosis	Left main coronary artery
LAD_coros	0=normal / <50%,1=>50% stenosis	left anterior descending artery
LCx_coros	0=normal / <50%,1=>50% stenosis	Left circumflex (LCX) is an artery of the heart.
RCA_coros	0=normal / <50%,1=>50% stenosis	right coronary artery (RCA) is an artery originating above the right cusp of the aortic valve, at the right aortic sinus in the heart
CVD risk factors		
Smoker	1= yes, 0 = no	A person who smokes any tobacco product.
Ex-smoker	1= yes, 0 = no	Someone who was able to quit smoking.
Hypertension	1= yes, 0 = no	when blood pressure in the arteries is persistently elevated
diabetes mellitus	1= yes, 0 = no	disease that prevents your body from properly using the energy from the food you eat
TC (total cholesterol)		The total amount of cholesterol in the blood.
LDL (low density lipoprotein)		a class of lipoproteins of relatively low density

University of Malaya

Table 3.1, Continued.

Variables Used	Categories	Explanation of Variables
HDL (high density lipoprotein)		a class of lipoproteins of relatively high density
FBS (mmol/L)		A test which determines how much glucose(sugar) is in blood sample after an overnight fast
newly_dm	1= yes, 0 = no	newly diagnosed diabetes mellitus
Dyslipidemia	1= yes, 0 = no	A condition where there is an abnormal amount of lipids, including cholesterol and / or triglycerides in the body.
Obesity	1= yes, 0 = no	The state of being grossly fat or overweight.
Alcohol	1= yes, 0 = no	An intoxicated Liquid that is produced by natural fermentation of sugars
LCLsimon_broome	1=>4.9, 2<4.9	Diagnoses familial hypercholesterolemia (FH) based on clinical, genetic and family history.
TCSimon_broome	1=>7.5, 2=<7.5	Diagnoses familial hypercholesterolemia (FH) based on clinical, genetic and family history.
Medicines		
Clopidogrel	1=yes, 0=no	A medicine used to prevent heart attacks and strokes in persons with heart disease (recent heart attack), recent stroke, or blood circulation disease.
Ticagrelor	1=yes, 0=no	A class of medications called antiplatelet medications, which prevents platelets (a type of blood cell) from collecting and forming clots that may cause a heart attack or stroke.
ARB	1=yes, 0=no	Angiotensin II Receptor Blokers
ACE (Angiotensin Converting Enzyme)	1=yes, 0=no	ACE is a central component of the renin–angiotensin system (RAS), which controls blood pressure by regulating the volume of fluids in the body
CCB (Calcium Chanel Blocker)	1=yes, 0=no	A drug that blocks the entry of calcium into the muscle cells of the heart and the arteries
B_blockers	1=yes, 0=no	Are medicines used to treat abnormal heart rhythms, specifically to prevent abnormally fast heart rates
Statins	1=yes, 0=no	A class of drugs that lowers the level of cholesterol in the blood
Statin_doz(mg)	1=10mg, 2=20mg,4=40mg,5=60 mg,6=80mg	Statin dosage

University of Malaya

Table 3.1, Continued.

Variables Used	Categories	Explanation of Variables
Statin med	1=lovastatin, 2=simvastatin,3=pravastatin, 4=atorvastatin, 5=rosuvastatin	Statin medication
ASA (drug caution code)	1=yes,0=no	A medication that indicates it contains acetylsalicylic acid (aspirin)
Nitrates	1=yes, 0=no	Nitrates are medicines that ease and prevent angina pains.
CVD comorbidity		
CCF (congistive cardiac failure)	1= yes, 0 = no	Inability of the heart to pump blood with normal efficiency
Ihd	1= yes, 0 = no	Ischemic Heart Disease
fx_IHD	1= yes, 0 = no	Fracture Ischemic Heart Disease
Non-CVD comorbidities		
Stroke	1= yes, 0 = no	Sudden death of brain cells due to lack of oxygen.
Coad (chonic obs airway diease)	1= yes, 0 = no	A chronic inflammatory lung disease that causes obstructed airflow from the lungs.
Ba (broncha asthma)	1= yes, 0 = no	A chronic inflammatory disease of the airways that causes periodic attacks of wheezing, coughing, shortness of breath and chest tightness.
Ckd(kidney failure)	1= yes, 0 = no	It is the gradual loss of kidney function.
Biomarker		
HbA1c (haemoglobin A1c)	Percentage = (0-15)	A test for diabetes patients which shows the average level of blood sugar
Creatinine		A compound, which is produced by metabolism of creatinine and excreted in the urine.
Troponin I		A cardiac and skeletal muscle protein useful in the laboratory diagnosis of heart attack
TG (triacylglycerol)		Are markers for several types of thermogenic
CK (Creatinine kinase)		Test which measures the amount of an enzyme

3.2 Data preprocessing

Data pre processing involves removing all the variables with missing values and outliers. Those variables with small percentage (less than 5%) of missing values was mantained. No data imputation was carried out and the actual data was used in this study. The pre-processing stage takes several processes including dealing with outliers, unbalanced data and missing data. All less significant variables were removed based on expertise opinion. The number of predictors was reduced based on Consultant cardiologist opinion. The initial selection of features was completed, where all predictors were kept if indicating possible future cause of mortality for example HbA1c, FBS,

University of Malaya

Troponin, and Creatinine. Predictors considered less important for the outcome (N = 10) were removed. Predictors with 80% missing values (N = 6) were also removed and predictors with zero variance (N=6) were also removed. The study ended up with a full set of 54 features (11 continuous, 43 categorical) without the output.

3.2.1 Classification and sample pre-processing

Prior to model development, data was centered and scaled as some variables have large variation or spread. To avoid biasness in results, stratified random sampling of data was used (Kuhn & Johnson, 2013). Data was split for model train (70%) and test (30%). 10-fold CV with three repeats were used to avoid over fitting for model development on training set. A 10 and 15 length tune-grid search ML was used for model parameter tuning to select model parameters with highest performance (Ramadhan et al., 2017; Kuhn & Johnson, 2013). Random down and up-sampling of the majority and minority class was performed as the data set was highly unbalanced (Menardi & Torelli, 2014).

3.2.2 Model tuning

To encounter overfitting, obtain better results, 10-fold CV with three iterations was used to develop on the train set. Within this resampling, A 10 and 15 length tune-grid search, (2) ROSE sampling where both random down and up-sampling of the majority and minority class was performed as the data set was highly unbalanced (Menardi & Torelli, 2014), (3) centering and scalling was applied.

Grid search is used to evaluate k-length of consistently value increase of the tuned parameters of the model. The value of parameters is later selected basing on train dataset that performs higher than others. The same train dataset is used in constructing final models on all train dataset and test dataset (Kuhn & Johnson, 2013; Kuhn, 2008). AUC metrics were used in the current study on train and test due to the problem of un-balanced

University of Malaya

dataset. Parameter tuning was also carried out on all models. Figure 3.1 illustrates ML models' parameters tuning process.

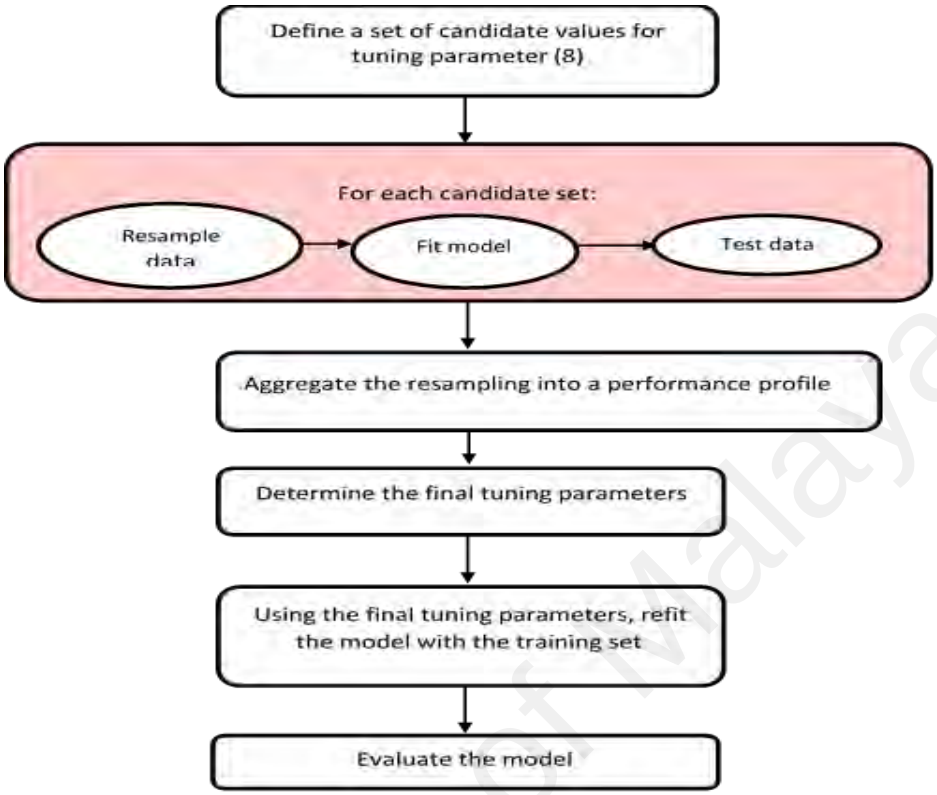

Figure 3.1: Parameter-Tuning Process.

3.2.3 Training and testing dataset

Training set is the portion of data in which the model is trained. In this study, 70 percent of data was used for training. To avoid performance error, 10-fold cross validation was used. This helps to minimize the over fitting and under fitting of the data. PCA was also used in this study to determine the relationship between survival and non-survival.

Testing set is the portion of data where the model is tested, it is often the dependent variable of the data. In this study, 30 percent of the data was used for testing. When cross validated data is tested, it performs better or worse depending on the model used. To ensure every model is functioning in its optimum, a technique named parameter tuning was used as explain in the previous section.

University of Malaya

3.2.4 Rose Algorithm: Balancing dataset

One of the main issues encountered with the dataset used in this study is that it is imbalanced. In particular, the dataset included 279 records (survival) and 23 records (non-survival). In general, the predication accuracy is significantly affected with imbalanced data. Classifiers algorithms for example LR, RF, SVM have the tendency of producing biased results if one of the classes have a higher number of instances. Due to this characteristic, classifiers often ignore minority class features treating them as noise. Therefore, there is a high likelihood of misclassification of the minority class to the majority class.

In order to handle the imbalanced dataset used in this study, ROSE algorithm from R software was implemented on the data set with "both" sampling method which is a combination of both oversampling and under sampling methods. Using this method, the majority class is under sampled without replacement and the minority class is oversampled with replacement.

3.3 ML Algorithm

Prediction and feature selection were performed using 10 feature selection methods on all 54 variables as explained in Table 3.1. RF (Breiman, 2001), DT (Quinlan, 1986), LR (Menard, 2002), and EN (Zou & Hastie, 2005) algorithms comprises of features selection and prediction capabilities. RFE (Jafarian et al., 2011), GA (Holland, 1992), BORUTA (Miron et al., 2010), LVQ (Kohonen, 1995), and CD using Euclidean distance (Cox, 1958) are feature selection algorithms without prediction capabilities and were combined with RF and SVM classifier algorithms for mortality prediction after ACS.

Feature selection comprises of variable ranking, identification and elimination of irrelevant and redundant information. This process of dimensionality reduction increases ML algorithm performance and prediction results. Feature selection was carried out using

University of Malaya

SBS method based on the ranked variables selected by separate feature selection algorithms in a descending order iteratively (Genuer et al., 2010). The prediction models were trained and tested for each iteration, and the models with highest performance were selected. Predictive performances of the prediction model were calculated and averaged using untouched testing dataset that was not down sampled for model validation. AUC was used as predictive performance metrics that is insensitive to class imbalances (Fawcett, 2006).

3.3.1 Random Forest (RF)

RF proposed by Breiman (2001) for classification and variable importance was implemented in this study. RF is an ensemble method that builds many decision trees randomly from boostrapping samples which are then clustered together by classification method (Breiman, 2001). Only a subset of predictor is randomly chosen from the full set of predictors, p at each tree node which is denoted by *mtry* (Díaz-Uriarte & De Andres, 2006). Gini index node of impurity calculated based on a set of predictors is used in RF to select best split at each node (Khalilia et al., 2011). Test set error estimate is obtained from growing a tree from bootstrap data which then is used to estimate variable importance (VarImp) which are the useful by-products of RF. RF algorithm implemented in this study was based on Breiman (2001). Varying value of *mtry* (mtry = 5; 7, 10, 15, 20) and number of trees *ntree* (500 – 4000) was used in this study to determine optimum RF model that produced the best results. The RF VarImp method was used to generate ranked variables that were then reduced using SBS iteratively. The algorithm of RF is shown below:

1. Each tree of RF is grown on a bootstrap sample of the training set.
2. At each node, n number of variables is chosen randomly out of N predictors when growing a tree.

University of Malaya

3. It is suggested, the value of n starts with n=√N and then increase it until the smallest error of the OOB is obtained. At each node, one variable with the best split is used for all value of n.
4. RF method is then applied for testing data for prediction.
5. This step is repeated using different mtry argument starting from (mtry = 5; 7, 10, 15, 20) where default value of classification RF is p1/2 = 7
6. The RF is repeated with different argument starting from ntree = 500; ntree = 1000; ntree = 1500; ntree = 2000; ntree = 2500; ntree = 3000 and ntree = 4000. This is done to examine the sensitivity to method argument mtry and ntree to better determine important variables and the stability of the variable importance score.
7. The RF variable importance method was used to generate ranked variables that were then reduced using sequential backward selection and prediction iteratively.

3.3.2 Support Vector Machines (SVM)

SVM proposed by Cortes and Vapnik, (1995) was implemented in this study using RBF kernel using e1071 package for parameter tuning. The tuning parameters for SVMs are the C parameter (cost), which regulates the margin width, and the gamma-parameter for the kernel calculation. The C or cost tuning parameter is essentially the regularization parameter for the kernels. SVM in this study uses receiving operating curve (ROC) variable importance to select and rank important variables. For two class problems, a series of cut-offs is applied to the predictor data to predict the class. The sensitivity and specificity are computed for each cut-off and the ROC curve is computed. The trapezoidal rule is used to compute the area under the ROC curve. This area is used as the measure of variable importance (Meyer, 2015). The parameter tuning that was used is sigma 0.5 and cost 10 using grid search for SVM classifier as shown in table 3.2.

University of Malaya

3.3.3 Decision Tree (DT)

DT was constructed using *caret* and rpart function (Therneau et al., 2015). DT has in built feature selection and classification method and it is an embedded approach. DT was implemented in this study according to (Barlow & Neville, 2001). DT were built for pruned and unpruned tree. Tree pruning was used to produce good predictions on the training set. The lowest error test rate was selected as the best pruned tree. Printcp (model) function was used to display cost-complexity parameter (cp). Prp (fit) was used to plot automatically generated variables as a DT diagram. A tune length value of 10 was used to avoid overfitting. The following are the steps that was carried out in DT.

1. Determine the size of the tree based on its relative error
2. Calculate the number of trees split, errors and standard deviation.
3. Variable importance was obtained.
4. DT were built for pruned and unpruned tree.
5. Rpart process is later used a test dataset for classification.

3.3.4 Logistic Regression (LR)

LR (Kleinbaum et al., 2014, 2008) was used for both prediction and feature selection. The method glm () with family binomial was used in the current study. The parameter tuning was carried out with tune length = 15. LR was constructed with package 'glmnet' (Friedman et al., 2010).

3.3.5 Elastic Net (EN)

EN optimizes the coefficients until the change of the coefficients is smaller than a predetermined toleration value. EN was constructed in this study using cv. glmnet () function as recommended by Friedman et al. (2010). EN was used for both feature selection and prediction. Parameter tuning with maxit = 1000000, alpha = 0.5, lambda =

University of Malaya

100 was done to improve on the model performance.

3.3.6 Genetic algorithm (GA)

Genetic algorithm was utilized in this study for feature selection only. The choice of parameter settings for GA in this study were experimentally determined through parameter optimization. The parameter settings play an important role in the development of accurate classification models (Tay et al., 2014, 2013). The details are; different population size: 20, 30, 50, 100, 275, 300 and 500; was determined using trial and error method until the best population size was met. Trial and error method were used to determine number of iterations starting from 50, 100, 200, 300 and 500. The final feature selection was conducted using population size = 275 and iteration = 200.

3.3.7 Learning vector quantization

LVQ (Grbovic & Vucetic 2009) (Nova et al., 2015) was used in this study only for feature selection. The LVQ is a classification model where the classification of a given vector is equivalent to find the class label of the nearest prototype of vector. The prototypes are the neurons learned with the LVQ in the learning phase. At the starting point neurons are initialized in a random way from the training set and such type of classification is equivalent which base on the prototypes constructed by the LVQ learning (Pedreira et al., 2006). LVQ is known to be simple since the system does not need expert for feature selection and feature importance cannot change during system operation thus a suitable method for feature selection in the current study (Thakare & Patil, 2014; Saulnier et al., 2011). LVQ has two parameters which increase its performance incase tunes and these are k and size. K is the number of instances to check when making predictions and size is the number of instances also known as codebooks in the model. These were used in this study to improve its performance with Size = 35 and k = 5 that

University of Malaya

were selected by grid search as mentioned in Table 3.2. LVQ was then combined with RF and SVM classifiers for prediction.

ML prediction model's parameter tuning was conducted to select model parameters with highest performance in this study (Kuhn et al., 2013). Table 3.2 illustrates selected parameters for optimized model performance in this study. All algorithms used similar number of folds = 3 and cross validation value k =10 and data were pre-processed for model development and prediction.

Table 3.2: Machine Learning Model Parameters.

Algorithm	Grid search	Parameter tune
RANDOM FOREST	mtry=c (1:15), ntree=c(500, 1000, 1500, 2000, 2500, 3000, 3500, 4000)	Number of trees = 1000, mtry = 7
RECURSIVE FEATURE ELIMINATION		
BORUTA		doTrace = 2
GENETIC ALGORITHM	Iteration = 50, 100, 200, 300, 500, Population size = 20, 30,50,100, 275, 300	Iteration =200, population size = 275
LEARING VECTOR QUANTITIZATION	size=c (5,10,15,20,25,30,35,40,45,50), k=c (3,5,7,10)	Size = 35 and k = 5.
Decision Tree		tune Length = 10
ELASTIC NET	lambda.grid <- seq(0, 100,150) alpha.grid <- seq(0, 0.5,1, length = 10)	maxit = 1000000, alpha = 0.5, lambda = 100.
LOGISTIC REGRESSION		tuneLength = 15
SVM Radial	Sigma = c (0.01, 0.5, 1, 1.5, 2.5, 3), C = c (.25, 5, 6, 7, 10)	Sigma = o.5, 0.25, cost = 10
SVM Polynomial	Degree = 0.5,1,2,3, scale = 0.5,1,1.5,2, cost = 0.1,0.25,1,1.5,2,5,8,10	Degree = 1, scale = 1, cost = 1
SVM Linear	Sigma = c (0.1, 0.5, 1, 2, 2.5, 3), C = c (.25, 5, 6, 7, 9, 10)	Gamma = 0.5, cost = 10.

University of Malaya

3.3.8 Self-Organizing Map (SOM)

Self-organizing map (SOM) was generated by using toolbox in MATLAB VER. (R2013, Math Works). SOM (Kohonen, 1985) was used in this study to ordinate factors affecting ACS from optimized features selected from best performing model. The Euclidian distance between the input factors are calculated and visualized as U-matrix (unified distance matrix) as a result from the trained SOM. The U-matrix represents the distance between neurons. The winning neuron is selected based on neuron that responds greatly to a given input vector where the winning neuron and maybe its neighbor can learn by altering the weights in a way to furthermore reduce the Euclidean distance among the weight and the input vector via the equation. SOM reduces data dimension and display similarity by producing 1 or 2 dimensions and group similar inputs together. Light colour representing clusters while dark clusters representing cluster separator. The SOM is built using variables with higher scores until a suitable error is achieved. SOM evaluation is represented by the topological and quantization error and the best map is expected to have the smallest average quantization error (Uriarte & Martín, 2005).

SOM algorithm implemented in this study was based on Kohonen (1985) and is as follows;

1. Randomly initialise all weights
2. Select input vector $\mathbf{x} = [x_1, x_2, x_3, x_4, x_5, x_6]$
3. Compare **x** with weights **w** for each neuron j to determine winner
4. Update winner so that it becomes more like **x**, together with the winner's neighbours
5. Adjust parameters: learning rate & 'neighbourhood function
6. Repeat from (2) until the map has converged (i.e. no noticeable changes in the weights) or pre-defined number of training cycles have passed

University of Malaya

3.4 Model Evaluation, Validation And Performance Measures

AUROC is used to evaluate the performance of the model where the highest area that can be achieved is one. AUROC is mostly used in a situation where the values of all the class are not known hence the performance of the models is compared based on the AUROC obtained.

AUROC was the basic measure of model performance though other indicators were also considered for model assessment as it is known to be unbiased by an unbalanced sample. Additional performance matrixes used in this study are classification accuracy sensitivity (true positive rate) and specificity (true negative rate). Accuracy is the correctly classified instances to total instances used in the test set.

True Positive (TP) refers to the number of high risk patients who are classified as high risk, whereas False Negative (FN) refers to the number of high risk patients who are classified as low risk patients. On the other hand, False Positive (FP) refers to the number of low risk patients who are classified as high-risk patients and False Negative (FN) refers to the number of low risk patients who are classified as low risk patients. All results of the different metrics are then averaged to return the final result.

3.5 Feature Selection

Feature selection was used in this study to decrease dimensionality of the dataset through eliminating unrelated and less significant features. This helps in easy understanding and interpretation of the model and reduces on train time and this helps in reducing overfitting of the model (Han et al., 2004).

Feature selection comprises of variable ranking, identification and elimination of irrelevant and redundant information. This process of dimensionality reduction increases ML algorithm performance and prediction results. Feature selection was carried out using

University of Malaya

SBS method out based on the ranked variables selected by separate feature selection algorithms in a descending order iteratively (Arauzo-Azofra et al., 2007; Genuer et al., 2010). The prediction models were trained and tested for each iteration, and the models with highest performance were selected. Predictive performances of the prediction model were calculated and averaged using untouched testing dataset that was not down sampled for model validation. AUC was used as predictive performance metric that is insensitive to class imbalances (Fawcett, 2006).

To identify the most significant features for ACS mortality prediction, variables that were not significant were omitted based on the ranked features by different feature selection methods. This reduces data dimensionality improved performances of the ML prediction model.

Feature selection methods without prediction capabilities (BORUTA, GA, LVQ, RFE and CD) were combined with RF and SVM classifier algorithms for mortality prediction after ACS and RF and SVM performance were later compared to determine the model which would perform better in prediction of mortality after ACS. DT, EN and LR were used for both feature selection and model development. Feature selection methods are; filter, wrapper and embedded as explained in chapter 2 was used in this study for feature selection and these are explained more below.

3.5.1 Filter Feature Selection Method

Filter methods ranks variables one by one basing on the univariate search method and selects variables with high rankings. In this study cluster dendrogram (CD) a filter feature selection method is used. CD calculates the correlation between features in terms of Euclidean distance (Agarwal et al., 2010). This method produces a tree like diagram to illustrate the arrangement of clusters called dendrogram. In a dendrogram, the height of the lines indicates the distance between the objects that are connected. Feature selection

University of Malaya

using CD was obtained by cutting the dendrogram at the desired level where each connected component forms a cluster and features were selected based on the levels of each cluster. Features selected from each CD level were then combined with SBS method and RF and SVM classifiers. Filter feature selection was implemented in the current study as described in Figure 3.2.

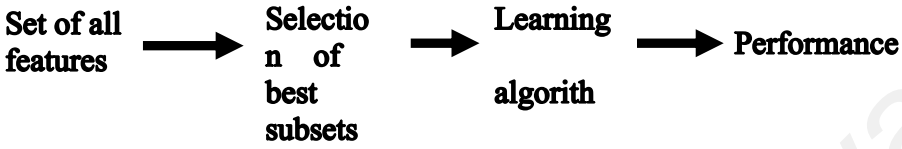

Figure 3.2: Process of Filter feature selection method.

3.5.2 Wrapper method

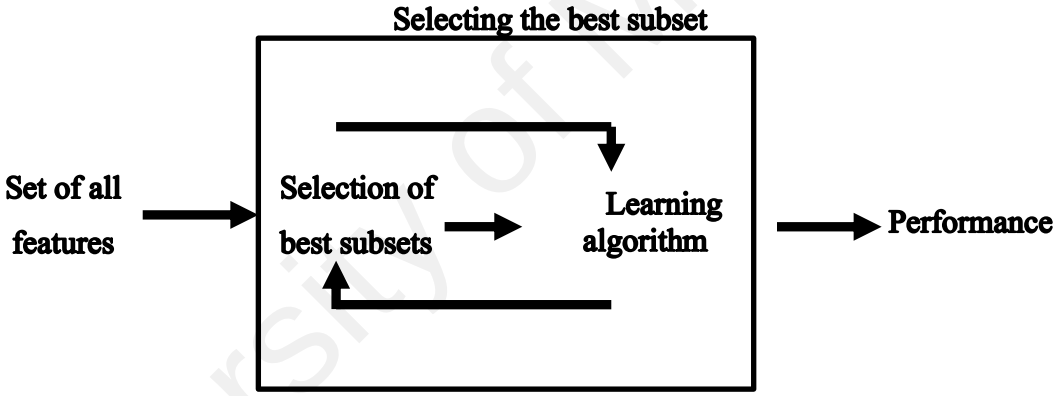

Figure 3.3: Process of Wrapper Method.

Figure 3.3 illustrates wrapper feature selection from the whole dataset where the process involves a step by step selection of features followed by Sequential backward selection (SBS) or forward method based on ranked variable.

A specific classifier is used to find the subset of features in wrapper feature selection methods with a good performing model. Feature selection takes into account the contribution to the performance of a given type of classifier. Wrapper methods implemented in this study were Sequential Backward Selection (SBS) (James, 2013),

University of Malaya

Recursive Feature Elimination (RFE), Boruta algorithm (Miron et al., 2010), RF and SVM.

RFE was implemented in this study as proposed by Jafarian et al. (2011). RFE method aims at finding a minimal and best performing set of variables. RFE feature selection method was used to fit SVM and RF model in this study. RFE removes least important features until the specified numbers of features were reached. Features were ranked using RFE feature importance function, and by recursively removing a small number of features at each iteration, RFE removes dependencies and collinearity that might occur in the model.

GA is also a wrapper method proposed by Holland (1992) and it was used in this study as explained in previous section.

SBS is performed in this study to identify variables that significantly affects mortality after ACS based on ranked variables by feature selection method in this study RF, SVM, LR, EN, GA, CD, LVQ. Feature reduction is carried out based on the ranked variables selected by different feature selection algorithms in a descending order iteratively (Genuer et al., 2010). SBS algorithm relies only on significance as a sufficient condition to remove insignificant variables from a model (Dunkler et al., 2014). The variable that causes significant increase in AUC in the testing dataset of the prediction model is deemed as important. The prediction models were trained and tested for each iteration, and the models with highest performance were selected. Model predictive performance was calculated and averaged using untouched testing dataset for validation purpose. Area under the Curve (AUC) was used as predictive performance metric that is insensitive to class imbalances (Fawcett, 2006).

This study has adopted the Boruta proposed by Miron, et al. (2010) to select most important features. The reason behind the selection of this approach is to be able to

University of Malaya

compare the significance of the real predictor variables with those of random shadow variables with the aid of statistical testing and multiple runs of RF.

Highlighted below is the stepwise command of Boruta algorithm:

1. Firstly, randomises the given data set by creating shuffled copies of all features otherwise known as shadow features.
2. Then, it trains a RF classifier on the extended data set and applies a feature importance measure (the default is Mean Decrease Accuracy) to evaluate the importance of each feature where higher means more important.
3. At each iteration, it checks whether a real feature has a higher importance than the best of its shadow features (i.e. whether the feature has a higher Z score than the maximum Z score of its shadow features) and constantly removes features which are deemed highly unimportant.
4. Finally, the algorithm stops either when all features gets confirmed or rejected or it reaches a specified limit of RF runs.

do Trace which simply means the verbosity level where 0 means no tracing, 1 means reporting attribute decision as soon as it is cleared and 2 means all of 1 plus additionally reporting each iteration. The default is 0. In this study do Trace=2 was used to improve Boruta's algorithms performance. RF and SVM for classification were constructed using features selected by Boruta algorithm.

University of Malaya

3.5.3 Embedded Method

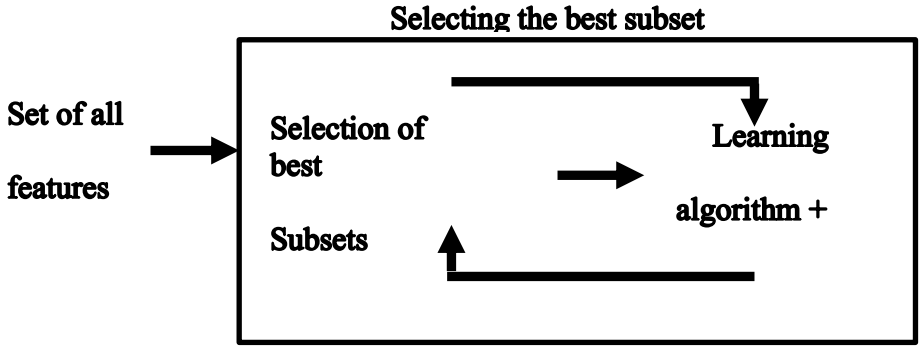

Figure 3.4: Process of Embedded Method.

The embedded feature selection methods as illustrated in Figure 3.4 perform variable selection as part of the learning procedure and are specific to a given learning machines. Embedded feature selection method comprises of the search for a maximum subset of characteristics that are incorporated into the classifier construction. This can be observed as a search in the combined space of feature subsets and hypotheses. Embedded method implemented in this study are DT proposed by Quinlan (1986), LR proposed by Menard (2002), and EN proposed by Zou and Hastie (2005). LVQ which is an embedded method proposed by Kohonen (1995) was used in this study only for feature selection and features obtained from this method were used on RF and SVM for prediction.

3.6 Software

The analysis and development of ML models was performed using R statistical software (R Core Team, 2014, version 3.4.3). SOM was developed by using MATLAB (version 16b). LR with package 'glmnet' (Friedman et al., 2010). RF classifier with 'randomForest' package (Liaw & Wiener, 2002); DT with the 'rpart' package (Therneau et al., 2015); EN with the 'glmnet' (Friedman et al., 2010); LVQ with 'class' (Venables & Ripley, 2002); and SVM with the 'e1071' package (Meyer et al., 2015). Model tuning was conducted with the 'caret' (Kumar, 2018) and 'e1071'packages depending on the

University of Malaya

model; visualizations were created with the 'ggplot2' (Wickham, 2009), 'corrplot' (Wei & Simko, 2016, gbm Gradient boosting machines with 'gbm' (Ridgeway, 2013)".

3.6.1 Additional Statistics

The data contains attributes such as patient number, gender, age, admission type, time in hospital, medical specialty of admitting physician, number of lab test performed, HbA1c test result, diagnosis, number of medication, diabetic medications, number of outpatient, inpatient, and emergency visits during hospitalization, etc." (Strack et al. 2014).

Result are expressed as average and standard deviation (SD) for parametric variable and as frequencies for non parametric variables. Correlation analysis were carried out to identify significant relationship between variables. Chi-Square test is used to carry out univariate analysis in order to identify important features, while data cleansing and statistical analysis were conducted using Welch's t-tests (p b 0.05) and Statistical Package for Social Sciences (SPSS) program version (v.21, IBM)

3.7 Summary of Design

This chapter provides the breakdown of the experiment carried out for the thesis. The strength and limitations of the experiments are also proposed in this chapter. The chapter starts with a brief description of the dataset including the variable types and data source. Data pre-processing techniques were briefly explained which were used in cleaning and normalizing the data to make it suitable for ML models. These pre-processing techniques include ROSE algorithm to balance dataset.

University of Malaya

CHAPTER 4: RESULTS

In this chapter, the results of implementations and empirical considerations are demonstrated. The results on various approaches carried out in this study are reported in detail.

4.1 Statistical Results

Table 4.1 illustrates summary statistic of the variables used in this study. Values which are mean ± SD or median, Confident interval (CI = 95%) derived from SPSS; Uncorrected P-values are from Welch's t-tests if variable is continuous, or Pearson-chi-square if categorical.

University of Malaya

Table 4.1: Summary Statistic of Variables used in this study**.**

Predictors (n = 55)	All cases (n =302)	Survivors (n = 279)	Non-survivors (n = 23)	p-value	Confident interval (CI= 5%)
		Socio-demographic characteristics			
Age (yrs)	56.72 ± 11.7	56.5 ± 11.5	58.7 ± 14.3	0.001	55.40-58.06
Age group					
29-55	131(43.3%)	123(44.1%)	8(34.78%)		
55-65	101(33.44%)	94(33.69%)	7(30.43%)		
Above 65	70(23.17%)	62(22.22%)	8(34.78%)		
Gender					
Male	206(68.2%)	189(67.7%)	17(73.9%)	0. 001	
Female	96(31.7%)	90(32.2%)	6(26.0%)		
Ethnicity					
Malay	159 (52.6%)	146 (52.3%)	13(56.5%)	0. 001	
Chinese	28(9.27%)	26(9.31%)	2(8.69%)		
Indian	104(34.43%)	97(34.7%)	7(30.43)		
Others	11(3.64)	10(3.58%)	1(4.34%)		
		CVD diagnosis and severity			
Nitrates	182(60.2%)	170(60.9%)	12(52.1%)	0. 001	
PCI_type	286(94.7%)	263(94.2%)	23(100%)	0. 001	
Combined_dm_acs _subtypes	122(40.3%)	112(40.1%)	10(43.4%)	0.612	
Thrombolysis	279(92.3%)	258(92.4%)	21(91.3%)	0. 001	
Stk_successful	279(92.3%)	258(92.4%)	21(91.3%)	0. 001	
LCLsimon_broome	281(93.0%)	259(92.8%)	22(95.6%)	0. 001	
TCSimon_broome	290(96.0%)	269(89.0%)	21(91.3%)	0. 001	
Acs_subtype				0. 001	
Unstable angina	170(56.29%)	161(57.7%)	9(39.1%)		
Nstemi	77(25.49)	67(24.01%)	10(43.47%)		
Stemi	51(16.88%)	47(16.84%)	4(17.39%)		
Others	4(1.32)	4(1.43%)	0(0%)		
		CVD risk factors			
Smoker	232 (76.8%)	212 (75.9%)	20 (86.9%)	0. 001	
Ex-smoker	270(89.4%)	248(88.8%)	22(95.6%)	0. 001	
Hypertension	230 (76.1%)	210 (75.2%)	20(86.9%)	0. 001	

University of Malaya

Table 4.1, Continued.

Predictors (n = 55)	All cases (n =302)	Survivors (n = 279)	Non-survivorss (n = 23)	p-value	Confident interval (CI = 95%)
Alcohol	292(96.6%)	270(96.7%)	22(95.6%)	**0.001**	
diabetes mellitus	180(59.6%)	167(59.8%)	13(56.5%)	0. 001	
newly_dm	294(97.3%)	271(97.1%)	23(100%)	0.005	
TC (mmol/L)	4.89 ± 1.31	4.92 ± 1.30	4.52 ± 1.48	0. 001	4.74-5.04
LDL (mmol/L)	3.04 ± 1.13	3.07± 1.13	2.70 ± 1.11	0. 001	2.92-3.17
HDL (mmol/L)	1.02 ± 0.27	1.03 ± 0.26	0.94 ± 0.28	0.103	0.99-1.05
Ticagrelor	293(97.0%)	270(96.7%)	23(100%)	0. 001	
FBS (mmol/L)	7.98 ± 3.43	7.97 ± 3.37	8.13 ± 4.19	0. 001	7.59-8.37
eGFR(mL/min)	66.55 ± 34.11	67.33 ± 33.77	56.99 ± 37.39	0. 001	62.6-70.4
ACE	204(67.5%)	191(68.4%)	13(56.5%)	0. 001	
CCB	225(74.5%)	206(73.8%)	19 (82.6%)	0. 001	
B_blockers	221 (73.1%)	206 (73.8%)	15(65.2%)	0. 001	
Treatment-modality	277(91.7%)	256(91.7%)	21(91.3%)	0. 001	
Clopidogrel	235(77.8%)	221(79.2%)	14(60.8%)	0. 001	
ARB	283(93.7%)	260(93.1%)	23(100%)	0. 001	
Statins	292(96.6%)	269(96.4%)	23(100%)	0. 001	
Statin_doz(mg)	164(54.3%)	150(53.7%)	14(60.8%)	0. 001	
Statin med	147(48.6%)	141(50.5%)	6(26.0%)	0. 001	
ASA	276(91.3%)	255(91.3%)	21 (91.3%)	0. 001	
Psychosocial characteristics					
LM_coros	293(97.0%)	272(97.4%)	21(91.3%)	0.003	
LAD_coros	267(88.4%)	247(88.5%)	20(86.9%)	0.001	
LCx_coros	271(89.7%)	251(89.9%)	20(86.9%)	0.019	
RCA_coros	271(89.7%)	252(90.3%)	19(82.6%)	0. 001	
LAD_stent	290(96.0%)	268(96.0%)	22(95.6%)	**0.001**	
LCx_stent	296(98.0%)	274(98.2%)	22(95.6%)	0.019	
RCA_stent	297(98.3%)	275(98.5%)	22(95.6%)	0.025	
CVD comorbidity					
Stroke	284(94.0%)	266(95.3%)	18 (78.2%)	0. 001	
Ihd	178(58.9%)	165(59.1%)	13(56.5%)	0. 001	
fx_IHD	267(88.4%)	245(87.8%)	22(95.6%)	0. 001	
Non-CVD comorbidities					
CCF	270(89.4%)	252(90.3%)	18(78.2%)	0. 001	
Coad	287 (90.0%)	265 (94.9%)	22(95.6%)	0. 001	
Ba	282(93.3%)	262(93.9%)	20(86.9%)	0. 001	
Obesity	296(98.0%)	275(98.5%)	21(91.3%)	0.014	
Ckd	274(90.7%)	255(91.3%)	19(82.6%)	0. 001	
Dyslipidemia	176(58.2%)	160(57.3%)	16(69.5%)	0. 001	
Biomarker					
HbA1c (mmol/mol)	7.46 ± 2.23	7.49 ± 2.26	7.10 ± 1.85	0. 001	7.20-7.71
Creatinine(µmol/L)	103.44 ± 61.5	101.5 ± 57.63	126.2± 96.22	0. 001	96.47-110.4
Troponin_1(ng/l)	12.38 ± 135.1	4.60 ± 18.47	106.7 ± 485.3	0.112	-2.91-27.69
TG (mg/dL mmol/L)	1.85 ±1.20	1.84 ± 0.95	1.96 ±2.89	0. 001	1.71-1.98
CK(u/l)	361.3 ± 735.5	344.8 ± 698.9	562.2 ± 1087.4	0. 001	278.0-444.6

University of Malaya

Values that are mean ± SD or median derived from SPSS, Confident interval (CI) and p-value. The meaning of the abbreviated variables are explained as follows; HDL=High density lipoprotein, TG= triglycerides, TC= total cholesterol, CK= Creatine kinase, LDL= low density lipoprotein, Egfr= estimated glomerular filtration rate, FBS=Fasting Blood Sugar and CCB= calcium channel blockers, dm =diabetes mellitus, ckd =kidney failure, ASA =Drug Caution Code, ba =broncha asthma, Coad =chronic obs airway disease, ARB =Angiotensin II Receptor Blockers, CCB =Calcium Chanel Blocker, ACE=Angiotensin Converting Enzyme, CCF=congestive cardiac failure, ihd=ischemic heart disease and fx_IHD=Fracture Ischemic Heart Disease.

Dataset of survivor vs. non-survivors was greatly unbalanced. There were no significance differences between survivors vs. non-survivor group ($p > 0.05$). Additional tests were carried out using PCA cluster analysis to confirm. PCA results indicated that there were no significant differences between survivors vs. non-survivor group in the dataset as shown in Figure 4.1.

University of Malaya

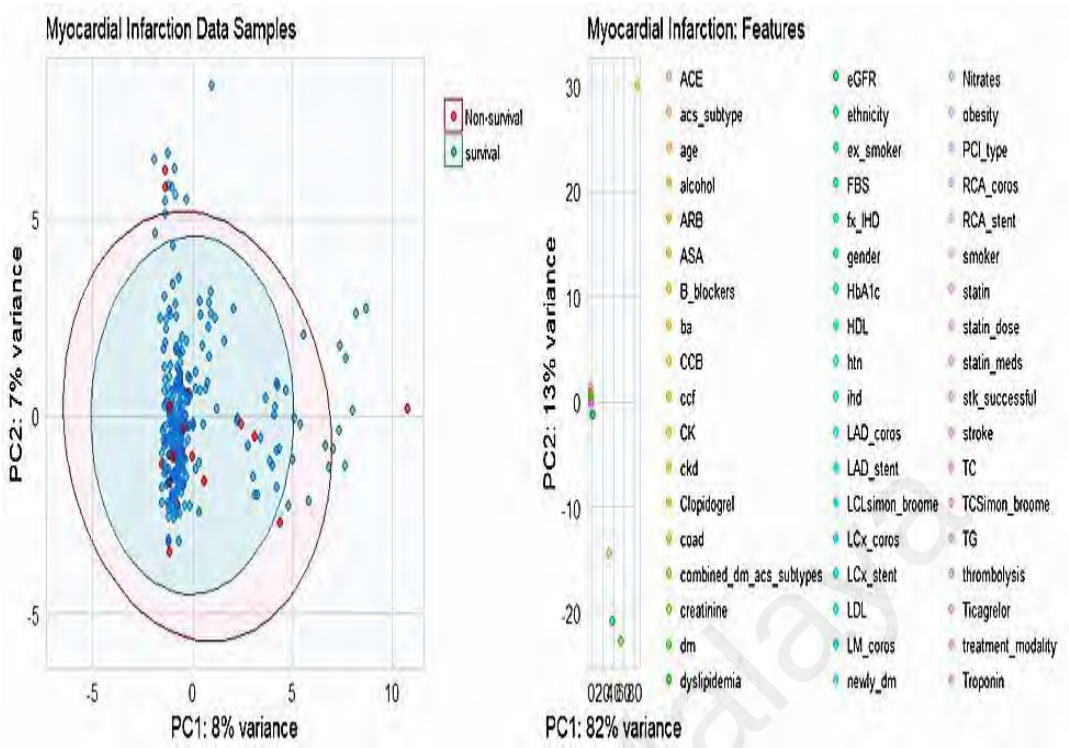

Figure 4.1: PCA cluster analysis results.

4.2 Feature selection

Feature selection was carried out using RF, Boruta, RFE, LVQ, GA, CD, DT, LR, EN and SVM. Classification models were developed using selected features. GA, LVQ, RFE, Boruta and CD were combined with RF and SVM classifier as they do not have inbuilt prediction capabilities. DT, LR, SVM, RF, EN are embedded method which were used both feature selection and classification as they all have inbuilt prediction.

4.2.1 Variable Importance

The variable importance ranked based on their significance for each method are, illustrated in this section. Variable importance was resulting from 100% of training samples. The features are ranked in descending order with the highest ranked in the first position.

University of Malaya

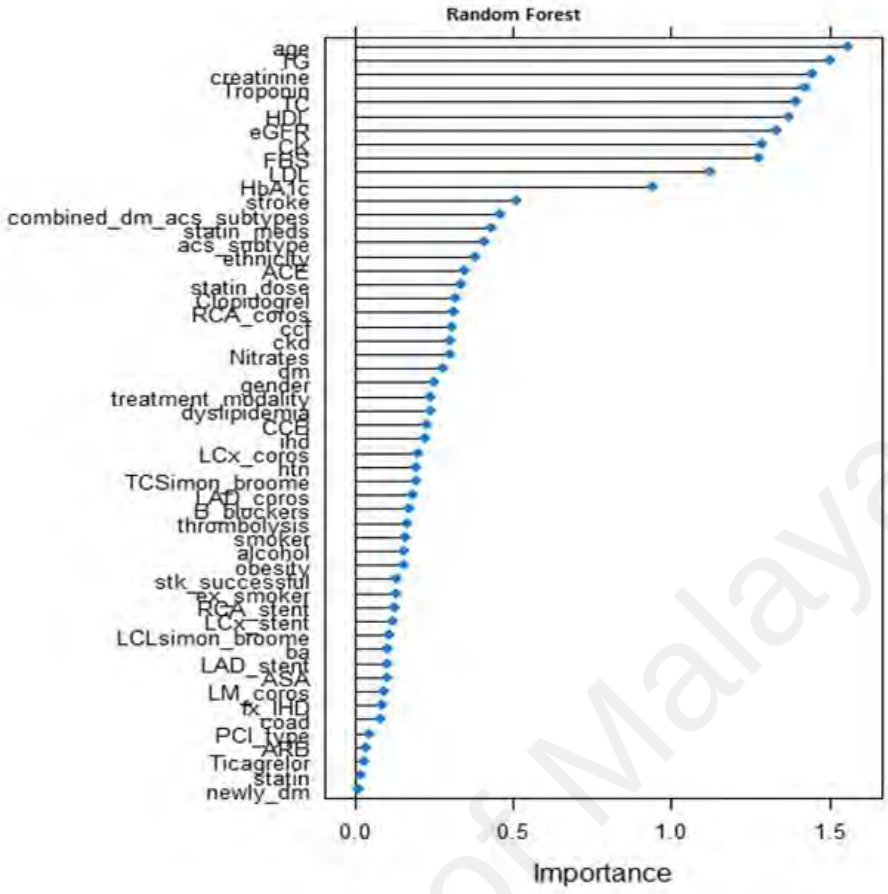

Figure 4.2: Random Forest Variable Ranking.

Illustration in Figure 4.2 shows RF VarImp method. The significant variables identified by RF algorithm are; Age, TG (triglycerides), creatinine, Troponin, TC (total cholesterol), HDL (High Density Lipoprotein), eGFR (estimated Glomerular Filtration Rate), CK (creatinine kinase), FBS (Fasting Blood Sugar), LDL (low density lipoprotein), HbA1c and stroke.

University of Malaya

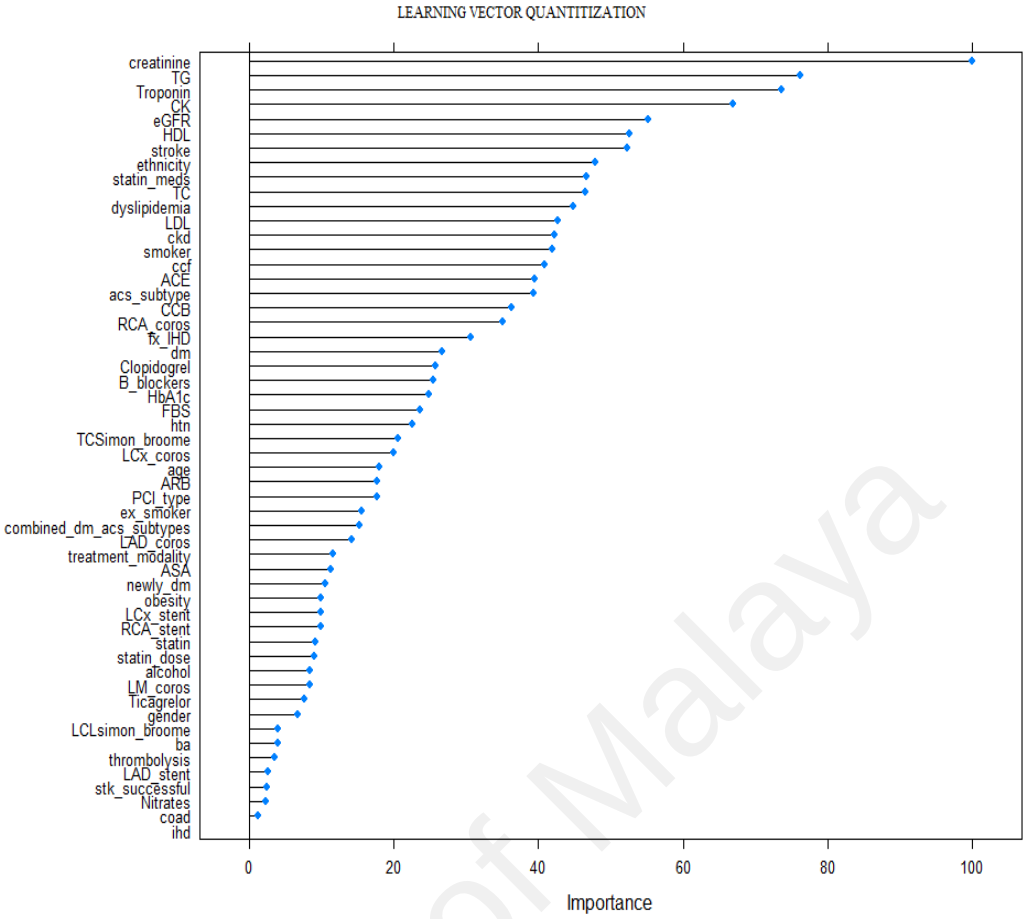

Figure 4.3: Learning Vector Quantization Variable Ranking.

Figure 4.3 illustrates the features selected by LVQ which are ranked in descending order with the highest-ranking variable is creatinine. LVQ was used to select features and then combined with RF and SVM for classification. Top ten significant variables identified by LVQ method are; Creatinine, TG, Troponin, CK, eGFR, HDL, Stroke ethnicity, statin_med and TC.

University of Malaya

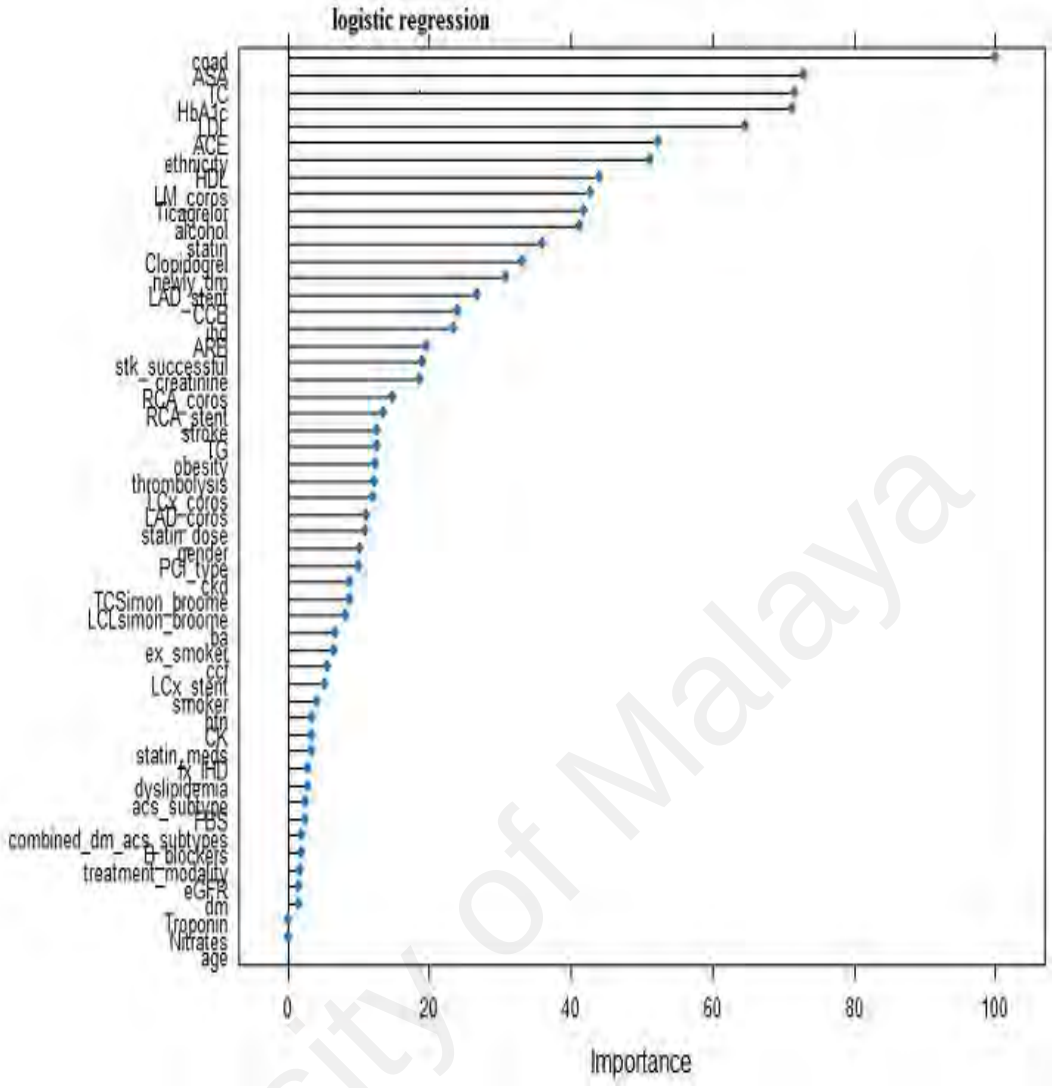

Figure 4.4: Logistic Regression Variable Ranking.

Figure 4.4 illustrates the variables selected by LR algorithm. Variables were ranked in descending order where Coad was ranked as the best variable. The top ten significant variables selected by LR are; Coad, ASA, TC, HBA1c, LDL, ACE, ethnicity, HDL, LM_coros and Ticegrel.

University of Malaya

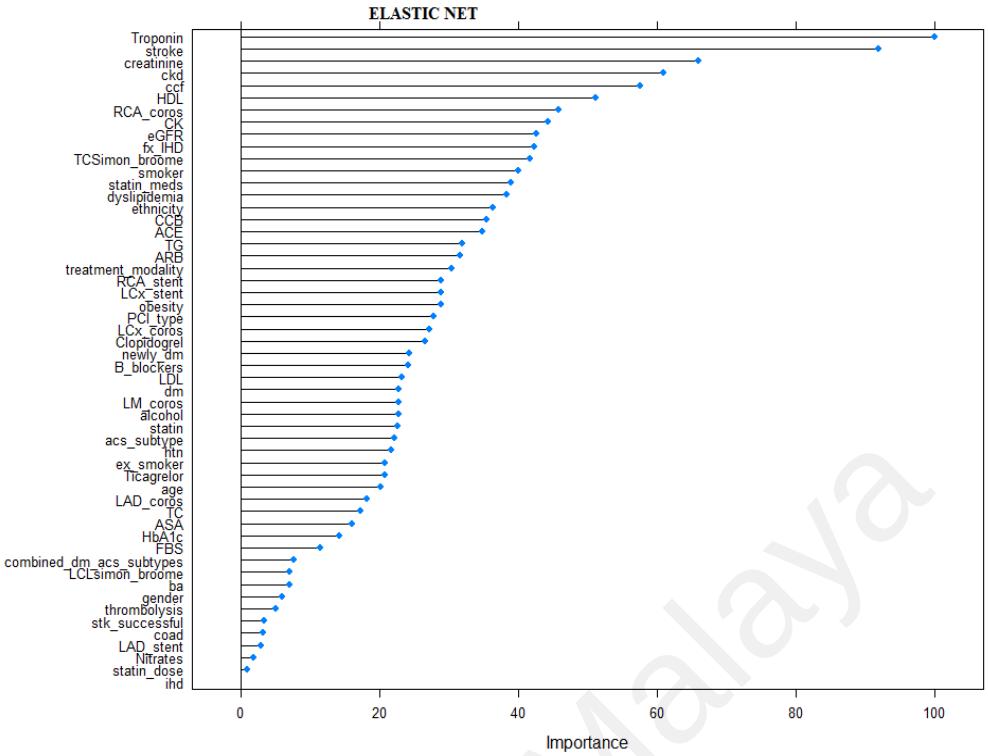

Figure 4.5: Elastic Net Variable Ranking.

Figure 4.5 illustrates the variables selected by EN. Variables were ranked in in descending order where Troponin was ranked as the best variable. The top ten variables selected by EN are; Troponin, Stroke, Creatinine, ckd, CCF, HDL, RCA_coros, CK, eGFR and fx_IHD.

University of Malaya

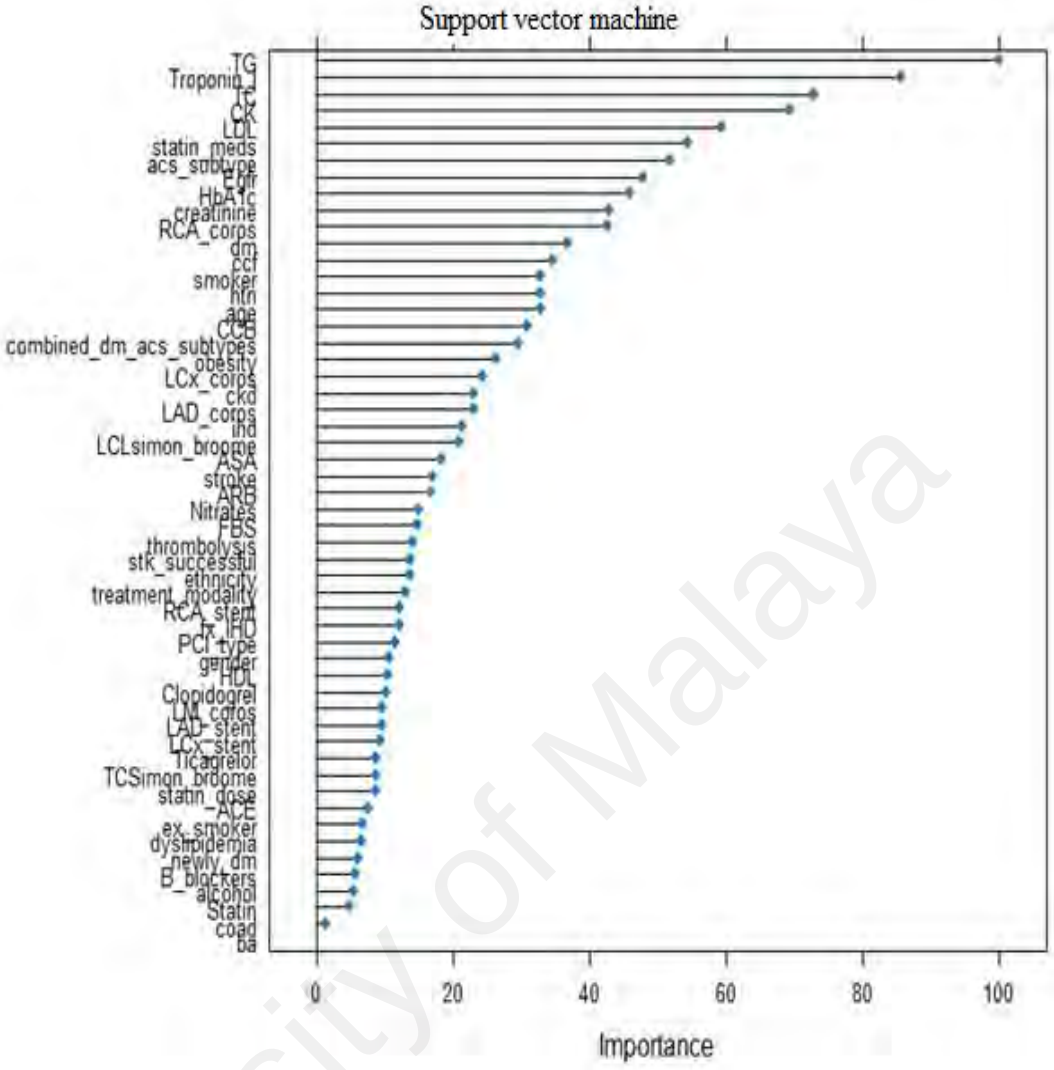

Figure 4.6: Support Vector Machine Variable Ranking.

Figure 4.6 illustrates the variables chosen by SVM. Variables were ranked in a descending order where TG was ranked as the best variable. The top ten significant variables selected by SVM VarImp are; TG, Troponin, TC, CK, LDL, statin_meds, acs_subtype, eGFR, HBA1c and Creatinine.

University of Malaya

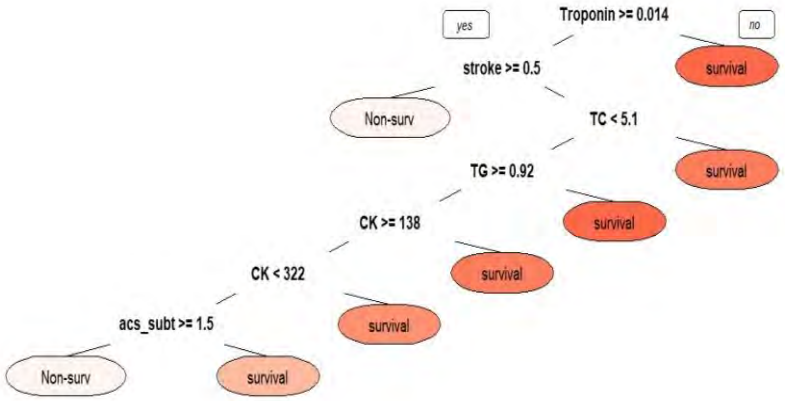

Figure 4.7: Decision Tree Variable Importance.

Figure 4.7 illustrates the variables chosen automatically by DT. Information gain was used in this study as it out performed Gini index due to its poor results. The pruned DT shows the variables on a descending order with the first being the best variable chosen by this model. The best variable selected by IG-DT in this study are Troponin, Stroke, TC, TG, CK and acs_subtype.

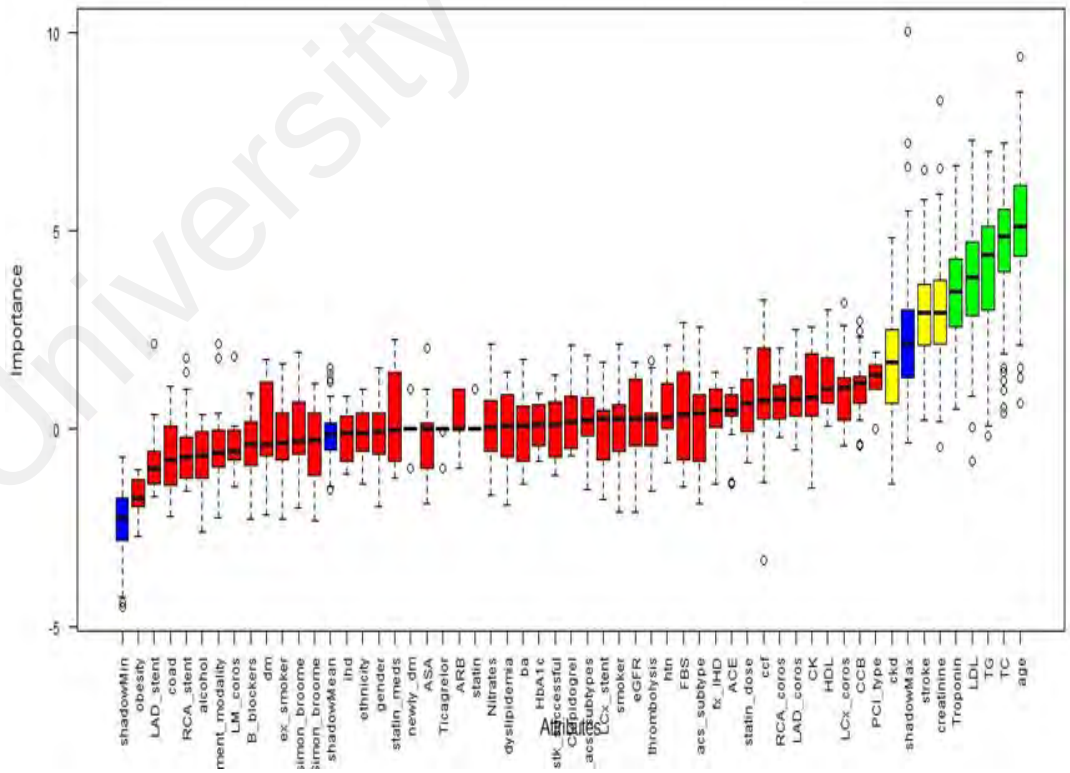

Figure 4.8: Boruta Variable Importance.

University of Malaya

Figure 4.8 illustrates boruta diagram showing variables selected by Boruta feature selection method. Boruta performed 99 iterations where out of 54 variables, 5 variables were selected as important: age, TC, TG, LDL, Troponin and 3 tentative variables left: Creatinine, Stroke and ckd. All the remaining variables in red color were not either designated as important or tentative meaning that boruta was not able to make the decision with the desired confidence. Green color represents important variables; yellow means the variables are tentative. After feature selection, boruta was combined with both RF and SVM for classification using features it selected.

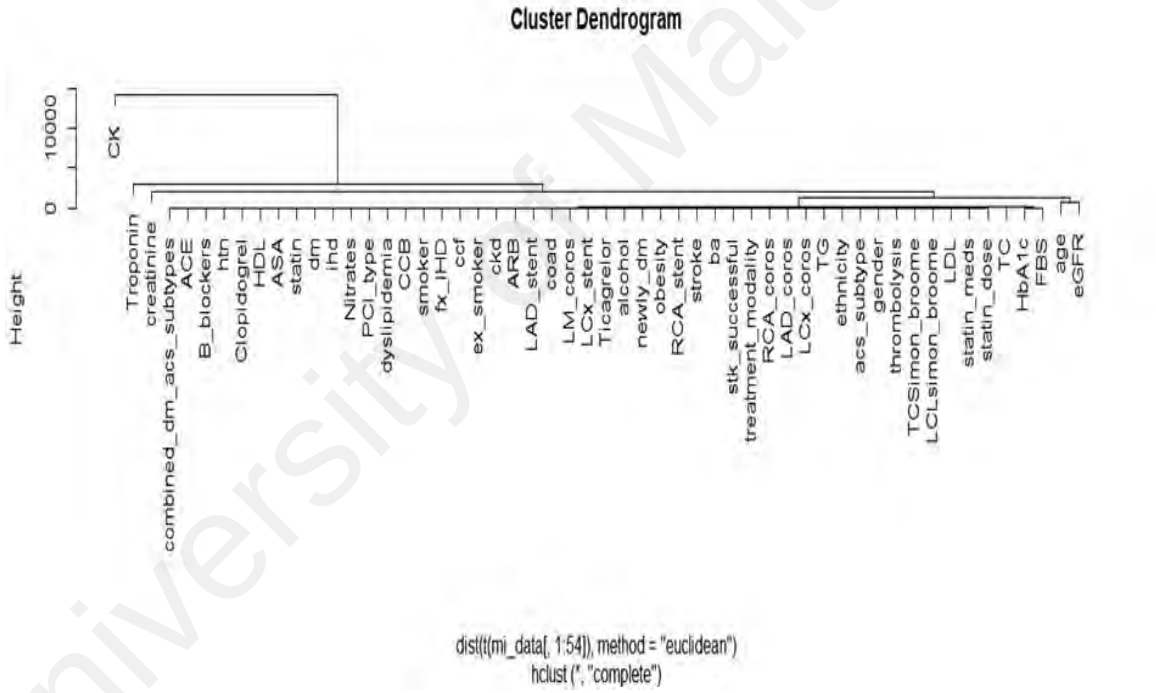

Figure 4.9: Cluster Dendrogram Feature Importance.

Figure 4.9 illustrates variables selected by Cluster Dendrogram (CD). The height axis displays the distance between observations and/or clusters. The horizontal bars indicate the point at which different clusters/observations are merged. Dendrogram were clustered into six clusters, which were ranked according to the percentage relevance of each variable. Variables within clusters were ranked according to similarity to each level. The first level contained only CK followed by troponin which was clustered in the second

University of Malaya

level followed by creatinine which was ranked in the third level and the fourth level contains age and eGFR, followed by HBA1c and FBS on the fifth level, TC on its level and the rest of the variables on one level.

4.3 Machine Learning Results

Feature selection was performed using 10 different feature selection methods on all 54 variables. RF, DT, LR, EN and SVM algorithms comprises of features selection and prediction capabilities. RFE, GA, BORUTA and CD are feature selection algorithms without prediction capabilities and were combined with RF and SVM classifier algorithm for prediction. LVQ is an embedded algorithm, which was used in the current study only for feature selection and was later combined with RF and SVM for classification. RBF is used in this study as it outperformed polynomial and linear kernels with the AUC (RBF = 0.7329, Polynomial = 0.728 and Linear = 0.692) when performed on full dataset. When used on features selected by RFE, GA, BORUTA, LVQ and CD, RBF shows to outperform polynomial and linear kernels. In this study, we have decided to consider only RBF as it outperformed other kernels as shown in Table 4.2, which illustrates the performance of three different SVM kernels when combined with wrapper and filter feature selection methods. As illustrated in Table 4.2, there was no significance difference between performance of linear kernel and RBF kernel when combined with RFE feature selection method. RBF outperformed using both BORUTA and GA with AUC value of (RBF = 0.76, 0.72, Linear = 0.54, 0.52 and Polynomial = 0.69, 0.68). Polynomial performed slightly better than RBF using CD as a feature selection method. The study concluded using SVM for classification and feature selection.

University of Malaya

Table 4.2: Comparing different SVM Kernels.

Models	RFE	BORUTA	GA	LVQ	CD
SVM-Polynomial	0.75	0.69	0.68	0.67	0.71
SVM-Linear	0.78	0.54	0.52	0.53	0.57
SVM-RBF	0.77	0.76	0.72	0.67	0.67

Feature selection to identify significant features that affects mortality was carried out using SBS method for RF, LR, LVQ, EN and SVM based on the ranked variables selected by separate feature selection algorithms in a descending order iteratively. The prediction models were trained and tested for each iteration, and the models with highest performance were selected. Predictive performances of the prediction model were calculated and averaged using untouched testing dataset that was not sampled for model validation. Area under the Curve (AUC) was used as predictive performance metric as it is insensitive to class imbalances. 20% method was used to eliminate variables where four variables were eliminated each time of running based on the best variables selected by ranked feature importance as a method for selecting variables from 54, 20, 16, 12, 8 and 4 deleting the least significant variable and retraining the model. Upon deletion of the variable if the AUC of the model reduces, the variable is deemed as significant. Methods like DT, RFE, BORUTA, CD and GA did not use SBS on the features selected since these methods chooses variables automatically. Table 4.3 illustrates the Sequential backward selection (SBS) where less significant variables were removed, and the model was retrained using significant variables. CD-RF performed slightly better than CD-SVM with AUC (CD-RF= 0.6948, CD- SVM =0.670) as illustrated in table 4.3. RBF performed better on all 54 variables with the AUC = 0.7329. When combined with other feature selection methods, it performed slightly similar to RF.

University of Malaya

Predictive performance of ACS mortality on different prediction models with varying number of variables based on AUC is given in Table 4.3.

Table 4.3: Performance Measure of ML Models Combined with FS and SBS.

Model	4 VARIABLES	8 VARIABLES	12 VARIABLES	16 VARIABLES	20 VARIABLES	54 VARIABLES
RFVarImp-SBS-RF	0.7942	0.7028	0.6998	0.6958	0.7259	0.6165
RFE-RF	0.7821	0.7821	0.7821	0.7821	0.7821	0.7821
BORUTA-RF	0.6767	0.6767	0.6767	0.6767	0.6767	0.6767
CD-RF	0.6486	0.689	0.689	0.689	0.689	0.689
GA-SBS-RF	0.751	0.751	0.751	0.751	0.751	0.6275
LVQ-SBS-RF	0.6948	0.6928	0.6576	0.6265	0.6205	0.6125
SVMVarImp-SBS-SVM	0.6325	0.594	0.566	0.6345	0.638	0.7329
RFE-SVM	0.7650	0.7650	0.7650	0.7650	0.7650	0.7650
BORUTA-SBS-SVM	0.7550	0.7188	0.7188	0.7188	0.7188	0.7188
CD- SBS-SVM	0.676	0.676	0.6485	0.6485	0.6485	0.6485
GA- SBS-SVM	0.6365	0.728	0.592	0.5060	0.5060	0.5060
LVQ- SBS-SVM	0.459	0.670	0.5903	0.5562	0.5843	0.5843
DT	0.6185	0.6185	0.5542	0.5542	0.5542	0.5542
EN- SBS-EN	0.6145	0.6225	0.6004	0.5763	0.492	0.5221
LR- SBS-LR	0.6365	0.5984	0.6245	0.506	0.4739	0.6064

Figure 4.10 illustrates AUC value with varying numbers of variables among different feature selection methods that were used in identifying significant variables.

University of Malaya

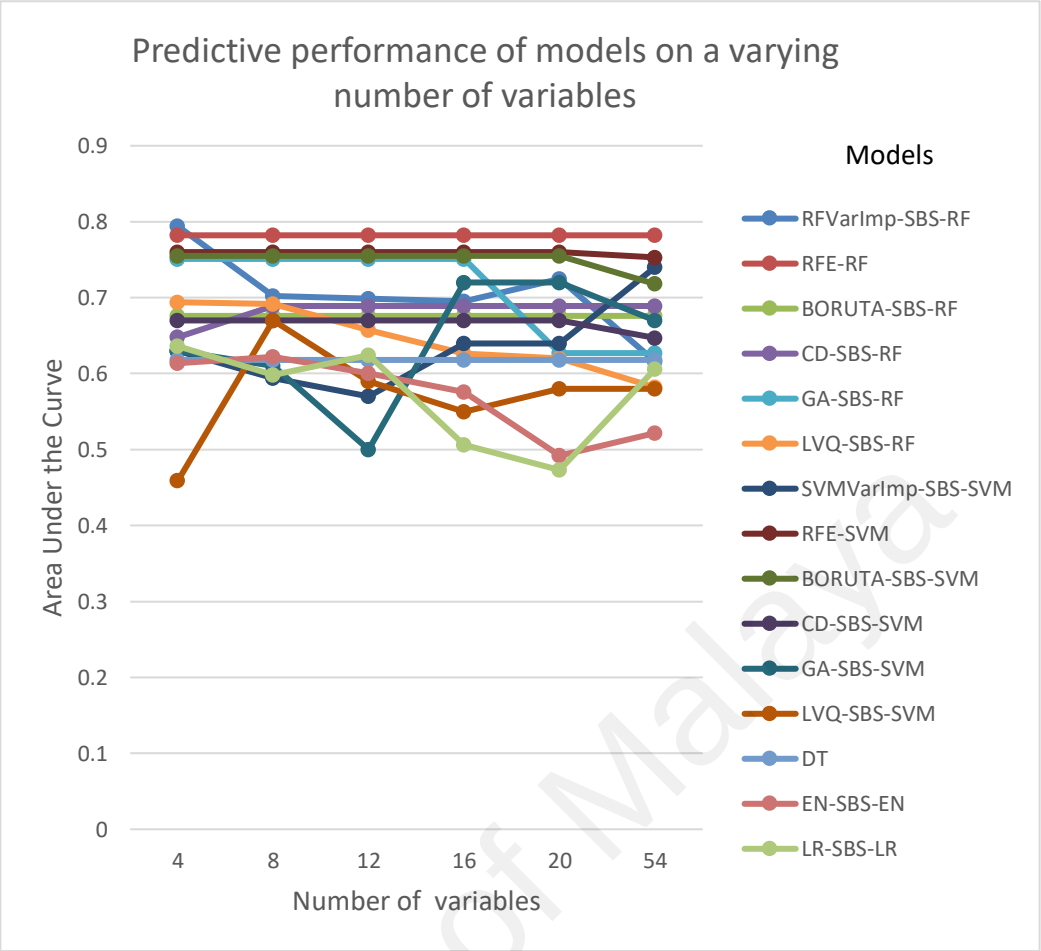

Figure 4.10: Predictive Performance of Classification Model using Varying number of variables.

Where RFVarImp= random forest variable importance, RFE= recursive feature elimination, CD= cluster dendrogram, GA =genetic algorithm, LVQ= Learning Vector Quantization, SVM VarImp = support vector machine variable importance, DT = decision tree, EN= elastic net and LR= logistic regression and SBS =sequential backward selection.

Table 4.4 shows the additional performance metrics of models on the best performing model using optimized parameters identified from feature selection. RF model developed on five predictors reported highest performance on the test set (AUROC = 0.79, Accuracy = 89%). The RF was followed by RFE-RF with (AUROC =0.78) from the variables that were automatically selected by RFE followed by RFE- SVM with AUC = 0.76 then GA-

University of Malaya

RF and BORUTA-SVM which performed similar with the AUC 0.75, SVM with the AUC = 0.73, GA-SVM = 0.728, LR = 0.6245, DT = 0.61and EN=0.61. DT and EN performed slightly similar with the AUC (DT=0.618, EN=0.6145). LR did not perform well in this study as compared to RF and SVM though it outperformed DT and EN on classification.

Theoretically, for a model that indicates survival, for one new patient at the time of first ACS, the best performing ML model RFVarImp-SBS-RF (5 predictors), the average mortality risk is reduced to 4.9% (NPV). If the model outcome is non-survival, the average risk of a patient being deceased is increased to 28.3% (PPV). This calculates to an average 5.9 risk ratios for non-survivors versus survivors by this model. Additional dataset of 102 patients comprising predictors: Age, TC, TG, Troponin, Creatinine, was tested and compared with TIMI score. TIMI score is a standard conventional method for ACS patients to evaluate 30 days mortality in Malaysian hospitals. The additional test performed for comparison purpose for the ML model reported an AUC value of 0.75 vs TIMI score with an AUC value of 0.60.

University of Malaya

Table 4.4: Additional Performance Metrics on Testing dataset for the best Model.

Model	Sense/Spec	PPV/NPV	Detection rate	Detection incidence	AUROC	Accuracy (95% CI)
RF VarImp-SBS-RF	0.331/0.939	0.283/0.951	0.022	0.078	0.794	0.898 (0.816,0.952)
RFE-RF	0.500/0.87	0.23/0.96	0.033	0.146	0.782	0.853(0.763, 0.919)
BORUTA-SBS-RF	0.00/0.951	0.00/0.929	0.00	0.044	0.6767	0.8876(0.803,0.944)
CD- SBS-RF	0.33/0.951	0.33/0.951	0.022	0.067	0.689	0.910 (0.8305,0.9604)
GA- SBS-RF	0.166/0.843	0.0714/0.933	0.011	0.157	0.751	0.797(0.69,0.87)
LVQ-SBS-RF	0.166/0.903	0.11/0.937	0.011	0.101	0.6948	0.53 (0.763,0.919)
SVM VarImp-SBS-SVM	0.00/.987	0.00/0.931	0.00	0.011	0.733	0.921(0.844,0.968)
RFE-SVM	0.166/0.674	0.0357/0.918	0.011	0.314	0.765	0.640(0.532, 0.739)
BORUT-SBS-SVM	0.166/0.638	0.032/0.913	0.011	0.348	0.755	0.606(0.497, 0.708)
CD- SBS-SVM	0.166/0.686	0.037/0.919	0.011	0.303	0.676	0.651(0.543, 0.749)
GA- SBS-SVM	0.00/0.530	0.00/0.880	0.000	0.438	0.728	0.797(0.69,0.87)
LVQ-SBS-SVM	0.166/0.843	0.0714/0.933	0.011	0.157	0.670	0.797(0.699, 0.875)
DT	0.66/0.759	0.166/0.969	0.269	0.044	0.618	0.752(0.65,0.838)
EN-SBS-EN	0.500/0.746	0.125/0.953	0.033	0.269	0.6145	0.730(0.625, 0.82)
LR-SBS-LR	0.166/0.590	0.028/0.907	0.011	0.393	0.6245	0.561(0.452, 0.66)

Results of trained models on 100% of testing data (n =302) by predictor set. For all models, Base Rate Incidence = 0.0674, and No Information Rate = 0.932 Sense= sensitivity, Spec= specificity, PPV= positive predictive value, NPV= negative predictive value, CI =confidence interval, NIR no information rate, RFVarImp= random forest variable importance, RFE= recursive feature elimination, Boruta, LVQ= Learning Vector Quantization, EN =elastic net, CD= cluster dendrogram, GA =genetic algorithm and SVMVarImp = support vector machine variable importance and SBS for sequential backward selection.

The Table 4.5 illustrates optimized variables for the best performing ML models. Age was selected in highest-ranking variable by (3/10) ML models followed by Troponin as the highest-ranking variable by (2/10) ML models and in second highest-ranking variable

University of Malaya

by (3/10) ML models. Different models selected different variables that resulted in optimum performance of the model. FBS and ethnicity were selected by 2/10 models, HBA1c were selected by 4/10 models followed by stroke, HDL and eGFR were all selected by 5/10 models. Creatinine, CK and LDL which were selected by 6/10 models followed by TG and TC which were selected by 7/10 models. The most important variable due to the number of times it was selected by different feature selection methods was troponin as it was chosen by 9/10 models that resulted in model optimum performance. LR obtained the best model on four variables which are; Coad, ASA, TC, HBA1c. Although Coad and ASA have not been chosen by any other feature selection method, TC was selected by 7/10 and HBA1c by 4/10 methods. The best results for EN was obtained using four variables which are; Troponin, stroke, creatinine and ckd.

Table 4.5: Optimized number of Features Selected by different Algorithms via SBS.

Algorithm	Optimum Features Selected
RFvarimp-SBS-RF/ LR	Age, TC, TG, Troponin, Creatinine
RFvarimp-SBS-EN/SVM	Age, TG, creatinine, Troponin, TC, HDL, Egfr, CK
SVMvarimp – SBS – SVM	TG, Troponin, TC, CK, LDL, statin_meds, ACS_SUBTYPE, Egfr, HbA1c, creatinine
SVMvarimp – SBS – RF/EN/LR	TG, Troponin, TC, CK
ENvarimp-SBS – EN/RF/SVM	Troponin, Creatinine, Stroke, ckd
ENvarimp-SBS –LR	Troponin, stroke, creatinine, ckd, ccf, HDL, RCA_coros, CK
LRvarimp-SBS-LR/ EN	TC, HBA1c, Coad, ASA
LRvarimp-SBS-RF/SVM	Coad, ASA, TC, HBA1c, LDL, ACE, ethnicity, HDL, LM_coros, Ticagrelor, alcohol, Statin
RFE-SBS – RF	Age, TC, TG, Troponin, Stroke
RFE-SBS - SVM/EN/LR	Age, TC, TG
BORUTA-SBS-RF	Age, TC, TG, Troponin, LDL
BORUTA-SBS-SVM/EN/LR	Age, TC, TG
CD-SBS- RF/SVM/EN/LR	Age, Troponin, Creatinine, Hba1c, CK, eGFR, FBS
LVQ- SBS- RF/SVM/EN	TG, Troponin, Creatinine, Stroke, CK, eGFR, HDL, Ethnicity
LVQ- SBS- LR	TG, Troponin, Creatinine, Stroke
GA- SBS- RF/SVM	TG, Troponin, FBS, Ethnicity, newly_dm, htn, ex_smoker, TCSimon_broome, LCLsimon_broome, LCx_coros
GA- SBS- EN/LR	TG, Troponin, FBS, Ethnicity

University of Malaya

Figure 4.11 shows Cluster explaining non-survival position on SOM map using features selected by RF best performing model. SOM map was used in this study to investigate relationship between predictors for the best performing model. SOM map performance was evaluated by quantization and topographic error (0.150, 0.056). The coloured scale illustrated in the U-matrix cluster map symbolizes vector distances (predictors are vector elements) (Kohonen, 2001). The blue colour represents the minimal distance (create clusters of vectors with similar features). The red colour represents maximal distance (vectors are dissimilar). The predictors represent the component planes by warm colours that correspond to high mean values and vice-versa. To ease interpretation a white dot was placed on u-matrix cluster and component planes (predictors) on the position that represents non-survival. Non-survival using SOM map is explained as patients with age (> 65), value of troponin I (>11.4 ng/L), creatinine value of (~70 µmol/L), TG (~0.988 mmol/L) and TC (~0.26). There were two notable cluster formed for survivals for patients with; i) Older patients age (>65 year) that survived post ACS, reported lower value of troponin I (~0.5 ng/L) and higher value of creatinine (>138 µmol/L) with similar TG (~0.988 mmol/L) and TC (<5.1 mmol/L). ii) Younger patients aged (< 55 year) with creatinine (<138 µmol/L) regardless of TG, TC and troponin I values were related to survival.

University of Malaya

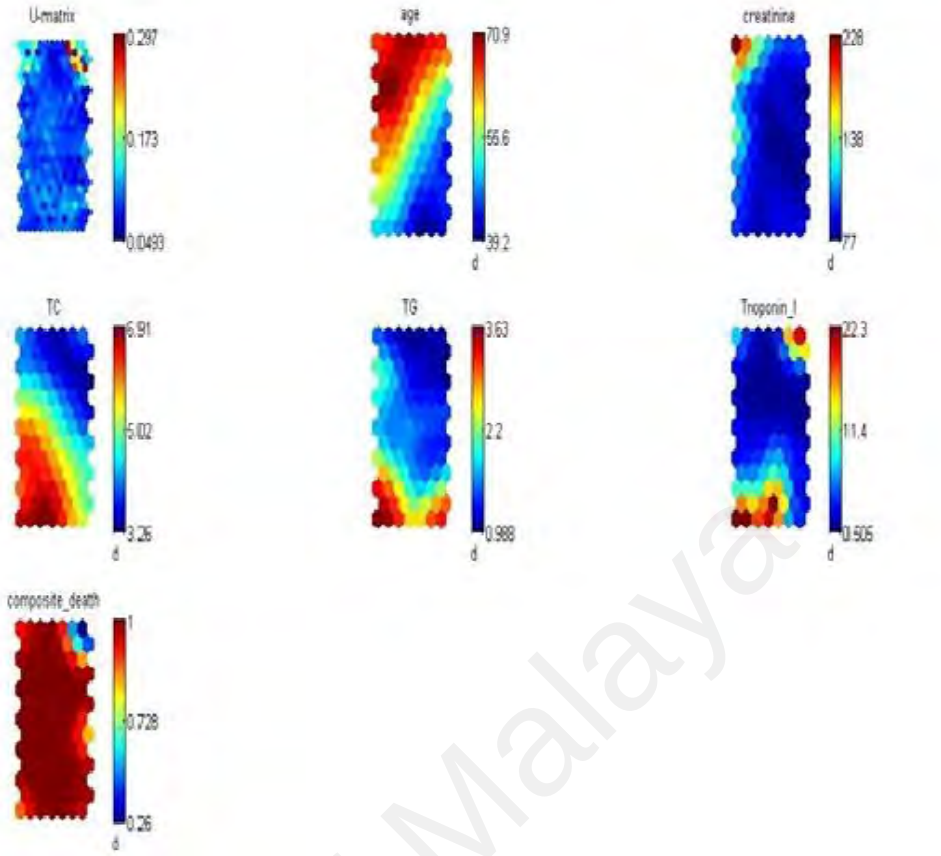

Figure 4.11: SOM map using Features Selected from the best Performing Model RF.

University of Malaya

CHAPTER 5: DISCUSSION

A Combination of feature selection and prediction methods were developed for mortality prediction post ACS in this study. High performances on untouched test set were reported for (7/10) models. Combination of RF VarImp - SBS- RF classifier model and RFE-RF model performance in this study using five predictors was almost similar to RF and SVM model with forward elimination for mortality prediction for a specific type of ACS by Shouval et al., (2017) using five predictors (AUC =0.79) and Wallert et al., (2017) using all predictors (39) (AUC = 0.81 and 0.84). Similar performance to the current study using ML models for mortality prediction for all cause of ACS by Motwani et al., (2016) reported AUC of 0.79 and Steele et al., (2018) reported AUC of 0.79 to 0.81 using RF and EN. Models developed using combination of feature selection and prediction methods reported better results compared to the models developed using all variables in this study. SVM and RF outperform DT, LR and EN models as reported in this study.

Feature selection algorithms are important especially in mortality prediction; however, selecting suitable algorithm or combination of algorithms is difficult and lacking in proper benchmarking. All models improved with the parameter optimisation in this study. Perez et al., (2017) reported combination of feature selection method with classification algorithms reported higher performance where combination of RFE- RF and RFE- SVM outperformed RF without any feature selection. RF and SVM model using full set of predictors without feature selection in this study reported lower AUC compared to RF and SVM with combination of SBS and feature selection algorithm such as LVQ, GA, BORUTA, CD and RFE. GA based feature selection has also been reported to achieve higher performance with combination of ML classifiers in heart failure related study (Mokeddem et al., 2013). LVQ with feature selection has been reported to perform

University of Malaya

better with the combination of ML classifier in heart disease prediction (Sonawane & Patil, 2014). RFE has been reported in previous studies as the one of the best feature selection method used on various clinical dataset and performs better when combined with ML classifier especially SVM and RF (Lin et al., 2017; Yang, 2017; Perez et al., 2017 & Chopra et al., 2017). BORUTA, GA and RFE have been reported to achieve higher performance when combined with ML classifiers like SVM and RF (Chopra et al., 2017). Cluster dendrogram (Euclidian distance-based method) has also been reported to be good feature selection method when combined with ML classifiers (Galili, 2015; Prokashgoswami & Mahanta 2013; Lu & Liang, 2008 & Zhang et al., 2017).

RF and SVM combined with wrapper type of feature selection algorithm resulted in high predictive performance compared to embedded and filter methods LR, EN and DT. Filter method does not combine the final learning algorithm in its stages compared to wrapper and the selected features can at the same time be used in other algorithms for more investigations (Ladha & Deepa, 2011). On the other hand, Saeys et al., (2007), mentioned the drawbacks of filter method that it poorly interacts with classifiers algorithms when used in the long run, and they added in their study that since the utmost filter methods are univariate in nature, these methods may not put much concern on values of other variables. Filter methods drawbacks are redundancy of the selected features and it ignores useful association among features.

RF and SVM are at more advantage compared to LR due to parameter optimization that enables algorithm to adjust to the data details that increases the model predictive performance (Liu et al., 2017; Huang et al., 2016). RF and SVM outperformed LR and EN in other studies even with the standard implementation and default parameters (Fernandez-Delgado 2014; Couronné et al., 2018). DT method did not perform well compared to RF and SVM in this study due inability to handle higher dimensionality of

University of Malaya

the predictors (54 predictors) and complex interactions. In mortality related studies, RF has been reported to achieve higher performance than DT and SVM classifiers (Sakr et al., 2017). In the current study, SVM out performed DT. Xing et al., (2007) where SVM outperformed DT in predicting the outcome of coronary heart disease, supported this.

The cross-validation approach used in this study increases the validity of the models as it reduces the risk of model over fitting. Also, the classification performance is highly impacted by data pre-processing and tuning of algorithms. The k-fold CV technique was used in this study to avoid over fitting. The study used 10-fold CV with three iterations for developing the model using train dataset. This method can be considered as an accurate indicator of performance of a classifier on dataset to avoid over-fitting of the model (Steele et al., 2018; Shouval et al., 2017; Wallert et al., 2017; Motwani et al., 2016).

ML model parameter tuning was carried out in the current study for parameter optimization that affects the model in order to enable the algorithm to perform the best. SVM-RBF was used in this study based on study by Son et al. (2010) and Hanifa et al. (2010). The authors reported that SVM-RBF outperformed SVM-linear and SVM-polynomial. Feature selection and parameter optimization of SVM has been reported in previous study to improve the model predictive performance and stability (Cho et al., 2017; Mohammed et al., 2017; Manurung et al., 2017; Syarif et al., 2016; Cao et al., 2014; Zhou et al., 2011; Dioşan et al., 2010). Feature selection and parameter optimization of SVM has been reported in previous study to improve the model predictive performance and stability (Cho et al., 2017; Mohammed et al., 2017; Manurung et al., 2017; Syarif et al., 2016; Cao et al., 2014; Zhou et al., 2011; Dioşan et al., 2010). In the current study, cost value= 10 and sigma value = 0.5 was chosen using grid search as the best optimal parameter for SVM-RBF model performance (Hric et al., 2011).

University of Malaya

Rankings for the significant variables for RF methods, depends on the mtry value used in splitting a node while developing RF. In this study default value of mtry that is p1/2 for classification was used as suggested by Brieman (2001). The number of trees for development of RF model in this study was set to ntree = 1000. Brieman (2001) reported that larger tree provides stable estimates of variable importance and proximity. Better performance has been reported using ntree= 1000 and default mtry by Probst et al., (2018). Iteration and population size were used in the current study and are noted to improve GA performance compared to other parameters such as crossover probability and mutation rate, that were not used in this study as it was reported be less sensitive and may not affect the model performance (Alajmi et al., 2014).

EN model tuning involves parameters maxit = 1000000, alpha = 0.5, lambda = 100 which improved the performance prediction of the model as stated by Li et al., (2018). This played an important role in improving its performance. Though the performance was not good, EN was able to select features that were selected by the best models. LVQ with SBS was used in this study only for feature selection and then combined with RF and SVM for classification. The parameters used in this study that improved its performance are; Size = 35 and k = 5 that were selected by grid search. LVQ has two parameters which increase its performance incase tunes and these are k and size. K is the number of instances to check when making predictions and size is the number of instances also known as codebooks in the model. In the current study, LVQ was able to select features that were selected by the best models. When combined with RF and SVM, it was able to perform with the AUC ~0.7 hence considered to be a good feature selection method. EN parameters tuning was based on recommendation Li et al., (2018) and LVQ by Grbovic and Vucetic (2009).

University of Malaya

Applications of feature selection algorithms also improve model performance using reasonable number of predictors by reducing predictor's dimensionality (Volmel et al., 2012). Model performance in this study increases with decreasing number of predictors compared to studies reported by (Wallert et al., 2017; Motwani et al., 2016) where the model performance is best using all (39, 54) predictors. This is not cost effective and using a larger number of predictors to arrive at outcome is expensive in terms of data collections and model interpretability for clinical applications. We have used sequential backward selection procedure as weaker predictor is selected when forward selection is used. In forward selection applied by (Wallert et al., 2017) the significance of the predictors is not assessed in the context of other not included predictors (Guyon et al., 2003).

Feature selection technique used, provides a ranked list of variables that were then narrowed down using SBS. Best performing ML model in this study selected five predictors Age, TG, creatinine, Troponin and TC. TG, troponin and TC were among high ranking variables selected across all models. Different feature selection algorithms in this study selected different combination of predictors for mortality predictions post ACS. In all cases, different features were selected by different methods. Model specific predictors that were highly ranked were Age, Hba1c, FBS, CK, eGFR, creatinine, ethnicity and history of stroke. Levels of FBS and HBa1c especially in non-diabetics are related with increased risk in ACS related mortality (Liang et al., 2016). Glucose levels support the relationship between hyperglycemia and increase risk in mortality for patients with STEMI in Asian population (Johansson et al., 2017 and Chen et al., 2017). Risk factors leading to worse outcomes after MI included comorbid diabetes, hypertension, older age, reduced renal function, and history of stroke (Wu et al., 2018).

University of Malaya

Prediction models developed with Boruta features selection reported low positive predictive value to predict non-survival when combined with RF and SVM as it correctly predicted only the survivals ignoring the non-survivals. LDL was among the top features selected by Boruta compared to other ML feature selection method. LDL levels are one of the indicators of decrease in risk of ACS mortality (Navarese et al., 2018).

Variables such as Coad and ASA were only selected by LR. Coad (chronic obstructive airway disease) Patients have a very high risk of MI than non-Coad patients. The common cause of Coad is smoking and shortness of breath in older age (Rothnie, et al., 2016). LR also chose ASA as the second top variable among other variables. ASA or aspirin has been reported by many studies as one of the medications that improve the outcome of patients with ACS hence aspirin users are at higher chances of survival than non-aspirin users (Dai & Ge 2012; Rich et al., 2010; Razzouk et al., 2010).

LCx, TCSimon_broome and LCLSomin_broome were among the top variables selected only by GA. The left circumflex artery (LCX) patients with a large LCX are more likely to have ACS due to a dominant left coronary artery present as STEMI (Waziri et al., 2016; Stribling et al., 2010). Different studies on the simon broome registery has always reported that patients with lower concentration of LCL and TC are at a decrease risk of ACS mortality than patients with higher TC and LCL levels according to simon broome register (Alonso et al., 2018; Latimer et al., 2016; Nanchen et al., 2016; Neil et al., 2008).

It has been well documented that age of the patient is an important criterion in prediction of mortality in different kinds of ACS (Wallert et al., 2017; Shouval et al., 2017; Motwani et al., 2016). The variable age has been identified from the RF VarImp as one of the important variables that affects patients' mortality after ACS. Predictive

University of Malaya

models for LR, DT, EN, SVM and RF was developed using selected variables. Apart from DT other classifiers were combined with SBS.

ML models are considered as black box models. In this study we have shown that clinical data can be visualized in a 2-dimensional representation using SOM technique. SOM was used in this study to understand the relationship between variables from best performing model RF (Age, TG, creatinine, Troponin and TC). This allows the clinician, if there is confidence in the original training data, to place a new patient within the context of previous or similar cases. Results of SOM techniques prove its ability with lowest quantization and topographic errors.

Non-survival from SOM map was related to older patients aged above 65 and higher level of troponin I (>0.01 ng/ml). Meanwhile survival from the same age group was related to lower level of troponin I (<0.01 ng/ml), TG, TC and creatinine. Cardiac troponin I levels were an independent predictor of all-cause mortality. Prognostic value of troponin needs to be considered with the patient's age. Older age (>65) was reported to be linked with a higher mortality for patients with troponin <0.01 ng/mL for troponin I and T (Cheng et al., 2015). Younger patients aged (< 55 year) with low values of creatinine and regardless of TG, TC and troponin I values were related to survival, showing that older age and creatinine levels have a considerable unfavorable consequence on mortality (Marenzi et al., 2015).

Rules provided by the DT algorithm were used to verify relationship between predictors post ACS mortality with results from SOM. DT algorithm however selected different predictors (troponin, TG, TC, CK and ACS subtype) than the best performing model used to construct SOM map. The similar variables selected by DT and RF model were troponin, TC and TG.

University of Malaya

Survival as explained by DT was related with lower values of troponin I and for patients with higher values of troponin without previous history of reported stroke, ACS subtype of NSTEMI, unstable angina and regardless of TC, TG and CK values. Non-survival from DT rules was related with higher level of troponin I and previous reported stroke. Patients with no previous reported stroke, non-survival was explained by ACS subtype of STEMI with higher values of troponin I, TG (>0.92 mmol/L), CK (>138 mmol/L) with lower TC (<5.1 mmol/L) values. This is consistent with (Engberding & Wenger, 2017) where overall mortality of patients was reported to be higher in STEMI than NSTEMI. Overall results were consistent with the SOM analysis. High troponin value and age was related to non-survival. Age is a significant predictor of mortality in ACS, it is also an independent risk factor for adverse outcomes after ACS (Kesavaraj & Sukumaran, 2017). Age was selected as factor that affects mortality post STEMI by ML models in previous studies by Shouval et al. (2017) and Wallert et al. (2017). Older patients usually have more complex cardiovascular disease, more comorbidities, and generally a more atypical clinical presentation (Kesavaraj & Sukumaran, 2017). Elderly patients in this study had a higher burden of cardiovascular risk factors. Compared to non-elderly patients, more elderly patients had a history of diabetes mellitus (38.3% vs. 61.6%), hypertension (38.2% vs. 61.7%), dyslipidaemia (41.2% vs. 58.7%), stroke (16.6% vs. 83.3%), ischemic heart disease (39.8% vs. 60.1%), chronic obstructive airways disease (20% vs. 80%), bronchial asthma (50% vs. 50%) and chronic kidney disease (10.7% vs. 89.2%). These findings conform to findings in the Malaysian population (Kesavaraj & Sukumaran, 2017).

Continuous data collection and digitization of electric health records as a part of clinical practice in Malaysia enables adaptation of ML predictive algorithms tailored to patient's risk grouping. Therefore, ML methods discussed in this study are needed to rank and select major risk factors associated with ACS mortality from clinical viewpoint.

University of Malaya

Conventional methods such as TIMI and GRACE are tools targeted for risk identification post STEMI with a definite timeline associated such as 30-days and six months. ML prediction models are not restricted or bound to timeline or specific group cardiovascular diseases. Furthermore, predictive performance of models constructed in this study was acceptable despite limited number of data samples. (b) Data is collected and archived continuously on everyday clinical repetitive at all hospitals in Malaysia. (c) Prediction models can be linked automatically in future clinical settings for each patient. (d) Enables better communication of individual patients' risk prediction outcome. This would indirectly increase patient awareness of risk and act as instigator to make life style changes and allows clinicians to better plan limited resources available.

The current analysis selected age, creatinine and cardiac markers for the best model RF that are similar to GRACE and TIMI. Variables related to medical history even though was provided during feature selections were not selected, as they were deemed insignificant. Data on vital signs at admission and ECG findings were not available for current analysis, which is considered as limitation and would be incorporated for future studies. Even though the association between the predictor and the outcome is vague to the clinician, SOM analysis suggests its direction by means of visualization. The ability to visualize the relationship of predictors to mortality post ACS offers an important advantage to determine relationship between predictors that may affects mortality which have not been reported in other similar ACS studies.

Results obtained from the combination of feature selection, prediction and visualization can be implemented in clinical practice involving an experimental strategy. Clinicians are allowed to select between prediction model and conventional drill. Assessment on subsequent personalized care and clinical outcomes needs to be carried

University of Malaya

out to determine effectiveness of these proposed methods in improving limited healthcare resources.

Many clinical research datasets have a large percentage of missing values that directly impacts their usefulness in yielding high accuracy classifiers when used for training in supervised machine learning. Limitation of this study was limited number of datasets. Even though lower number of sample size was used in the current study, no data imputation was carried out, as it is known not to improve model performance (Steele et al., 2018). The current study is also subjected to the quality of the data collection and measurement method that was obtained from a single institutional national registry system.

Another limitation was insufficient variables to conduct comparison with conventional methods such as TIMI and GRACE and choice of outcomes. Outcomes such that can be used are cardiovascular disease specific or time-based outcome. However, our main objective was to compare and determine application of feature selection and prediction together with visualization approach in understanding association of risk factors on mortality after ACS. Different combinations of models were developed and comparison with LR showed that combination of ML models performed better.

Evaluating the combination of various features selection methods (filter, wrapper, embedded) with predictive machine learning methods and visualization for this clinically relevant problem was the strength of this study. The application SOM method that allows visualization of association between mortality risk factors that have not yet been reported in any other mortality related studies. The high performance of RF model was achieved by dimensionality reduction of variables using feature selection method that enables model interpretation using SOM method from a clinical point of view. This allows better

University of Malaya

communication and increases awareness of patients that enables behavioral modifications, and better management of limited resources by clinicians.

Comprehensive set of ACS dataset consisting of predictors that enables development of mortality prediction model using numerous unrelated clinical features which are normally recorded by clinicians. The models developed can be beneficial in corresponding to assessment and maintenance of patients' health.

Electronic health records contain large amounts of information about patients' medical history and are becoming more valuable for research. Expert selection of variables, fine-tuning of variable transformations and interactions, in datasets are time-consuming and could lead to biasness in analysis. Furthermore, usage of variables that are not significant for predicting patient outcomes can compromise model performance. A heuristic feature selection method allows identification of significant or important variables. Our study shows that prognostic modelling can be applied, and it is important in clinical practice for patient management and research using machine-learning approaches that allows automatic variable selection techniques that can handle large numbers of predictors. These reduce the amount of human intervention required in fitting prognostic models.

However, for future study models using various data imputation method should be explored when limited and missing data is present. We also aim to evaluate deep learning method with the combination of feature selection and visualization. Future work would also look into model hyper parameters setting using Bayesian optimization instead of random or grid search that has been implemented in the current study. Finally, larger dataset obtained from the Malaysian National Cardiovascular Disease Registry can be used to validate constant models and improve health basing on the new data with possibility of developing models that focuses on specific type of ACS such as STEMI and NSTEMI mortality.

University of Malaya

CHAPTER 6: CONCLUSION

In conclusion, we demonstrated the ability of application of ML algorithms for feature selection, prediction and visualization for mortality predictions in ACS patients.

A combination of applying RF and SOM techniques prove its suitability to be an extremely powerful tool for selecting significant variables, prediction and visualization. It is evident from this work that it is possible to create, a compressed data representation that can be used as tool where the abundance of data obscures straightforward diagnostic reasoning in ACS related mortality study. A conclusion can be drawn that using such a map in conjunction with presentation of mortality related with ACS can be a useful screening mechanism for detecting patients of high ACS risks. As this study is based on small clinical dataset, it can form a useful tool for placing a patient within a clinical setting, permitting and achieving consensus between health practitioners and assess the specific risks of patients when used for system validation.

Additionally, in terms of predictive power, this study can be compared with other published models reporting better results. Factors make these models less generalizable than those proposed in this study. In conclusion, this work built and compared prediction models based on nine heavily used ML algorithms, ranking them by performance (AUC). It found RF method to perform better. When selecting features that determine and predict mortality after ACS. Furthermore, this prediction yielded the finding that age, TG, TC and Troponin were factors that increased the risk of mortality. These insights can be used to design disease-specific interventions to decrease the mortality of high-risk patients. Some of the risk factors found in this study could be used as targets for disease-specific interventions. For example, improved care coordination for patients with multiple conditions and specific follow-up for the most severe patients.

University of Malaya

REFERENCES

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics, 2(*4), 433-459.

Adhikari, G., & Baral, D. (2018). Clinical profile of patients presenting with acute myocardial infarction. *International Journal of Advances in Medicine*, *5*(2), Article #228.

Agarwal, P., Alam, M. A., & Biswas, R. (2010). A hierarchical clustering algorithm for categorical Attributes. In *2010 Second International Conference on Computer Engineering and Applications*, *2*, 365-368. IEEE.

Agresti, A. (2002). Categorical Data Analysis (2nd ed.). Hoboken, NJ: Wiley.

Agresti, A. (2014). Categorical Data Analysis (3rd ed.). Hoboken, NJ: Wiley.

Ahmed, M. H., Awadalla, H., Elmadhoun, W. M., Osman, M., Noor, S. K., & Almobarak, A. O. (2017). Prevalence and risk factors for acute coronary syndrome among sudanese Individuals with Diabetes: A population-based study. *Cardiology Research, 8*(5), 184-189.

Alajmi, A., & Wright, J. (2014). Selecting the most efficient genetic algorithm sets in solving unconstrained building optimization problem. *International Journal of Sustainable Built Environment, 3*(1), 18-26.

Al-Ghamdi, A. S. (2002). Using logistic regression to estimate the influence of accident factors on accident severity. *Accident Analysis & Prevention, 34*(6), 729-741.

Alhassan SM, Ahmed HG, Almutlaq BA, Alanqari AA,……. Alshammari RK. (2017) Risk factors associated with acute coronary syndrome in Northern Saudi Arabia. In search of a perfect outfit. *Journal of Cardiol Coronary Research, 8*(3), Article #00281.

University of Malaya

Allyn, J., Allou, N., Augustin, P., Philip, I., Martinet, O., Belghiti, M., Ferdynus, C. (2017). A comparison of a machine learning model with EuroSCORE II in predicting mortality after elective cardiac surgery: A Decision Curve Analysis. *PLoS ONE, 12*(1), Article # e0169772

Alonso, R., Isla, L. P., Muñiz-Grijalvo, O., Diaz-Diaz, J. L., & Mata, P. (2018). Familial hypercholesterolaemia diagnosis and management. *European Cardiology Review, 13*(1), 14.

Álvarez, J. L., Marí, J. M., Ramon, M. M., & Valls, G. C. (2018). Support vector machine and kernel classification algorithms. Digital signal processing with kernel methods, 433-502.

Amma, N. G. (2012). Cardiovascular disease prediction system using genetic algorithm and neural network. In *2012 International Conference on Computing, Communication and Applications* (pp. 1-5). IEEE.

Antman, E. M., Cohen, M., Bernink, P. J., Mccabe, C. H., Horacek, T., Papuchis, G., . . . Braunwald, E. (2000). The TIMI Risk Score for Unstable Angina/Non–ST Elevation MI. *Jama,* 284(7), 835.

Arauzo-Azofra, A., Benitez, J. M., & Castro, J. L. (2007). Consistency measures for feature selection. *Journal of Intelligent Information Systems, 30*(3), 273-292.

Asadi, H., Dowling, R., Yan, B., & Mitchell, P. (2014). Machine learning for outcome prediction of acute ischemic stroke post intra-arterial therapy. *PLoS ONE, 9*(2), Article # e88225

Aye, M., Cabot, J. S., & Sazali, M. (2015). Study of coronary risk factors in rural Malaysia population. *International Journal of Medical Biology, 2*, 1-7.

Bao, Z., Pi, D., & Sun, Y. (2007). Nonlinear model predictive control based on support vector machine with multi-kernel. *Chinese Journal of Chemical Engineering, 15*(5), 691-697.

University of Malaya

Barlow, S. T., & Neville, P. (2001). Case Study: Visualization for decision tree analysis in data mining. *IEEE Symposium on Information Visualization, 2001. INFOVIS,* 149-152

Bęćkowski, M., Gierlotka, M., Gąsior, M., Poloński, L., Zdrojewski, T., Dąbrowski, R., Szwed, H. (2018). Risk factors predisposing to acute coronary syndromes in young women ≤45 years of age. *International Journal of Cardiology, 264*, 165-169.

Belle, V. V., Pelckmans, K., Suykens, J. A., & Huffel, S. V. (2009). Feature selection in survival least squares support vector machines with maximal variation constraints. *In International Work-Conference on Artificial Neural Networks, 5517,* 65-72. Berlin, Heidelberg: Springer.

Ben-Hur, A., Ong, C. S., Sonnenburg, S., Schölkopf, B., & Rätsch, G. (2008). Support vector machines and kernels for computational Biology. *PLoS Computational Biology, 4*(10), Article # e1000173

Bhatia, S., Prakash, P., & Pillai, G.N. (2008). SVM based decision support system for heart disease classification with integer-coded genetic algorithm to select critical features. In *Proceedings of the world congress on engineering and computer science* (pp. 34-38).

Body, R., Carley, S., Mcdowell, G., Ferguson, J., & Mackway-Jones, K. (2009). Can a modified thrombolysis in myocardial infarction risk score outperform the original for risk stratifying emergency department patients with chest pain? *Emergency Medicine Journal, 26*(2), 95-99.

Boersma, E., Pieper, K. S., Steyerberg, E. W., Wilcox, R. G., Chang, W., Lee, K. L., . . . Simoons, M. L. (2000). Predictors of outcome in patients with acute coronary syndromes without persistent ST-Segment Elevation: *Results From an International Trial of 9461 Patients. Circulation, 101*(22), 2557-2567.

Bowden, R., Mitchell, T., & Sarhadi, M. (1997). Cluster based nonlinear principle component analysis. *Electronics Letters, 33*(22), 1858.

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984) Classification and Regression Trees., (8, pp. 452-456). Wadsworth, NY. Chapman and Hall.

University of Malaya

Breiman, L. (2001). Random Forests. Machine Learning, *45*(1), 5-32.

Cao, Q., Yu, L., & Cheng, M. (2014). A brief overview on parameter optimization of support vector machine. *DEStech Transactions on Materials Science and Engineering*, (smne), 275-279

Castro-Dominguez, Y., Dharmarajan, K., & Mcnamara, R. L. (2018). Predicting death after acute myocardial infarction. *Trends in Cardiovascular Medicine, 28*(2), 102-109.

Chan, M. Y., Du, X., Eccleston, D., Ma, C., Mohanan, P. P., Ogita, M., Jeong, Y. (2016). Acute coronary syndrome in the Asia-Pacific region. *International Journal of Cardiology, 202*, 861-869.

Chang, C., Verhaegen, P. A., & Duflou, J. R. (2014). A Comparison of Classifiers for Intelligent Machine Usage Prediction. In *2014 International conference on intelligent environments* (pp. 198-201). IEEE.

Chandralekha, M., & Shenbagavadivu, N. (2018). Performance Analysis of Various Machine Learning Techniques to Predict Cardiovascular Disease: An Emprical Study. *Applied Mathematics & Information Sciences, 12*(1), 217-226.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering, 40*(1), 16-28.

Chapman, A. R., Adamson, P. D., & Mills, N. L. (2016). Assessment and classification of patients with myocardial injury and infarction in clinical practice. *Heart, 103*(1), 10-18.

Chaudhary, V., Bhatia, R., & Ahlawat, A. K. (2014). A novel self-organizing map (SOM) learning algorithm with nearest and farthest neurons. *Alexandria Engineering Journal, 53*(4), 827-831.

University of Malaya

Chawla, N. V. (2005). Data mining for imbalanced datasets: An overview. In data mining and knowledge *discovery handbook*, (8, pp.875-886). Boston, MA: Springer.

Cheng, J. M., Helming, A. M., Vark, L. C., Kardys, I., Uil, C. A., Jewbali, L. S., . . . Akkerhuis, K. M. (2015). A simple risk chart for initial risk assessment of 30-day mortality in patients with cardiogenic shock from ST-elevation myocardial infarction. *European Heart Journal: Acute Cardiovascular Care, 5*(2), 101-107.

Chen, C., Yen, D. H., Lin, C., Tsai, S., Chen, S., Sheu, W. H., & Hsu, C. (2017). Glycated hemoglobin level is an independent predictor of major adverse cardiac events after nonfatal acute myocardial infarction in nondiabetic patients. *Medicine, 96*(18), Article #e6743.

Chen, X., & Ishwaran, H. (2012). Random forests for genomic data analysis. *Genomics,* 99(6), 323-329.

Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin (2003). A practical guide to support vector classification, Technical report, *Department of Computer Science, National Taiwan University,* 1-16.

Chicco, D. (2017). Ten quick tips for machine learning in computational biology. *BioData Mining, 10*(1), 35.

Cho, M., & Hoang, T. T. (2017). Feature selection and parameters optimization of SVM using particle swarm optimization for fault classification in power distribution systems. *Computational Intelligence and Neuroscience, 2017*, 1-9.

Chopra, A., Dimri, A., & Pradhan, T. (2017). Prediction of factors affecting amlodipine induced pedal edema and its classification. In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 1684-1689). IEEE.

Christenson, E., & Christenson, R. H. (2013). The role of cardiac biomarkers in the diagnosis and management of patients presenting with suspected acute coronary syndrome. *Annals of Laboratory Medicine, 33*(5), 309.

University of Malaya

Clarke, B., Fokou´e, E. & Zhang, H. (2009). Principles and theory for data mining and machine learning. *Springer Series in Statistics*, Springer, New York.

Couronné, R., Probst, P., & Boulesteix, A. (2018). Random forest versus logistic regression: A large-scale benchmark experiment. *BMC Bioinformatics, 19*(1), 270

Cox, D. R. (1958). Two further applications of a model for binary regression. *Biometrika, 45*(3-4), 562-565.

Collinson, P., & Lindahl, B. (2015). Type 2 myocardial infarction: The chimaera of cardiology? *Heart,* 101(21), 1697-1703.

Copeland, D. C. (2017). Quantitative analysis and qualitative case study research. Princeton University Press.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273-297.

Collazo, R. A., Pessôa, L. A., Bahiense, L., Pereira, B. D., Reis, A. F., & Silva, N. S. (2016). A comparative study between artificial neural network and support vector machine for acute coronary syndrome prognosis. *Pesquisa Operacional, 36*(2), 321-343.

Dagostino, R. B., Vasan, R. S., Pencina, M. J., Wolf, P. A., Cobain, M., Massaro, J. M., & Kannel, W. B. (2008). General cardiovascular risk profile for use in primary care: T*he Framingham Heart Study. Circulation, 117*(6), 743-753.

Dai, Y., & Ge, J. (2012). Clinical use of aspirin in treatment and prevention of cardiovascular disease. *Thrombosis, 2012*, 1-7.

Dash, M., & Liu, H. (2003). Consistency-based search in feature selection. *Artificial Intelligence*, *151*(1-2), 155-176.

University of Malaya

Datla, M. V. (2015). Bench marking of classification algorithms: decision trees and random forests - a case study using R. In *2015 international conference on trends in automation, communications and computing technology (I-TACT-15)* (pp. 1-7). IEEE.

DeSalvo, K. B., Bloser, N., Reynolds, K., He, J., & Muntner, P. (2006). Mortality prediction with a single general self-rated health question. *Journal of General Internal Medicine, 21*(3), 267–275.

Dietrich, S., Floegel, A., Weikert, C., Prehn, C., Adamski, J., Pischon, T., . . . Drogan, D. (2016). Identification of serum metabolites associated with incident hypertension in the european prospective investigation into cancer and nutrition–potsdam study. *Hypertension, 68*(2), 471-477.

Dioşan, L., Rogozan, A., & Pecuchet, J. (2010). Improving classification performance of support vector machine by genetically optimising kernel shape and hyper-parameters. *Applied Intelligence, 36*(2), 280-294.

Doak, J. (1992). An evaluation of feature selection methods and their application to computer security (Technical Report CSE-92–18). Davis, CA: University of California, Department of Computer Science.

Dohare, A. K., Kumar, V., & Kumar, R. (2018). Detection of myocardial infarction in 12 lead ECG using support vector machine. *Applied Soft Computing, 64*, 138-147.

Dong, G., & Wang, X. (2009). Application of decision tree construction algorithm based on decision classify-entropy. *Journal of Computer Applications, 29*(11), 3103-3106.

Dunkler, D., Plischke, M., Leffondré, K., & Heinze, G. (2014). Augmented backward elimination: A pragmatic and purposeful way to develop statistical models. *PLoS ONE, 9*(11), e113677.

Du, M., Wang, S. M., & Gong, G. (2011). Research on decision tree algorithm based on information entropy. *Advanced Materials Research, 267*, 732-737.

University of Malaya

Engberding, N., & Wenger, N. K. (2017). Acute coronary syndromes in the elderly. *F1000Research, 6*, 1791.

Faber, J., & Fonseca, L. M. (2014). How sample size influences research outcomes. *Dental Press Journal of Orthodontics, 19*(4), 27-29.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters,27*(8), 861-874.

Fesmire, F. M., Martin, E. J., Cao, Y., & Heath, G. W. (2012). Improving risk stratification in patients with chest pain: the erlanger hearts$_3$ score. *American Journal of Emergency Medicine*, *30*(9), 1829-1837.

Feder, S. L., Schulman-Green, D., Geda, M., Williams, K., Dodson, J. A., Nanna, M. G., Chaudhry, S. I. (2015). Physicians' perceptions of the Thrombolysis in Myocardial Infarction (TIMI) risk score in older adults with acute myocardial infarction. *Heart & Lung: The Journal of Acute and Critical Care, 44*(5), 376-381.

Fernandez-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014).Do we need hundreds of classifiers to solve real world classification problems?. *Journal of Machine Learning Research, 15(*1), 3133–3181.

Fox, K. A., Dabbous, O. H., Goldberg, R. J., Pieper, K. S., Eagle, K. A., Werf, F. V., . . . Granger, C. B. (2006). Prediction of risk of death and myocardial infarction in the six months after presentation with acute coronary syndrome: Prospective multinational observational study (GRACE). *Biomedical Journal, 333*(7578), 1091.

Friedman, J., Hastie, T. & Tibshirani, (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33*, 1-22.

Fusaro, V. A., Mani, D. R., Mesirov, J. P., & Carr, S. A. (2009). Prediction of high-responding peptides for targeted protein assays by mass spectrometry. *Nature Biotechnology, 27*(2), 190-198.

University of Malaya

Galili, T. (2015). Dendextend: An R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics, 31*(22), 3718-3720.

Gareth James (2013). Introduction to Statistical Learning. New York, Springer.

Gaspar, P., Carbonell, J., & Oliveira, J. L. (2012). On the parameter optimization of Support Vector Machines for binary classification. *Journal of Integrative Bioinformatics, 9*(3), 33-43

Genuer, R., Poggi, J., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters, 31*(14), 2225-2236.

George, G.V.S.; Raj, V.C (2011). Review of Feature Selection Techniques and the Impact of SVM for cancer classification using gene expression profile. *International Journal of Computer Science & Engineering Survey (IJCSES), 2(3), 16-27.*

Goldstein, B. A., Navar, A. M., & Carter, R. E. (2016). Moving beyond regression techniques in cardiovascular risk prediction: Applying machine learning to address analytic challenges. *European Heart Journal*, *38*(23), 1805-1814

Gray, H. H., & Henderson, R. A. (2011). The GRACE scores performance in predicting in-hospital and 1-year outcome. *Heart, 97*(18), 1461-1462.

Green, M., Björk, J., Forberg, J., Ekelund, U., Edenbrandt, L., & Ohlsson, M. (2006). Comparison between neural networks and multiple logistic regression to predict acute coronary syndrome in the emergency room. *Artificial Intelligence in Medicine, 38*(3), 305-318.

Grech ED, Ramsdale DR (2003a). Acute coronary syndrome: unstable angina and non-ST segment elevation myocardial infarction. *Biomedical Journal, 326*, 1259-61.

Grech ED, Ramsdale DR. (2003b). Acute coronary syndrome: ST segment elevation myocardial infarction. *Biomedical Journal, 326*, 1379-81.

University of Malaya

Grbovic, M., & Vucetic, S. (2009). Learning vector quantization with adaptive prototype addition and removal. *2009 International Joint Conference on Neural Networks, (*pp. 994-1001). IEEE.

Gregorutti B, Michel B, Saint-Pierre P (2013). Correlation and variable importance in random forests. *Statistical Computational, 27*, 659–78.

Guo, P., Luo, Y., Mai, G., Zhang, M., Wang, G., Zhao, M., . . . Zhou, F. (2014). Gene expression profile based classification models of psoriasis. *Genomics, 103*(1), 48-55.

Guyon, I., Elisseeff, A., & Kaelbling, L. P. (Ed.). (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*(7-8), 1157-1182.

Habermann, J. K., Doering, J., Hautaniemi, S., Roblick, U. J., Bündgen, N. K., Nicorici, D., . . . Ried, T. (2009). The gene expression signature of genomic instability in breast cancer is an independent predictor of clinical outcome. *International Journal of Cancer, 124*(7), 1552-1564.

Hachesu, P. R., Ahmadi, M., Alizadeh, S., & Sadoughi, F. (2013). Use of data mining techniques to determine and predict length of stay of cardiac patients. *Healthcare Informatics Research, 19*(2), 121.

Hajar, R. (2017). Risk factors for coronary artery disease: Historical perspectives. *Heart Views, 18*(3), 109.

Hamilton, B., Kwakyi, E., Koyfman, A., & Foran, M. (2013). Diagnosis and management of acute coronary syndrome. *African Journal of Emergency Medicine, 3*(3), 124-133.

Han, T., Goodenough, D., Dyk, A., & Chen, H. (2004). Hyperspectral feature selection for forest classification. In *IGARSS 2004. 2004 IEEE International Geoscience and Remote Sensing Symposium* , 2, 1471-1474. IEEE.

University of Malaya

Haneef, M., Joseph, A., Noone, M., & Babu, A. (2010). Risk factors among patients with acute coronary syndrome in rural Kerala. *Indian Journal of Community Medicine, 35*(2), 364.

Hanifa, S. M., & Raja, K. S. (2010). Stroke risk prediction through non-linear support vector classification models. *International Journal of Advanced Research in Computer Science, 1*(3), 47-53.

Helwan, A., Uzun, D., Abiyev, R., & Bush, J. (2017). One-Year survival prediction of myocardial infraction. *International Journal of Advanced Computer Science and Applications, 8*(6), 173-178.

Hess, E. P., Agarwal, D., Chandra, S., Murad, M. H., Erwin, P. J., Hollander, J. E., . . . Stiell, I. G. (2010). Diagnostic accuracy of the TIMI risk score in patients with chest pain in the emergency department: A meta-analysis. *Canadian Medical Association Journal, 182*(10), 1039-1044.

Hodzic, E., Perla, S., Iglica, A., & Vucijak, M. (2018). Seasonal incidence of acute coronary syndrome and its features. *Materia Socio Medica, 30*(1), 10.

Hoogendoorn, M., Hassouni, A. E., Mok, K., Ghassemi, M., & Szolovits, P. (2016). Prediction using patient comparison vs. modeling: A case study for mortality prediction. *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, (pp. 2464-2467). IEEE

Holland, J. H. (1992). Genetic algorithms. *Scientific American, 267*(1), 66-72.

Hoo, F. K., Boo, Y. L., Foo, Y. L., Lim, S. M., & Ching, S. M. (2016). Acute coronary syndrome in young adults from a Malaysian tertiary care centre. *Pakistan Journal of Medical Sciences, 32*(4), 841.

Hozawa, A., Folsom, A. R., Sharrett, A. R., & Chambless, L. E. (2007). Absolute and attributable risks of cardiovascular disease incidence in relation to optimal and borderline risk factors. *Archives of Internal Medicine, 167*(6), 573.

University of Malaya

Hric, M., Chmulik, M., & Jarina, R. (2011). Model parameters selection for SVM classification using particle swarm optimization. In *Proceedings of 21st International Conference Radioelektronika 2011* (pp. 1-4). IEEE.

Huang, B. F., & Boutros, P. C. (2016). The parameter sensitivity of random forests. *BMC Bioinformatics, 17*(1), 331.

Huang, J., & Ling, C. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering, 17* (3), 299-310.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics, 2*(3), 841-860.

Jafarian, A., Ngom, A., & Rueda, L. (2011). A novel recursive feature subset selection algorithm. *2011 IEEE 11th International Conference on Bioinformatics and Bioengineering,* (pp. 78-83). IEEE.

Jenhani, I., Amor, N. B., & Elouedi, Z. (2008). Decision trees as possibilistic classifiers. *International Journal of Approximate Reasoning, 48*(3), 784-807.

Jiang, Z., & Shekhar, S. (2017). Spatial Information Gain-Based Spatial Decision Tree. Spatial Big Data Science, (10, pp. 57-76). Cham, Switzerland: Springer.

Jollife, I. T. (2010). Principal component analysis. (2nd edn.).New York, NY: Springer-Verlag.

Jovic, A., Brkic, K., & Bogunovic, N. (2015). A review of feature selection methods with applications. 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), (pp. 1200-1205). IEEE.

University of Malaya

Johansson, S., Rosengren, A., Young, K., & Jennings, E. (2017). Mortality and morbidity trends after the first year in survivors of acute myocardial infarction: A systematic review. *BMC Cardiovascular Disorders, 17*(1), 53.

Kabir, M. M., Islam, M. M., & Murase, K. (2010). A new wrapper feature selection approach using neural network. *Neurocomputing, 73*(16-18), 3273-3283.

Karatzoglou, A., Meyer, D., & Hornik, K. (2006). Support vector machines in R. *Journal of Statistical Software, 15*(9), 1-28.

Karetnikova, V., Gruzdeva, O., Uchasova, E., Osokina, A., & Barbarash, O. (2016). Glucose levels as a prognostic marker in patients with ST-segment elevation myocardial infarction: A case–control study. *BMC Endocrine Disorders, 16*(1), 31.

Kayani, W. T., & Ballantyne, C. M. (2018). Improving outcomes after myocardial infarction in the US population. *Journal of the American Heart Association, 7*(4), 1-3.

Kenny, G. P., Sigal, R. J., & Mcginn, R. (2016). Body temperature regulation in diabetes. *Temperature, 3*(1), 119-145.

Kesavaraj, G., & Sukumaran, S. (2013). A study on classification techniques in data mining. 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), (pp. 1-7). IEEE

Khalilia, M., Chakraborty, S., & Popescu, M. (2011). Predicting disease risks from highly imbalanced data using random forest. BMC *Medical Informatics and Decision Making, 11*(1), 51.

Kira, K. & Rendell, L. (1992). A practical approach to feature selection. In *Machine Learning Proceedings 1992* (pp. 249-256). Morgan Kaufmann.

University of Malaya

Kishk, Y. T., Helmy, H. A., Abdelmegid, M. A., & Abdel-Nour, M. M. (2017). Incidence of impaired glucose tolerance assessed by glycated Hemoglobin and fasting plasma Glucose in patients with acute coronary syndromes and its impact on clinical and angiographic outcomes. *The International Annals of Medicine, 1*(10), 205-215.

Kleinbaum, D. G., Kupper, L. L., Nizam, A., & Rosenberg, E. S. (2014). Applied regression analysis and other multivariable methods. (Vol. 601). Belmont, CA: Duxbury press.

Kleinbaum, D.G. & Kupper, L.L. & Nizam, A & Muller, K.E. (2008). Logistic regression analysis. Applied Regression Analysis and Other Multivariate Methods (4th ed.). Belmont: Thomson Higher Education.

Kohonen,T.(1986). Learning vector quantization for pattern recognition. Technical Report TKK-F-A601, Helsinki University of Technology, Espoo, Finland. 2C. (n.d.).

Kolman, L., Hu, Y., Montgomery, D. G., Gordon, K., Eagle, K. A., & Jackson, E. A. (2009). Prognostic value of admission fasting glucose Levels in patients with acute coronary syndrome. *The American Journal of Cardiology, 104*(4), 470-474.

Kononenko, I., & Kukar, M. (2007). Data Preprocessing. Machine Learning and Data Mining, 181-211.

Kohonen, T. (1995). Learning vector quantization. The Handbook of Brain Theory and Neural Networks, (Vol.95, pp. 537–540). Cambridge, MA: *MIT Press*.

Kohonen, T. (2001). Self-Organizing Maps. Springer Series in Information Sciences. (Vol.30). Self-organizing. Berlin: Springer.

Kong, X., Hu, C., & Duan, Z. (2017). Generalized Principal Component Analysis. *Principal Component Analysis Networks and Algorithms,* 185-233.

University of Malaya

Korjus, K., Hebart, M. N., & Vicente, R. (2016). An efficient data partitioning to improve classification performance while keeping parameters interpretable. *PLoS ONE,11*(8), Article # e0161788.

Kursa MB, Rudnicki WR (2010). Feature selection with the Boruta package. *Journal of Statistical Software, 36*, 1–13.

Kumbhar, P., & Mali, M. (2016). A survey on feature selection techniques and classification algorithms for efficient text classification. *International Journal of Science and Research, 5*(5), 1267-1275.

Kuhn, M. (2008). Building Predictive Models in RUsing the caret Package. *Journal of Statistical Software, 28*(5), 1-26.

Kumar, A., & Cannon, C. P. (2009). Acute coronary syndromes: Diagnosis and management, part I. *Mayo Clinic Proceedings, 84*(10), 917-938.

Kuhn, M., & Johnson, K. (2013). Applied predictive modelling. New York: Springer.

Kumar, A. (2018). Pre-processing and modelling using caret package in R. *International Journal of Computer Applications, 181*(6), 39-42.

Kuo, P., Wu, S., Chien, P., Rau, C., Chen, Y., Hsieh, H., & Hsieh, C. (2018). Derivation and validation of different machine-learning models in mortality prediction of trauma in motorcycle riders: A cross-sectional retrospective study in southern Taiwan. *BMJ Open, 8*(1), Article #e018252

Ladha, L & Deepa, T. (2011). Feature selection methods and algorithms. *International Journal on Computer Science and Engineering, (3)*1787-1797.

Lagerqvist, B. (2005). FRISC score for selection of patients for an early invasive treatment strategy in unstable coronary artery disease. *Heart, 91*(8), 1047-1052.

University of Malaya

Latimer, J., Batty, J. A., Neely, R. D., & Kunadian, V. (2016). PCSK9 inhibitors in the prevention of cardiovascular disease. *Journal of Thrombosis and Thrombolysis, 42*(3), 405-419.

Lee, K., Lee, M., & Kim, D. (2017). Utilizing random forest QSAR models with optimized parameters for target identification and its application to target-fishing server. *BMC Bioinformatics, 18*(16), 567.

Lee, G. K., Lee, L. C., Liu, C. W., Lim, S. L., Shi, L. M., Ong, H. Y., . . . Yeo, T. C. (2010). Framingham risk score inadequately predicts cardiac risk in young patients presenting with a first myocardial infarction. Ann Acad Med Singapore, 39, 163-7.

Lee, C. H., & Yoon, H. (2017). Medical big data: Promise and challenges. *Kidney Research and Clinical Practice, 36*(1), 3-11.

Little, M. A., Varoquaux, G., Saeb, S., Lonini, L., Jayaraman, A., Mohr, D. C., & Kording, K. P. (2017). Using and understanding cross-validation strategies. Perspectives on Saeb *GigaScience,6*(5), gix020

Libby P (2001). Current concepts of the pathogenesis of the acute coronary syndromes, *Circulation, 104*, 365-72.

Li, X., & Claramunt, C. (2006). A Spatial Entropy-Based Decision Tree for Classification of Geographical Information. *Transactions in GIS, 10*(3), 451-467.

Li, X., Xie, S., Zeng, D., & Wang, Y. (2018). Efficient ℓ 10-norm feature selection based on augmented and penalized minimization. *Statistics in Medicine,37*(3), 473-486.

Liang, H., Guo, Y. C., Chen, L. M., Li, M., Han, W. Z., Zhang, X., & Jiang, S. L. (2016). Relationship between fasting glucose levels and in-hospital mortality in Chinese patients with acute myocardial infarction and diabetes mellitus: A retrospective cohort study. *BMC Cardiovascular Disorders, 16*(1), 156

University of Malaya

Liu, C. H., Chamberlain, B. P., Little, D. A., & Cardoso, Â. (2017). Generalising random forest parameter optimisation to include stability and cost. *Machine Learning and Knowledge Discovery in Databases Lecture Notes in Computer Science,* 102-113.

Lin, X., Li, C., Zhang, Y., Su, B., Fan, M., & Wei, H. (2017). Selecting feature subsets based on SVM-RFE and the overlapping ratio with applications in bioinformatics. *Molecules, 23*(1), 52.

Liu, H., & Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering, 17*(4), 491-502.

L. I. e. K. Han. (1999). Model selection and model averaging for neural networks. Department of Statistics, Carnegie Mellon University Pittsburgh.

Lloyd-Jones, D., Adams, R. J., Brown, T. M., Carnethon, M., Dai, S., Simone, G. D., . . . Wylie-Rosett, J. (2010). Executive summary: Heart disease and stroke statistics--2010 Update: *A Report from the American Heart Association. Circulation, 121*(7), 948-954.

Loprinzi, P. D., Cardinal, B. J., Winters-Stone, K., Smit, E., & Loprinzi, C. L. (2012). Physical Activity and the risk of breast cancer recurrence: A literature review. *Oncology Nursing Forum, 39*(3), 269-274.

Lu, Y., & Liang, L.R. (2008). Hierarchical clustering of features on categorical data of Biomedical Applications.In *CAINE*. (pp.26-31).

Lu, H. T., & Nordin, R. B. (2013). Ethnic differences in the occurrence of acute coronary syndrome: Results of the Malaysian national cardiovascular disease (NCVD) database registry (March 2006 - February 2010). *BMC Cardiovascular Disorders, 13*(1), 97.

Ma, C.-P., Wang, X., Wang, Q.-S., Liu, X.-L., He, X.-N., & Nie, S.-P. (2016). A modified HEART risk score in chest pain patients with suspected non-ST-segment elevation acute coronary syndrome. *Journal of Geriatric Cardiology, 13*(1), 64–69.

University of Malaya

Manurung, J., Mawengkang, H., & Zamzami, E. (2017). Optimizing support vector machine parameters with genetic algorithm for credit risk assessment. *Journal of Physics: Conference Series, 930*, 012026.

M. Blachnik (2009). Comparison of various feature selection methods in application to prototype best rules. *Computer Recognition Systems, 3*(57), 257-264.

Mansoor, H., Elgendy, I. Y., Segal, R., Bavry, A. A., & Bian, J. (2017). Risk prediction model for in-hospital mortality in women with ST-elevation myocardial infarction: A machine learning approach. *Heart & Lung: The Journal of Acute and Critical Care, 46*(6), 405-411.

Mastoi, Q., Wah, T. Y., Raj, R. G., & Iqbal, U. (2018). Automated diagnosis of coronary artery disease: A review and workflow. *Cardiology Research and Practice, 2018*, 1-9.

Malek, S., Gunalan, R., Kedija, S., Lau, C., Mosleh, M. A., Milow, P., . . . Saw, A. (2018). Random forest and self organizing maps application for analysis of pediatric fracture healing time of the lower limb. *Neurocomputing, 272*, 55-62.

Masetic, Z., & Subasi, A. (2016). Congestive heart failure detection using random forest classifier. *Computer Methods and Programs in Biomedicine, 130*, 54-64.

Marenzi, G., Cabiati, A., Cosentino, N., Assanelli, E., Milazzo, V., Rubino, M., . . . Bartorelli, A. (2015). Prognostic significance of serum creatinine and its change patterns in patients with acute coronary syndromes. *American Heart Journal, 169*(3), 363-370.

Menardi, G. and Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data mining and Knowledge Discovery, 28*, 92–122.

Menard, S. W. (2002). Applied logistic regression analysis. Thousand Oaks, CA: Sage Publications.

University of Malaya

Mendis, S., Thygesen, K., Kuulasmaa, K., Giampaoli, S., Mahonen, M., Blackett, K. N., & Lisheng, L. (2010). World Health Organization definition of myocardial infarction: 2008-09 revision. *International Journal of Epidemiology, 40*(1), 139-146.

Mendoza, M. R., Fonseca, G. C., Loss-Morais, G., Alves, R., Margis, R., & Bazzan, A. L. (2013). RFMirTarget: Predicting human MicroRNA target genes with a random forest classifier. *PLoS ONE, 8*(7), Article #e70153.

Meier, C. K., & Oyama, M. A. (2009). Myocardial infarction. *Small Animal Critical Care Medicine*, 174-176.

Mendis, S., Puska, P. and Norrving, B., Eds. (2011) Global Atlas on cardiovascular disease prevention and control. World Health Organization, Geneva.

Mirza, A. J., Taha, A. Y., & Khdhir, B. R. (2018). Risk factors for acute coronary syndrome in patients below the age of 40 years. *The Egyptian Heart Journal*, *70*(4), 233-235.

Miron B. Kursa, Witold R. Rudnicki (2010). Feature selection with the boruta package. *Journal of Statistical Software, 36*(11), 1-13.

Mokeddem, S., Atmani, B., & Mokaddem, M. (2014). A new approach for coronary artery diseases diagnosis based on genetic algorithm. *International Journal of Decision Support System Technology, 6(4)*, 1-15.

Mokeddem, S., Atmani, B., & Mokaddem, M. (2013). Supervised feature selection for diagnosis of coronary artery disease based on genetic algorithm. *Computer Science & Information Technology, 13*(5), 6046.

Mohamadlou, H., Panchavati, S., Calvert, J., Lynn-Palevsky, A., Barton, C., Fletcher, G., Das, R. (2018). Multicenter validation of a machine learning algorithm for 48 hour all-cause mortality prediction. *Biological Review* (bioRxiv), 427054.

University of Malaya

Motwani, M., Dey, D., Berman, D. S., Germano, G., Achenbach, S., Al-Mallah, M. H., . . . Slomka, P. J. (2016). Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: A 5-year multicentre prospective registry analysis. *European Heart Journal*, *38*(7), 500-507.

Mohammed, L. B., & Raahemifar, K. (2017). Improving support vector machine classification Accuracy based on Kernel parameters optimization. In *Proceedings of the Communications and Networking Symposium* (p. 10). Society for Computer Simulation International.

Myers, J., Souza, C. R., Borghi-Silva, A., Guazzi, M., Chase, P., Bensimhon, D., . . . Arena, R. (2014). A neural network approach to predicting outcomes in heart failure using cardiopulmonary exercise testing. *International Journal of Cardiology, 171*(2), 265-269.

Nauta, S. T., Deckers, J. W., Boon, R. M., Akkerhuis, K. M., & Domburg, R. T. (2012). Risk factors for coronary heart disease and survival after myocardial infarction. *European Journal of Preventive Cardiology, 21*(5), 576-583.

Navarese, E. P., Robinson, J. G., Kowalewski, M., Kolodziejczak, M., Andreotti, F., Bliden, K., Gurbel, P. A. (2018). Association between baseline LDL-C Level and total and cardiovascular mortality after LDL-C Lowering. *Journal of American Medical Analysis (Jama), 319*(15), 1566.

Nanchen, D., Gencer, B., Muller, O., Auer, R., Aghlmandi, S., Heg, D., . . . Rodondi, N. (2016). Prognosis of patients with familial hypercholesterolemia after acute coronary syndromes. *Circulation, 134(*10), 698-709.

Neil, A., Cooper, J., Betteridge, J., Capps, N., Mcdowell, I., Durrington, P., . . . Humphries, S. E. (2008). Reductions in all-cause, cancer, and coronary mortality in statin-treated patients with heterozygous familial hypercholesterolaemia: A prospective registry study. *European Heart Journal, 29*(21), 2625-2633.

Nikam, S. , Shukla P., & Shah M., (2017) Cardiovascular disease prediction using genetic algorithm and neuro-fuzzy system. *International Journal of Latest Trends in Engineering and Technology, 8*(2), 104-110.

University of Malaya

Nova, D., & Estévez, P. A. (2015). A review of learning vector quantization classifiers. *Neural Computing and Applications, 25*(3-4), 511-524.

Özçift, A. (2011). Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis. *Computers in Biology and Medicine, 41*(5), 265-271.

Palaniappan, S., & Awang, R. (2008). Intelligent heart disease prediction system using data mining techniques. 2008 IEEE/ACS International Conference on Computer Systems and Applications (pp. 108-115). IEEE.

Park, H. (2013). An introduction to logistic regression: From basic concepts to interpretation with particular attention to nursing domain. *Journal of Korean Academy of Nursing, 43*(2), 154.

Pal NR, Pal SK (1993). A review on image segmentation techniques. *Pattern Recognition 26*(9), 1277-1294.

Paluszek, M., & Thomas, S. (2016). An Overview of Machine Learning. *MATLAB Machine Learning,* (pp. 113-141). Berkeley, CA: Apress.

Pedreira, C. (2006). Learning vector quantization with training data selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28*(1), 157-162.

Perez-Riverol, Y., Kuhn, M., Vizcaíno, J. A., Hitz, M., & Audain, E. (2017). Accurate and fast feature selection workflow for high-dimensional omics data. *PLoS ONE, 12*(12), Article # e0189875.

Pham, H. N., & Triantaphyllou, E. (2008). The impact of overfitting and overgeneralization on the classification accuracy in data mining. *Soft Computing for Knowledge Discovery and Data Mining*, (pp. 391-431). Boston, MA: Springer.

University of Malaya

Pollack, C. V., & Braunwald, E. (2008). Guidelines for the management of patients with unstable Angina and Non–ST-Segment Elevation Myocardial Infarction: Implications for Emergency Department Practice. *Annals of Emergency Medicine, 51*(5), 591-606.

Powers, D. M. (2011). Evaluation: From precision, recall and f-measure to roc, informedness, and markedness & correlation. *Journal of Machine Learning Technologies, 2*(1), 37-63.

Probst, P., Wright, M., & Boulesteix, A. (2018). Hyperparameters and Tuning Strategies for Random Forest. *WIREs Data Mining Knowledge Discovery*, *2*(1), 1-19.

Prokashgoswami, J., & Mahanta, A. K. (2013). Categorical data clustering based on an alternative data representation technique. *International Journal of Computer Applications, 72*(5), 7-12.

Quinlan, J. R. (1993). C4. 5: Programs for machine learning, 1. Morgan Kaufmann.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning,1*(1), 81-106.

Ramadhan, M. M., Sitanggang, I. S., Nasution, F. R., & Ghifari, A. (2017). Parameter tuning in random forest based on grid search method for gender classification based on voice frequency. *DEStech Transactions on Computer Science and Engineering,*(Cece). C625-629).

Razzouk, L., Mathew, V., Lennon, R. J., Aneja, A., Mozes, J. I., Wiste, H. J., . . . Farkouh, M. E. (2010). Aspirin use is associated with an improved long-term survival in an unselected population presenting with unstable angina. *Clinical Cardiology, 33*(9), 553-558.

Rich, J. D., Cannon, C. P., Murphy, S. A., Qin, J., Giugliano, R. P., & Braunwald, E. (2010). Prior aspirin use and outcomes in acute coronary syndromes. *Journal of the American College of Cardiology, 56*(17), 1376-1385.

University of Malaya

Ricci, B., Cenko, E., Vasiljevic, Z., Stankovic, G., Kedev, S., Kalpak, O., Bugiardini, R. (2017). Acute coronary syndrome: The risk to young women. *Journal of the American Heart Association, 6*(12), Article #e007519.

Ridker, P. M., Paynter, N. P., Rifai, N., Gaziano, J. M., & Cook, N. R. (2008). C - reactive protein and parental history improve global cardiovascular risk prediction: The Reynolds Risk Score for Men. *Circulation, 118*(22), 2243-2251.

Ross, E. G., Shah, N. H., Dalman, R. L., Nead, K. T., Cooke, J. P., & Leeper, N. J. (2016). The use of machine learning for the identification of peripheral artery disease and future mortality risk. *Journal of Vascular Surgery, 64*(5), 1515-1522.

Rothnie, K. J., Smeeth, L., Pearce, N., Herrett, E., Timmis, A., Hemingway, H., . . . Quint, J. K. (2016). Predicting mortality after acute coronary syndromes in people with chronic obstructive pulmonary disease. *Heart, 102*(18), 1442-1448.

Rodondi, N., Locatelli, I., Aujesky, D., Butler, J., Vittinghoff, E., Simonsick, E., & Bauer, D. C. (2012). Framingham risk score and alternatives for prediction of coronary Heart Disease in older adults. *PLoS ONE, 7*(3), Article # e34287.

Saeys, Y., Inza, I., & Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics, 23*(19), 2507-2517.

Salappa, A., Doumpos, M., & Zopounidis, C. (2007). Feature selection algorithms in classification problems: An experimental evaluation. *Optimization Methods and Software, 22*(1), 199-212.

Salari, N., Shohaimi, S., Najafi, F., Nallappan, M., & Karishnarajah, I. (2013). Application of pattern recognition tools for classifying acute coronary syndrome: An integrated medical modeling. *Theoretical Biology and Medical Modelling, 10*(1), 57.

Sakr, S., Elshawi, R., Ahmed, A., Qureshi, W. T., Brawner, C., Keteyian, S., . . . Al-Mallah, M. H. (2018). Using machine learning on cardiorespiratory fitness data for predicting hypertension: The Henry Ford ExercIse Testing (FIT) Project. *PLoS ONE, 13*(4), Article #e0195344.

University of Malaya

Sakr, S., Elshawi, R., Ahmed, A. M., Qureshi, W. T., Brawner, C. A., Keteyian, S. J., . . . Al-Mallah, M. H. (2017). Comparison of machine learning techniques to predict all-cause mortality using fitness data: The Henry ford exercIse testing (FIT) project. *BMC Medical Informatics and Decision Making, 17*(1), 174.

Saulnier, D. M., Riehle, K., Mistretta, T., Diaz, M., Mandal, D., Raza, S., Versalovic, J. (2011). Gastrointestinal microbiome signatures of pediatric patients with irritable bowel syndrome. *Gastroenterology, 141*(5), 1782-1791.

Sanguansat, P. (2012). Two-Dimensional principal component analysis and its extensions. Principal Component Analysis. *International Conference on Pattern Recognition, Hong Kong, China,* 2, 1246–1249.

Scholkopf, b. (2018). Learning with kernels: Support vector machines, regularization, optimization, and beyond. S.l.: *MIT PRESS.*

Schmid, M., Wright, M. N., & Ziegler, A. (2016). On the use of Harrell's C for clinical risk prediction via random survival forests. *Expert Systems with Applications, 63*, 450-459.

Sen, S. K. (2017). Predicting and diagnosing of heart disease using machine learning Algorithms. *International Journal of Engineering and Computer Science*, *6*(6), 56.

Shouval, R., Hadanny, A., Shlomo, N., Iakobishvili, Z., Unger, R., Zahger, D., . . . Beigel, R. (2017). Machine learning for prediction of 30-day mortality after ST elevation myocardial infraction: An Acute Coronary Syndrome Israeli Survey data mining study. *International Journal of Cardiology, 246*, 7-13.

Six, A. J., Backus, B. E., & Kelder, J. C. (2008). Chest pain in the emergency room: Value of the HEART score. *Netherlands Heart Journal, 16*(6), 191-196.

Son, Y., Kim, H., Kim, E., Choi, S., & Lee, S. (2010). Application of support vector machine for prediction of medication adherence in heart failure patients. *Healthcare Informatics Research, 16*(4), 253.

University of Malaya

Sonawane, J. S., & Patil, D. R. (2014). Prediction of heart disease using learning vector quantization algorithm. 2014 conference on IT in business, industry and government (CSIBIG) (pp. 1-5). IEEE.

Steele, A. J., Denaxas, S. C., Shah, A. D., Hemingway, H., & Luscombe, N. M. (2018). Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. *PloS ONE, 13*(8), Article # e0202344.

Stebbins, A., Mehta, R. H., Armstrong, P. W., Lee, K. L., Hamm, C., Werf, F. V., . . . Granger, C. B. (2010). A Model for Predicting Mortality in Acute ST-Segment Elevation Myocardial Infarction Treated With Primary Percutaneous Coronary Intervention: Results From the Assessment of Pexelizumab in Acute Myocardial Infarction Trial. *Circulation: Cardiovascular Interventions, 3*(5), 414-422.

Stribling, W. K., Kontos, M. C., Abbate, A., Cooke, R., Vetrovec, G. W., & Lotun, K. (2010). Clinical outcomes in patients with acute left circumflex/obtuse marginal occlusion presenting with myocardial infarction. *Journal of Interventional Cardiology, 24*(1), 27-33.

Sundaram, V., & T, S. (2012). Classification rules by decision tree for disease prediction. *International Journal of Computer Applications, 43*(8), 6-12.

Suykens, J. A. (2001). Support vector machines: A nonlinear modelling and control perspective. *European Journal of Control, 7*(2-3), 311-327.

Syarif, I., Prugel-Bennett, A., & Wills, G. (2016). SVM parameter optimization using grid search and genetic algorithm to improve classification performance. *TELKOMNIKA (Telecommunication Computing Electronics and Control), 14*(4), 1502.

Tapas, N., Lone, D., Reddy, D., & Kuppili, V. (2017). Prediction of cardiac arrest recurrence using ensemble classifiers. *Indian Academy of Sciences, 42*(7), 1135–1141.

University of Malaya

Thakare, V. S., & Patil, N. N. (2014). Image texture classification and retrieval using self-organizing map. *2014 International Conference on Information systems and Computer Networks (ISCON)* (pp. 25-29). IEEE.

Therneau, T., Atkinson, B., & Ripley, B. (2015). RPART: Recursive partitioning for classification, regression and survival trees. *R Package Version, 4,* 1-9.

Thirugnanam, M., Mukherji, D., Padalia, N., & Naidu, A. (2013). A heart disease prediction model using SVM-Decision trees-logistic regression (SDL). *International Journal of Computer Applications, 68*(16), 11-15.

Thygesen, K., Alpert, J. S., Jaffe, A. S., Simoons, M. L., Chaitman, B. R., & White, H. D. (2012). The writing group on behalf of the joint ESC/ACCF/AHA/WHF task force for the universal definition of myocardial infarction. *Third Universal Definition of Myocardial Infarction Circulation, 16*, 2020-2035.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. *Series B (Methodological)*, 267-288.

Tuckova, J. (2013). The possibility of kohonen self-organizing map applications in medicine. *2013 IEEE 11th International Workshop of Electronics, Control, Measurement, Signals and Their Application to Mechatronics*. (pp. 1-6). IEEE.

Tušar, T., Gantar, K., Koblar, V., Ženko, B., & Filipič, B. (2017). A study of overfitting in optimization of a manufacturing quality control procedure. *Applied Soft Computing, 59*, 77-87.

Uriarte, E. A., & Martín, F. D. (2005). Topology preservation in SOM. *International Journal for Applied Mathematics and Computer Sciences, 1*(1), 19-22.

VanHouten, J. P., Starmer, J. M., Lorenzi, N. M., Maron, D. J., & Lasko, T. A. (2014). Machine learning for risk prediction of acute coronary syndrome. AMIA *Annual Symposium Proceedings, 2014,* 1940–1949.

University of Malaya

Vedanthan, R., Seligman, B., & Fuster, V. (2014). Global perspective on acute coronary syndrome. *Circulation Research, 114*(12), 1959-1975.

Verikas, A., Gelzinis, A., & Bacauskiene, M. (2011). Mining data with random forests: A survey and results of new tests. *Pattern Recognition, 44*(2), 330-349.

Volmel, .K.H., Tuma, P., Precek, J., Hutyra, M. (2012).Machine learning methods for mortality prediction in patients with ST elevation myocardial infarction. Proceedings of WUPES, 204–213.

Wallert, J., Tomasoni, M., Madison, G., & Held, C. (2017). Predicting two-year survival versus non-survival after first myocardial infarction using machine learning and Swedish national register data. *BMC Medical Informatics and Decision Making, 17*(1), 99.

Wang, Y., Li, J., Zheng, X., Jiang, Z., Hu, S., Wadhera, R. K. & Jiang, L. (2018). Risk factors associated with major cardiovascular events 1 Year after acute myocardial infarction. *JAMA Network Open, 1*(4), Article # e181079.

Waziri, H., Jørgensen, E., Kelbæk, H., Fosbøl, E. L., Pedersen, F., Mogensen, U. M., . . .Wachtell, K. (2016). Acute myocardial infarction and lesion location in the left circumflex artery: Importance of coronary artery dominance. *EuroIntervention, 12*(4), 441-448.

Webb, G. I. (2017). Overfitting. Encyclopedia of Machine Learning and Data Mining, 947-948.

Wickham, H., & Sievert, C. (2016). Ggplot2: Elegant graphics for data analysis. Houston, TX: Springer.

Wiens, J., & Shenoy, E. S. (2017). Machine learning for healthcare: On the verge of a major shift in healthcare epidemiology. *Clinical Infectious Diseases, 66*(1), 149-153.

University of Malaya

Witten, I. H., Frank, E., & Hall, M. A. (2011). Ensemble learning. *Data Mining: Practical Machine Learning Tools and Techniques. (The Morgan Kaufmann Series in Data Management Systems)*, 351-373.

Wu, C., Singh, A., Collins, B., Fatima, A., Qamar, A., Gupta, A., . . . Blankstein, R. (2018). Causes of troponin elevation and associated mortality in young patients. *The American Journal of Medicine, 131*(3), 284-292.

Xing, Y., Wang, J., Zhao, Z., & Gao, A. (2007). Combination Data Mining Methods with New Medical Data to Predicting Outcome of Coronary Heart Disease. In *2007 International Conference on Convergence Information Technology (ICCIT 2007)* (pp. 868-872). IEEE.

Yang, X. (2017). Identification of risk genes associated with myocardial infarction based on the recursive feature elimination algorithm and support vector machine classifier. *Molecular Medicine Reports*, *17*(1), 1555-1560.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68*(1), 49-67.

Zaccardi, F., Webb, D. R., Yates, T., & Davies, M. J. (2015). Pathophysiology of type 1 and type 2 diabetes mellitus: A 90-year perspective. *Postgraduate Medical Journal, 92*(1084), 63-69.

Zhang, Z., Murtagh, F., Poucke, S. V., Lin, S., & Lan, P. (2017). Hierarchical cluster analysis in clinical research with heterogeneous study population: Highlighting its visualization with R. *Annals of Translational Medicine, 5*(4), 75-75.

Zhou, J., Maruatona, O. O., & Wang, W. (2011). Parameter optimization for support vector machine classifier with IO-GA. In *2011 First International Workshop on Complexity and Data Mining* (pp. 117-120). IEEE.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(*2), 301-320.

University of Malaya

Zuhdi, A., Ahmad, W., Zaki, R., Mariapun, J., Ali, R., Sari, N., Hian, S. K. (2016). Acute coronary syndrome in the elderly: The Malaysian national cardiovascular disease database- acute coronary syndrome registry. *Singapore Medical Journal, 57*(04), 191-197.

University of Malaya