A METHODOLOGY BUILDING FOR MULTI-LAYER FEED-FORWARD NEURAL NETWORK (MLFFNN): AN APPLICATION IN BIOMETRY MODELLING

MOHAMAD NASARUDIN BIN ADNAN

UNIVERSITI SAINS MALAYSIA

2023

A METHODOLOGY BUILDING FOR MULTI-LAYER FEED-FORWARD NEURAL NETWORK (MLFFNN): AN APPLICATION IN BIOMETRY MODELLING

by

MOHAMAD NASARUDIN BIN ADNAN

Thesis submitted in fulfilment of the requirements
for the Degree of
Master of Science

May 2023

ACKNOWLEDGEMENT

I am grateful to Allah for the health, protection, and guidance in completing this master's study. The journey has not been an easy one. However, with His permission, it has been smooth and possible. It is a great honour to have this opportunity to extend my gratitude and appreciation to acknowledge the contributions of the many people who supported me during my master's study.

My first debt of gratitude goes to my supervisor, Associate Professor Ts. Dr. Wan Muhamad Amir Bin W Ahmad for his constant guidance, encouragement, patience, continuous support, comments, and endurance during this master's study. I have benefited enormously from his biostatistics knowledge, experience, and expertise. I am also grateful to him for the opportunities to attend biostatistics workshops before and during my master's study.

I dedicate this thesis to my family members, especially my parents, for their continuous support and encouragement throughout this study, and also to each of my siblings, whom I value very much for constantly providing me with love, hope, and continuance courage to endure the challenges I have throughout this study. I would also like to thank my senior, Farah Muna Binti Mohamad Ghazali, who provided helpful suggestions and encouragement in completing this study.

Last but not least, my appreciation goes to the Ministry of Higher Education Malaysia for Fundamental Research Grant Scheme with Project, [Code: FRGS/1/2022/STG06/USM/02/10] and the Universiti Sains Malaysia (USM) for awarding me the GRA-Assist, which covered the study fees has enables me to complete my master's study successfully.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

ANN	Artificial Neural Network
BLR	Binary Logistic Regression
BMI	Body mass index
DFA	Discriminant Function Analysis
F1X	Mental and Behavioral Disorders
F2X	Schizophrenia and Delusional Disorders
F4X	Somatoform Disorders
F6X	Adult Personality Disorders
GDP	Gross Domestic Product
HDL	High-Density Lipoprotein
HPV	Human Papilloma Virus
LM	Linear Model
LM	Logistic Model
MAD	Mean Absolute Deviance
MLFFNN	Multi-Layer Feed-Forward Neural Network
MLinearR	Multiple Linear Regression
MLogisticR	Multiple Logistic Regression
MLP	Multi-Layer Perceptron
MLPNN	Multi-Layer Perceptron Neural Network
MSE	Mean Square Error
MSE.f	Mean Square Error Forecasting
MSE.lm	Mean Square Error Linear Model
MSE.net	Mean Squared Error Neural Network
OLR	Ordered Logistic Regression

OSCC	Oral Squamous Cell Carcinoma
PMSE	Predicted Mean Square Error
WHO	World Health Organization

LIST OF APPENDICES

PEMBINAAN METODOLOGI UNTUK RANGKAIAN NEURAL HADAPAN SUAPAN BERBILANG LAPISAN (MLFFNN): APLIKASINYA DALAM PEMODELAN BIOMETRI

ABSTRAK

Penyelidikan ini bertujuan untuk membangunkan suatu kaedah hibrid bagi Rangkaian Neural Hadapan Suapan Berbilang Lapisan (MLFFNN) dengan dua pendekatan yang berbeza iaitu; (i) Regresi Logistik Berganda (MLogisticR) untuk kaedah pertama, (ii) Regresi Linear Berganda (MLinearR) untuk kaedah kedua. Kaedah hibrid yang dibangunkan adalah dirumuskan berdasarkan butstrap, regresi dan MLFFNN. Pada kaedah pertama, ketepatan kaedah yang dibangunkan diukur pada nilai Min Ralat Kuasa Dua Rangkaian Neural (MSE.net), Purata Sisihan Mutlak (MAD), dan juga peratusan kejituan. Manakala bagi kaedah kedua, Min Ralat Kuasa Dua Rangkaian Neural (MSE.net) dan R^2 akan digunakan untuk menilai prestasi kaedah yang dicadangkan. Semua komponen tersebut berfungsi sebagai kayu ukur untuk menentukan tahap ketepatan dan kecekapan model yang dibangunkan. Perisian sedia ada hanya menghasilkan suatu keputusan yang terbatas. Keperluan terhadap penghasilan keputusan yang lebih baik dengan bersertakan bukti yang kukuh merupakan fokus utama dalam kajian ini. Matlamat utama penyelidikan ini adalah bertujuan untuk membina kaedah hibrid serta menjanakan suatu keputusan secara numerik dan juga visualisasi (perwakilan grafik). Keputusan daripada kedua-dua kajian kes menunjukkan bahawa kaedah hibrid telah berjaya meningkatkan ketepatan, keberkesanan dan juga kecekapan anggaran parameter dalam hasil akhir analisis. Penemuan kajian ini menyumbang kepada pembangunan metodologi penyelidikan, dan mencadangkan keputusan yang lebih tepat untuk proses membuat keputusan.

A METHODOLOGY BUILDING FOR MULTI-LAYER FEED-FORWARD NEURAL NETWORK (MLFFNN): AN APPLICATION IN BIOMETRY MODELLING

ABSTRACT

This research aims to develop a hybrid method for Multi-Layer Feed-Forward Neural Network (MLFFNN) with two different approaches; (i) Multiple Logistic Regression (MLogisticR) for the first method, (ii) Multiple Linear Regression (MLinearR) for the second method. The developed hybrid method is based on bootstrap, regression, and MLFFNN. In the first method, the accuracy of the developed method is measured based on the value of the Mean Squared Error Neural Network (MSE.net), Mean Absolute Deviance (MAD), and the accuracy percentage. While for the second method, Mean Squared Error Neural Network (MSE.net) and R2 will be used to evaluate the performance of the proposed method. All those components serve as a yardstick to determine the accuracy and efficiency of the developed model. Existing software only produces limited results. The main focus of this study is the need for better decision-making with solid evidence. The main goal of this research is to build a hybrid method and generate a numerical result and visualization (graphical representation). The results from both case studies show that the hybrid method has successfully improved the accuracy, effectiveness, and efficiency of parameter estimation in the final results of the analysis. The findings of this study contribute to the development of a comprehensive research methodology in future and suggest more accurate results for the decision-making process.

CHAPTER 1

INTRODUCTION

1.1 Preview of the Chapter

The thesis is thoroughly summarized in this chapter. It clearly discusses the background of the study. The research problem and motivation are covered in the next section, followed by the study's rationale. The conceptual framework presents a flowchart of the entire investigation. The research hypothesis is covered in the following section. The objective section of the study includes general and specific goals. This chapter focuses on the study's scope, methodology, and contribution while considering the research objectives. The study's limitations are discussed in the final section of this chapter, followed by the thesis organization, which describes how the current analysis was conducted in detail.

1.2 Background of Study – A review of Statistical Modelling

Statistical applications are now recognized and accepted as an essential component of research in nearly every scientific field, including economics, education, business management, biology, and many others. In biological sciences, statistical methods are known as biostatistics, one of the branches of applied statistics concerned with applying statistical techniques to infer relationships and draw conclusions (Furrer, 2023).

In any scientific experiment, it is necessary to investigate the relationship between variables to forecast future values and determine which factors cause an outcome. Thus, regression analysis is used to accomplish these goals (Roger, 2017; William *et al.,* 2022). The least-squares method also referred to as the original

regression, was invented by Legendre and Gauss in 1805. Francis Galton created the phrase "regression" in a study report looking at the likelihood of having tall parents and/or children. He discovered that children's average heights "regressed" towards the community's average height (Krashniak & Lamm, 2021). Karl Pearson discovered that regressing tall and short children to the population's mean height resulted in the same results (Ahmad *et al.,* 2017).

Biostatistics was used to analyze the early experiments that served as the foundation for the regression method because they were historically related to the biological sciences. Galton (1875) tested the relationship between daughter and mother seeds' weights using sweat peas. The graph shows a straight line with a positive slope. Consequently, he concluded that the straight regression line was accurate. There has been ongoing research since Galton's regression invention and regression methodology advancements. However, the definition of regression today differs from Galton's definition, in which "regression analysis" is defined as "the study of the dependence of one variable on one or more explanatory variables to estimate and predict a mean or average of the population" (Ahmad *et al.,* 2018).

Regression analysis is a technique that can be used in almost any field or discipline, including finance, social, physical, chemical, and health sciences. Its primary purpose is to forecast the dependant variable's value when there are independent variables' values and assess how closely related they are. Most medical researchers used regression analysis to make accurate diagnoses. Regression models are divided into two models: i) linear regression; when two variables have a relationship in a straight line, ii) non-linear regression; when two variables do not have a relationship in a straight line. The most popular kind of regression analysis is linear

regression. It can estimate the strength of one dependent variable (scaled or continuous) and one independent variable (nominal, ordinal, or scale) (Ross, 2020).

Multiple linear regression is a more complex version of simple linear regression that involves multiple explanatory variables (Ross, 2020). It was used in research with data that was linearly distributed. In the medical case study, several studies applied multiple linear regression; for instance, studied done by Nawi *et al.,* in 2018, the multiple linear regression model was applied to predict the patients' diabetes mellitus by using patients' body mass index (BMI), weight, cholesterol level, height, and systolic blood pressure as the independent variables (Nawi *et al.,* 2018).

Another regression model is logistic regression, which determines the relationship between a categorical variable and one or more categorical independent variables. In 2020, Yaqoob *et al.,* was used multiple logistic regression to investigate the prevalence of Human Papilloma Virus (HPV) 16 in forty-one specimens from patients who were diagnosed with Oral Squamous Cell Carcinoma (OSCC) and its association with the sociodemographic characteristics among the Kelantan population. There is a higher prevalence in affect Kelantan population compared to the 3% global prevalence by World Health Organization (WHO) (Yaqoob *et al.,* 2020).

The study proposed by Ahmad *et al.,* (2016) showed that hybrid methodology work effectively but not yet explore in detail, and high-accuracy results could be obtained from the hybrid statistical methodology. The hybrid methodology aims to improve the models' accuracy and predictability by combining two or more approaches. It produces better results than standalone ones (Ahmad *et al.,* 2016). One study by Adnan *et al.,* (2023) on 200 simulated forensic data was done using a robust hybrid methodology between applied linear regression model (ALRM) and multilayer perceptron (MLP) and it shows that this hybrid method was successfully. Thus, the

current study aims to integrate the MLFFNN with Multiple Logistic Regression (MLogisticR) for medical cases and MLFFNN with Multiple Linear Regression (MLinearR) for forensic crime cases. In addition, bootstrapping was applied to improve the regression models' predictability and accuracy. This study is expected to help the researchers, especially regarding validation and prediction (Ahmad *et al.*, 2020).

1.3 Research Motivation and Research Problem/Statement

Research done in the past has shown the important links between the medical and forensic sciences regarding statistical modelling. Medical and forensic sciences require extensive research to identify the causal factor, and selecting a reliable and valid biometric model is essential for precisely mapping the issue. This research will utilize a recently constructed hybrid model to ascertain the association between studied parameters. The term "cause-and-effect" refers to deciding how independent variables influence dependent ones. Because of this, it is necessary to make significant modification changes to both methods to bridge the research gap. This study attempts to bridge the functional gap between the two current methodologies by developing the hybrid model. The performance of the developed method will be assessed using the reading of Mean Square Error Forecasting (MSE.F), Mean Absolute Deviance (MAD), and accuracy. These two result readings are obtained after these two processes have been synchronized. The statistical method used to estimate the model is most important for producing an accurate and exhaustive map. The combination of several statistical techniques has not yet been widely explored, especially in studies on chronic diseases and forensic cases in Malaysia. Therefore, it has the high potential to solve health sciences problems more accurately.

However, in a previous study, most researchers focused on a single model technique, which had limited potential for solving the problems associated with chronic disease and criminal cases. This is because, in some cases, the proposed method is not feasible. Thus, statistical models incorporating hybrid techniques would enable a more precise estimation of the modelling purpose. In addition, the parsimony rule with the fewest assumptions possible is also being considered in this study. Linear models or logistic models with MLFFNN are the focus of this study because it is believed that this integrated model can be applied to various space-time combinations. In addition, the proposed integrated model can be used to investigate the existence of "cause-and-effect" and relate it with several potential factors contributing to the studied problem.

1.4 The Rationale of the Study

In medical and forensic sciences, specific research methodology on the method development is still lacking especially involving integrating the method or procedures. This study proposes two methodologies proposed; MLFFNN with MLogisticR and MLFFNN with MLinearR (see Figure 1.1). Demand for highly accurate models and optimized output is increasing nowadays. Therefore, this study focuses on integrating high-accuracy models with high reliability to maximize the relevance of the findings. Most research-based methodology development lacks a model validation mechanism, incorporating modelling techniques like bootstrapping. As a consequence, solutions to this issue must be developed. This problem can be solved by integrating the current methodology or refining or modifying current methods.

1.5 Conceptual Framework of the Research Study

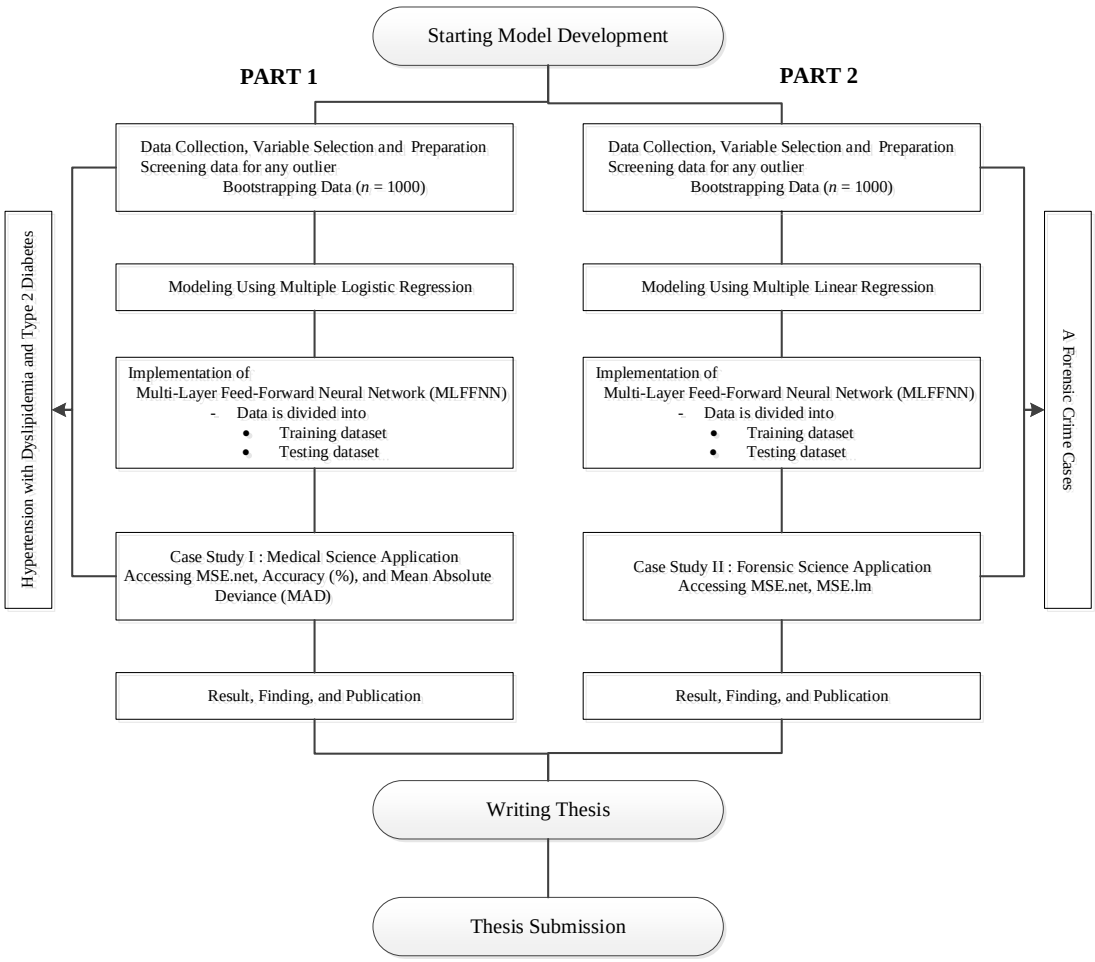

Figure 1.1: Conceptual framework of the study

This study's conceptual framework aimed to develop an extensive methodology by integrating MLFFNN with MLogisticR for medical health cases (Case I) and MLFFNN with MLinearR for forensic crime cases (Case II), as in Figure 1.1.

1.6 Research Hypothesis

This research is crucial to the fundamental development of statistical computing methodology. The following are the hypothesis:

i. The integrated model can improve the level of analysis in terms of efficacy and efficiency.

ii. The integrated model can serve as a benchmark for biometry modelling.

1.7 Research Objectives

1.7.1 General Objectives

To develop a methodology for a Multi-Layer Feed-Forward Neural Network (MLFFNN) with application in medical health and forensic crime case.

1.7.2 Specific Objectives

i. To develop an integrated Multi-Layer Feed-Forward Neural Network (MLFFNN) model.

ii. To determine and validate the best variable selection using the developed integrated Multi-Layer Feed-Forward Neural Network (MLFFNN) model.

1.8 Scope and Methodology

The scope of the study underlying is performing biometry modelling with integrating other statistical procedures. In this thesis, an alternative method (known as the integrated method) has been constructed and introduced, which involves a combination of several main statistical methods that focus on more accurate research findings. Two methods have been approached in this research; methodology-based building for MLogisticR and methodology-based building for MLinearR.

The first methodology in this study focuses on data bootstrapping, MLogisticR, MLinearR, and MLFFNN, which are the four components that are used in model building. Second, the proposed approach will be used in two different case studies;

medical and forensic crime studies. The developed method focuses on the integrated MLogisticR for medical studies, while the forensic crime study will be on the integrated MLinearR. Both of the cases will be evaluated through the obtained result by looking at the performance of their accuracy and predictivity.

1.9 The Contribution of the Study

The contribution can be summarized into two sections. The first contribution focuses on developing an integrated methodology that can apply in medical and forensic crime cases for future predictions. This study will significantly contribute to methodology development consisting of two different proposed methods; MLogisticR and MLinearR.

The second contribution is to the R-syntax for statistical tools. This enables the researcher to design the need of the R-syntax for improved result performance and determine the best variable selection. In addition, this syntax offers a thorough output for the researcher's conclusions, particularly for decision-making.

Thirdly, this work adds to the scientific literature work and allows other researchers to continue future research work. This will assist the new researcher in comprehending the methodology-based difficulty while designing new methodologies or doing research. On top of that, this study can suggest a superior method for modelling purposes with highly predictable and accurate.

1.10 Limitation of Study

The limitation of the study had been determined as four components only, which are

a. Bootstrapping

- Only involving case resampling of data which aims the refining of parameters estimated

b. Multiple Logistic Regression and Multiple Linear Regression

- Only using the method enter, all the variables included are already clean and have clinically significant importance from the previous literature.

c. Multi-Layer Feed-Forward Neural Network

- This methodology was used only to validate the factors obtained in (b); calculate the accuracy, and assess the methodology's predictability. The predicted results (testing data) will be compared to the actual value reflecting the proposed methods' goodness.

The second limitation is the selection of the software used. Since R is a free software program and user-friendly, all the developed methods will only be utilized with this program. The data itself came in as the third limitation. Only secondary data were employed for better understanding and visualization in this research. The first case was on Hypertension with Dyslipidemia and Type 2 Diabetes, and the second was from the forensic crime cases. All data used sources were gathered from textbooks.

1.11 Organization of the Thesis

This thesis comprises five chapters and is organized in the following manner. The first chapter is an introduction, focusing on the study's background and the context within which this study is being carried out. This chapter also discusses the study's objectives, rationale, scope and methodology, significance, and limitations of the current study.

In the second chapter, a literature review of the statistical methods that will be integrated is presented. This chapter also reviews and explores the previous statistical methods used in medical health and forensic crime cases. Besides that, this chapter will review those methods and also their limitations.

The third chapter is methodology, contains comprehensive information on the procedure and the statistical models, namely MLFFNN, MLogisticR, MLinearR, and bootstrap. In short, this chapter describes the study design, the location of the study, the study period, and case studies. This chapter also included a flow chart of the study based on the proposed statistical logistic and linear modelling.

Next, the fourth and fifth chapters present the results and discussion of each case study. A detailed discussion of each analysis was done in this chapter. The last chapter focuses on the conclusions derived from the results of the findings in detail. This chapter also briefly discussed the aspects of the analysis and interpretation of the model building by giving a few recommendations to produce better research in the future.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview

Sections 2.2 to Section 2.4 discussed in general terms Logistic Regression, MLogisticR, and MLinearR. The approach of bootstrapping is also discussed in Section 2.5. The accessibility and application of MLFFNN in science fields are also discussed in Section 2.6. Sections 2.7 and 2.8 briefly explain both case studies with the previous statistical methods applied. Furthermore, The necessity of a combination strategy is emphasized in Section 2.8 due to the limitations of statistical approaches. The last section provides concluding remarks.

2.2 History of Regression in Science Fields

Sir Francis Galton is a British scholar who first proposed the concept of regression in a paper entitled "Typical Laws of Heredity" in 1877 in England. His research focuses more on causality, which involves the relationship between variables. He experimented in 1886 to determine the growth rate of peas concerning the size of their seeds. The seeds have been divided into seven different groups. One hundred produced seeds' diameters were measured, and the average value was recorded. The findings showed that the growth rate of small-sized bean seeds was very high compared to that of large-sized bean seeds. He also discovered that most average diameters for peanuts "regress" within the normal average range. His modelling revealed two key findings: 1) The diameter of the peanuts has a linear relationship with the diameter of the peanut trees. 2) The average diameter of the peanuts is "regress" within the normal average range (Dhakal, 2018).

He also studied the relationship between the height of children with their parents. A set of 205 data was used. He discovered that children are marginally taller than their parents if both of their parents' height measurements are low. However, their children's height will be slightly shorter than their parents if one or both of their parents are tall. He named this phenomenon as "Regression Toward Mediocrity". This phenomenon was referred as "revert" at first, then changed to "regress". He discovered the term "regression" to describe related events. Due to this, the basic idea of linear regression was developed (Suarez *et al.*, 2017).

2.3 Biometry Modelling using Multiple Logistic Regression Model

One study was done by Adwere (2011), on high school students using MLogisticR to predict cigarette smoking behaviour from selected predictors from the 2009 CDC Youth Risk Behaviour Surveillance Survey. The target student's behaviour of interest was based on frequent cigarette use. The study included five predictor variables which are a) race, b) initial cigarette smoking age, c) frequency of cocaine use, d) feeling sad, and e) physically inactive. The results show the model was statistically significant, which considered those variables together, and the strongest predictors were race, physically inactive, and frequency of cocaine use (Adwere, 2011).

MLogisticR was applied by Ghazali *et al.*, (2020) to determine the independent variables related to an elderly health status at receiving home care. Then they validated by Artificial Neural Network (ANN) through Multi-Layer Perceptron Neural Network (MLPNN). Four independent variables involve; gender, education level, duration of stay in receiving home care, and status of physiological stress or acute disease. Results

show those variables are significantly related to the elderly health status (Ghazali *et al.,* 2020).

2.3 Biometry Modelling using Multiple Linear Regression Model

In a study by Ahmad *et.al.,* (2016), triglyceride levels were used as a dependent variable in MLinearR models. They regressed it with the hip circumference, blood parameters, lipid-lowering medication, and weight (Ahmad *et al.,* 2016). Other study by Nawi *et.al.,* (2018), diabetes mellitus (dependent variable) was diagnosed in patients using body mass index (BMI), height, weight, cholesterol level, and systolic blood pressure as independent variables (Nawi *et al.,* 2018). Furthermore, in 2019, Khan *et.al.,* determined the knowledge scores of medical and non-medical students, along with age, university, study year, and discipline (Khan *et.al.,* 2019).

Several studies have used MLinearR models to predict disease prevalence and health-related physical fitness. The estimated time needed to complete surgical procedures was forecasted by researchers using a linear regression model to predict patient flow in healthcare facilities. Several studies have used MLinearR models to predict disease prevalence and health-related physical fitness (Edelman *et al.,* 2017; Baihaqi *et al.,* 2018; Shah *et al.,* 2020). Two studies used the MLinearR model to calculate the newborn mortality rate or estimate life expectancy based on the gross domestic product (GDP) and population size (Palupi & Rizki, 2020; Rubi *et al.,* 2021).

2.5 Bootstrapping Approach

In 20th century, the American statistician Bradley Efron developed the idea of bootstrapping (Efron, 1993). In the past four decades, the application of bootstrapping has expanded to include a variety of bootstrapping techniques, including parametric

and Bayesian bootstrapping (LaFontaine, 2021). It took many years for the bootstrapping methodology in statistics to gain widespread acceptance. Most people did not understand how the procedures worked or accepted the underlying assumption (Champkin, 2010). Bootstrap resampling methods save time when taking the initial sample compared to manual analysis. The software capabilities prevented many from conducting exhaustive research or empirical tests. The bootstrap method of statistical inference randomly resamples data up to 10,000 times (with replacement, which means that as soon as an item is sampled, it is immediately replaced). A random sample can estimate population distribution (LaFontaine, 2021).

Furthermore, to more accurately represent the unidentified population, this random sample should use bootstrapping (LaFontaine, 2021). Bootstrap does not rely on a hypothetical sampling distribution, in contrast to statistical significance testing. The bootstrap method uses constant resampling with replacement to generate an empirical distribution for a sample statistic, which calculates means, standard errors, and confidence intervals. The bootstrap technique relies on sampling or resampling with replacement, so there is no limit to sample combinations (Efron *et al.*, 1993).

2.6 Accessibility of Multi-Layer Feed-Forward Neural Network (MLFFNN)

The fundamental components of artificial neural networks (ANN) are known as “neurons” (Pazikadin *et al.*, 2020). The biological brain and nervous system served as inspiration for the ANN's name, design, and structure. (Moayedi *et al.*, 2020). Despite the fact that there are numerous varieties of neural networks, their fundamental principles are still the same (Hasson *et al.*, 2020). Neural networking architecture is classified into two groups based on the connection between neurons: “feed-forward

neural network" and "recurrent neural network". If the synaptic connections between the neurons only allow information to be transmitted forward, the network is described as a "feed-forward neural network (MLFFNN)". However, the network is called a "recurrent neural network" when the synaptic connections transfer the information from the output to input neurons (Li *et al.,* 2018). The MLFFNN is composed of three layers, including an input layer, one or more hidden layers, and an output layer (Faris *et al.,* 2019). Usually, layers are used to organize the neurons in an MLFFNN. The direction of signal flow from the input layer to the output layer is directional but not within the same layer. In medical studies, the MLFFNN has been used as a validation approach for the independent variables, for example, in MLinearR. MLFFNN has two primary components: training and testing data sets. The original data are divided into 60% or more for training the network and the rest for network testing (Rachmadi *et al.,* 2018; Mijwel *et al.,* 2019).

Two MLFFNN models were developed by the authors, one model using all five independent variables (age, sex, dimensions of various body parts, serum cholesterol level, and blood pressure) and another model with only four independent (height, weight, arm length, and pulse). In the regression model, both models were found to be significant. Comparing the mean square error (MSE) of the two MLFFNN models, it was determined that the model with only four independent variables was more effective (Mohamed *et al.,* 2011).

The survival of oral cancer' patients was studied in 2018 by Ahmad et al. It was discovered that using MLFFNN networking to test the model was a great way to make a model that could generate more precise forecasts and decisions in the future (Ahmad *et al.,* 2018). Mostly, the researchers have used the MLFFNN to validate the parameters linked to the dependent variables. Some have used the mean square error

(MSE) to verify their regression models. On the other hand, other researchers have relied on the accuracy, sensitivity, and specificity of neural networks to validate their independent variables without developing a regression model (Ferenci *et al.,* 2018; Güldoğan *et al.,* 2020; Çiçek & Küçükakçali, 2020).

2.7 Application of Statistical Methods

2.7.1 Case Study I: Hypertension with Dyslipidemia and Type 2 Diabetes

In 2021, Chai *et al.,* developed Multi-Layer Perceptron Neural Network (MLPNN) model for adolescent hypertension classification, focusing on using simple anthropometric and sociodemographic data in Sarawak, Malaysia. Eleven anthropometric and sociodemographic data were collected. Only five parameters, weight, weight-to-height ratio, age, sex, and ethnicity, were selected and used as the input for the developed model. Out of 2461 data collected, 741(30.1%) were hypertensive and 1720(69.9%) were normal. The developed model was well-generalized by analyzing the performance metrics obtained from the training, validation, and testing data sets (Chai *et al.,* 2021).

In 2020, Ahmad et al. studied hypertension with dyslipidemia and type 2 diabetes mellitus towards the biochemical profiles among patients at Hospital Universiti Sains Malaysia. The independent t-test determined the differences between biochemical profiles regarding hypertension and dyslipidemia characteristics. Of all those 23 variables, only three factors show significant differences: fasting blood sugar, high-density lipoprotein (HDL), and urea reading. The Ordered Logistic Regression (OLR) and Multi-Layer Perceptron (MLP) were used to determine the variable selection to predict hypertension status. At the first stage, the researchers screened five variables (smoking status, alcohol consumption, total cholesterol, fasting blood

glucose, and triglyceride levels) through OLR and bootstrap methodology. After considering 1500 of the bootstrap methods, it was found that smoking status, total cholesterol, and triglycerides were significantly associated with hypertension. After that, these three variables were selected based on the level of significance of 0.25 for OLR and used for the input of the MLP model. The model's accuracy was evaluated through the predicted mean square error (PMSE) value. The obtained value was 0.1564. Thus, there was a strong relationship between hypertension, smoking status, total cholesterol, and triglyceride levels (Ghazali *et al.,* 2020).

2.7.2 Case Study II: A Forensic Crime Case

Studies in Norway from 1996 to 2017 used MLogisticR to determine the factors associated with the deaths undergoing forensic autopsy. The independent variables were age, sex, place of death, police district, population size, urbanity level of the municipality, and distance to the autopsy facility. From those variables, only three variables (age, place of death, and police district) were significantly associated with the deaths undergoing the forensic autopsy (Christian *et al.,* 2021). In China, the researchers studied the influencing factors of criminal responsibility in patients with mental disorders from 2010 to 2020 using the Chi-square test and nominal regression analysis. Out of 437 patients, there were 361 males from 15 years to 91 years. After the assessment, it shows that the influencing factors of criminal responsibility were a crime in public places, a crime in the victim's residence, a crime in the suspect's residence, forensic psychiatric diagnosis of F1X (mental and behavioural disorders), F2X (schizophrenia and delusional disorders), F4X (somatoform disorders), and F6X (adult personality disorders), the criminal object of property, and case types of theft (Sun *et al.,* 2023).

A sample of 213 adult male-on-female homicides with sexual or unknown motives was drawn from a United Kingdom (UK) database. This study used the bivariate statistical approach to explore the relationship between 13 pre-conviction variables and 29 crime scene behaviours. The presence or absence of particular pre-convictions was then predicted using binary Logistic Regression models based on a combination of offense-related behaviours. Analysis revealed 16 statistically significant correlations between key criminal behaviours and previous convictions. However, these correlations were frequently "less likely" to result in prior convictions. Furthermore, this analysis failed to find any correlations for various other variables. The findings show that predicting an offender's criminal history is possible based on their behaviours during an offense, though not all pre-convictions may be equally suitable (Almond *et al.,* 2021).

Two studies done by Matthew in 2020 about Establishing the Factors Related to Domestic Burglary used regression analysis. The first study examines the associations between burglary propensity to burglary-related cognitive distortions, general criminal cognitive distortions, empathy, and human needs. In this study, 306 burglaries were studied. The study shows that burglary propensity is associated with general criminal cognitive distortions and empathy. The second study examines the associations between burglary behaviours with burglary proclivity, burglary empathy, cognitive burglary distortions, and emotional reactivity. A sample of 51 burglaries simulated as first-time burglary offenses were examined. Results show that all those variables were associated with burglary behaviours (Matthew, 2020).

The comparison of statistical methods for estimating the sex in the forensic case was studied by Rani *et al.,* 2022. They were looking at the accuracy of two models, discriminant function analysis (DFA) and Binary Logistic Regression (BLR),

in estimating the sex from human external ear anthropometry. The participants (233 males and 264 females) were obtained from the Himachal Pradesh state of North India. A total of 12 anthropometric measurements were taken independently on each participant's right and left ear. The variable sex was discriminated by using DFA and BLR. However, the predictive percentage of sex estimation from both methods was substantially the same as DFA (76.3%) and BLR (76.2%). In contrast, the correct predicted percentage value was 0.1% higher in DFA than BLR (Rani *et al.,* 2022).

2.8 Review of Methodology and Limitations

According to the literature review, combining MLFFNN with MLogisticR and MLinearR in medical and forensic studies is less common among researchers. Validating and modelling datasets for future predictions is critical for increasing the significance of research outputs and the accuracy of results. Thus, it is crucial to integrate the methodology that enables medical and forensic researchers to predict future trends for future planning. Combining the bootstrapping method with regression modelling and MLFFNN is an alternative method to solve the problem of having a small sample. The robustness of the bootstrap method (through the production of data replication through a sampling process) in regression modelling can be seen through the results of research conducted by Efron *et al.,* (1993), Booth *et al.,* (1998), and Davidson *et al.,* (2000) (Efron *et al.,* 1993; Booth *et al.,* 1998; Davidson *et al.,* 2000). In general, this bootstrap method is introduced to estimate a statistical distribution through an empirical distribution approach to improve the accuracy of statistical estimation values.

2.9 Concluding Remarks

In the review of relevant literature, the integrated statistical methods or tools, including bootstrap, MLogisticR, MLinearR, and MLFFNN, have been tested and employed countless in the literature. However, the performance of integrated based on their predictability, success, and accuracy has been proven. Most researchers apply only a single statistical method and directly lead to the results. However, there is a chance that the outcome might be improved by utilizing certain integrated models. Owing to the limitations of applying a single statistical method, the potential of the improved methodology should be investigated from the previous study. Thus, this study is designed to integrate two different methods. The first method

(i) The First method considers a combination of bootstrap and MLogisticR with MLFFNN.

(ii) The second method considers a combination of bootstrap and MLinearR with MLFFNN.

All the proposed methods will use the R-syntax programming language.

CHAPTER 3

METHODOLOGY

3.1 Overview of Propose Methodology

This study was divided up into two main sections. The first section involved the development of a computational methodology in R-syntax using R software. All the steps are mentioned clearly in this chapter. The second phase of the research is to apply the developed method using two datasets. The data used were all secondary sources obtained from the textbook mentioned in the references section.

3.2 Study Design

This study is a computational biostatistical study design focused on methodology development. The study mainly focuses on integrating and harmonizing the statistical tool. All the tools are being synchronized for excellent and optimized results.

3.3 Study Population

The secondary data from the books are used as the study population in this study in order to test the methodology. The data being used in this study are health record and crime cases that focus on the hypertension associated with dyslipidemia and type 2 diabetes mellitus, and total crime case, respectively. The data will be inserted in R software to create an accurate model of health issue and crime cases and to evaluate the model's accuracy.

3.4 Study Period

The total duration of the study is one year. It started in April 2022 and ended in March 2023.

3.5 Study Location

The study has taken place in the School of Dental Sciences (PPSG), Universiti Sains Malaysia (USM), Kubang Kerian, 16150 Kelantan.

3.6 Reference and sources Population

Since this is a computational study, no reference or source population exists. The research focuses solely on the computational approach and results for demonstrating enhanced performance.

3.7 Ethical Considerations

The Human Ethics Committee of Universiti Sains Malaysia, where the study was conducted, provided ethical approval (Appendix A). The Jawatankuasa Etika Penyelidikan Manusia, USM, Malaysia (JEPeM code: USM/JEPeM/22040245) was followed in the conduct of this study**.**

3.8 Sampling Frame

In this study, no sampling frame was used. Therefore, there are no criteria for inclusion and exclusion.

3.9 Software Used in Research

This study is only employed R software exclusively for research purposes.

3.10 The variables

This study uses two different case studies as examples. Below is a list of the variables used in this specific case study.

Case Study I: Hypertension with Dyslipidemia and Type 2 Diabetes

This study looked at data (secondary data) from patients who attend Hospital Universiti Sains Malaysia's Klinik Rawatan Keluarga (KRK). Thirty-nine patients were included in the study. Table 1.1 summarises the research variables' data descriptions.

Table 1.1: Data description of the selected blood profile

Variable	Code	Description
Hyper	Y	Hypertension [0 = Yes, 1 = No]
Marital	X_1	Marital status [0 = Married, 1 = Single, 2 = Widow/widower]
Sysbp	X_2	Systolic blood pressure
Tc	X_3	Total cholesterol
Hdl	X_4	High-density lipoprotein
Alp	X_5	Alkaline phosphatase
Urea	X_6	Urea reading

Case Study II: A Forensic Crime Case

The second case study is a forensic case study for which data was taken from the textbook. The complete description of the variable is provided in Table 1.2.

Table 1.2: Data description for the criminal case, which focuses on the number of times respondents were victimized in a year

Variables	Code	Descriptions
Total victim	Y	The number of times respondents was victimized in a year.
Gender	X_1	Gender [1 = Male, 2 = Female]
Age	X_2	Age in years
Marital	X_3	Marital status [1 = Married, 2 = Divorced, 3 = Separated, 4 = Other]
Socclass	X_4	Social class [1 = Daily base working, 2 = Middle class, 3 = Upper middle class, 4 = Not working]
Adult	X_5	Number of adults in the household
Children	X_6	Number of children in the household
Burglary	X_7	Number of times victimized by burglary
Sexual	X_8	The number of times victimized by a sexual offense.
Report	X_9	Number of victimizations reported to the police
Location	X_{10}	Location of household [1 = Urban , 2 = Sub Urban , 3 = Rural]

3.11 The Four Main Components for the Methodology Building using the R-syntax

This section discusses the research's methodology, including the R programming language, bootstrapping for logistic and linear regression, and model validation using MLFFNN. The syntax for all the suggested methodologies will be R. More information is provided in the following subsections; Section 3.10.1 until Section 3.10.4.

3.11.1 Data Bootstrapping

Bootstrap first calculates sample statistics from a random subset of the population. A pseudo-population is then established by repeatedly drawing from the same set of initial samples, after which the bootstrap draws many substitution samples. The capability of the bootstrap to generate a sample of the same size as the initial sample, with some findings repeated multiple times and others eliminated. Random sampling with substitution yields samples that are not identical to the original sample. While drawing a large number of samples with replacement, the bootstrap calculates statistics for each sample (Efron & Tibshirani, 1993).

3.11.2 Multi-Layer Feed Forward Neural Network (MLFFNN)

In this study, MLFFNN is generally grouped into layers that are divided into input layers, hidden layers, and output layers. In this research, the output node is fixed at one since only one dependent variable exists. In MLFFNN the values $\hat{y}$ are given by $\hat{Y} = g_i\left(\sum_{j=1}^{H} w_j h_j + w_0\right)$ where w_j an output weight from the hidden node j to the output node is w_0 the bias for the output node, g which is an activation function. The