A COMBINATION OF METHODOLOGY BUILDING FOR MULTI-LAYER FEED-FORWARD NEURAL NETWORK (MLFF) AND LINEAR MODELING: AN APPLICATION IN BIOMETRY MODELING

MUHAMMAD KHAIRAN SHAZUAN BIN JUSOFF

UNIVERSITI SAINS MALAYSIA

2024

A COMBINATION OF METHODOLOGY BUILDING FOR MULTI-LAYER FEED-FORWARD NEURAL NETWORK (MLFF) AND LINEAR MODELING: AN APPLICATION IN BIOMETRY MODELING

by

MUHAMMAD KHAIRAN SHAZUAN BIN JUSOFF

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

January 2024

ACKNOWLEDGEMENT

First and foremost, all praises and blessings belong to Allah the Almighty, the Most Gracious and Most Merciful, for granting me strength, wisdom, and guidance throughout my study journey.

First and foremost, I would like to sincerely express my gratitude to my supervisor, Associate Professor Ts. Dr. Wan Muhamad Amir Bin W Ahmad for his supervision, constant guidance, and positive encouragement throughout being a postgraduate student under his supervision. I would also like to thank him for the exposure and opportunities to participate in online biostatistics workshops during my master's study. May Allah bless you with good health, happiness, and prosperity.

I am forever thankful to my family members, Jusoff bin Hassan (father), Hanisah binti Abdullah (mother), Nor Azwani binti Jusoff (sister), and Muhammad Khairul bin Jusoff (brother). Their faith and the sacrifices made on my behalf have driven me to complete my study. Apart from that, I would also like to acknowledge my colleague, Mohamad Nasarudin bin Adnan, for the help and shared experiences that made my study journey much smoother.

I am also taking this chance to humbly thank you to those who have contributed directly or indirectly to my study journey. I pray that Allah showers His blessings upon everyone who is involved.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Page

LIST OF SYMBOLS

$\bar{x}$	Mean
$\hat{y}$	Expected Independent Variable
x	Independent variable
y	Dependent variable
n	Number of sample (n=1,2,3,…,N)
r^2	Coefficient of determination
β	Coefficient value
f(x)	Neural Network function

LIST OF ABBREVIATIONS

ADALINE	Adaptive Linear Element
AI	Artificial Intelligence
AIC	Akaike’s Information Criterion
ANN	Artificial Neural Network
ART1	Adaptive Resonance Theory
BIC	Bayesian Information Criterion
BMI	Body Mass Index
CASI	Cognitive Ability Screening Instrument
CKD	Chronic Kidney Disease
COVID-19	Coronavirus disease 2019
CVD	Cardiovascular Disease
ENN	Ensemble Neural Network
FBS	Fasting Blood Sugar
FFNN	Feed-forward Neural Network
GDM	Gestational Diabetes Mellitus
GFR	Glomerular Filtration Rate
GRNN	General Regression Neural Network
HBA1C	Glycated Haemoglobin
IFG	Impaired fasting glucose
IGT	Impaired Glucose Tolerance
INNS	International Neural Network Society
KRK	Klinik Rawatan Keluarga
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
ML	Machine Learning
MLE	Maximum Likelihood Estimator
MLFF	Multi-Layer Feed Forward Neural Network
MLP	Multilayer Perceptron
MLR	Multiple Linear Regression
MMRE	Mean Magnitude of Relative Error
MNAE	Mean Normalised Average Error

MSE	Mean Square Error
MSE.lm	Mean Square Error for Linear Model
MSE.net	Mean Square Error for Neural Network
PPSG	School of Dental Sciences
qSOFA	Quick Sequential Organ Failure Assessment
RMSE	Root Mean Squared Error
RTRL	Real-time recurrent learning
SNN	Single Neural Network
USM	Universiti Sains Malaysia

LIST OF APPENDICES

GABUNGAN PEMBINAAN METODOLOGI UNTUK RANGKAIAN NEURAL HADAPAN SUAPAN BERBILANG LAPISAN (MLFF) DAN PEMODELAN LINEAR: APLIKASI DALAM PEMODELAN BIOMETRI

ABSTRAK

Biostatistik, juga dikenali sebagai biometri ialah bidang statistik yang memfokuskan kepada aplikasi kaedah statistik dalam bidang bioperubatan dan sains kesihatan. Biostatistik boleh membantu penyelidik dan pengamal perubatan dalam mengenalpasti faktor risiko, penilaian keberkesanan rawatam dan banyak lagi. Walaubagaimanapun, biostatistik belum diterima sepenuhnya oleh pengamal perubatan profesional kerana beberapa sebab. Salah satu sebab utama ialah bidang perubatan adalah mencabar, dan mengekalkan tahap ketepatan yang tinggi adalah kritikal. Di samping itu, banyak kajian terdahulu tertumpu kepada teknik pemodelan individu yang mempunyai keupayaan yang terhad untuk mengkaji bidang perubatan yang dinamik dan kompleks. Kajian ini bertujuan untuk membina model biometri yang menggabungkan beberapa teknik statistik iaitu bootstrap, Rangkaian Neural Hadapan Suapan Berbilang Lapisan (MLFF) dan Regresi Linear Berganda (MLR). Kajian ini akan mencadangkan dua model berbeza: (i) Model MLFF-MLR hibrid dengan bootstrap dan (ii) Model MLFF-MLR tanpa bootstrap. Kedua-dua model tersebut akan dibandingkan menggunakan Min Ralat Kuasa Dua Rangkaian Neural (MSE.net) dan Min Ralat Kuasa Dua Model Linear (MSE.lm). Model dengan nilai MSE.net dan MSE.lm yang lebih rendah akan dianggap lebih baik. Hasil analisis daripada kedua-dua model menunjukkan model MLFF-MLR dengan bootstrap adalah lebih tepat. Penyelidikan ini menyumbang kepada kandungan ilmu dengan meneroka potensi pemodelan biometri dan boleh menjadi rujukan kepada penyelidik masa hadapan.

A COMBINATION OF METHODOLOGY BUILDING FOR MULTI-LAYER FEED-FORWARD NEURAL NETWORK (MLFF) AND LINEAR MODELING: AN APPLICATION IN BIOMETRY MODELING

ABSTRACT

Biostatistics, also known as biometry, is a field of statistics that focuses on the application of statistical methods to the field of biomedicine and health sciences. Biostatistics can assist researchers and healthcare professionals in identifying risk factors, evaluating intervention effectiveness and many more. However, biostatistics has not been entirely embraced by medical professionals due to several reasons. One of the main reasons is that the medical field is challenging, and maintaining a high level of accuracy is critical. In addition, many previous studies focused on individual modeling technique that has limited ability to capture the dynamic and complexities in the medical field. This study aims to develop a biometry model that combines several statistical techniques, namely bootstrap, Multi-Layer Feed-Forward Neural Network (MLFF) and Multiple Linear Regression (MLR). This study will propose two distinct models: (i) Hybrid MLFF-MLR model with case resampling and (ii) Hybrid MLFF-MLR model without case resampling. The two models will be compared using the Mean Square Error of Neural Network (MSE.net) and the Mean Square Error of the Linear Model (MSE.lm). The model with lower MSE.net and MSE.lm values will be deemed superior. The analysis results from both models show that the hybrid MLFF-MLR model with case resampling yields a more accurate output. This research contributes to the body of knowledge by exploring the potential of biometry modeling and can be a reference for future researchers in the same field.

CHAPTER 1

INTRODUCTION

1.1 Chapter Overview

This chapter gives an overview of the general information of this research. The second section in this chapter highlights the workflow and application of Multi-Layer Feed Forward Neural Network (MLFF) in analysing medical data. The third section describes the problems when applying statistical techniques to medical data. The fourth section elucidates the significance of the study, while the fifth section highlights the conceptual framework of the study. The sixth section briefly discusses the research hypothesis. Next, seventh section explains the study objectives, encompassing both general and specific objectives. Meanwhile, the eighth section highlight the scope of this study, while the following section outlines the contributions of the research. Moreover, the tenth section discuss the study's limitation, followed by the eleventh section which highlight the organization of this thesis.

1.2 Background of the Study

Statistics plays a crucial role in the medical and health field, helping researchers and practitioners make sense of complex data, draw meaningful conclusions, and make evidence-based decisions. Biostatistics, or biometry, is a specialized branch of statistics focusing on applying statistical methods in the biomedical and health sciences. Biostatistics is critical in clinical trials, epidemiology, observational studies, survival analysis, diagnostic testing, medical imaging, health services research, and meta-analysis (Cadarso-Suárez & González-Manteiga, 2007). Using these statistical techniques, biostatistics enables researchers and healthcare professionals to make evidence-based decisions, improve patient outcomes, evaluate treatment effectiveness, identify risk factors, and inform public health policies. Given the increasing complexity and quantity of health-related data, the emphasis on accelerating clinical and translational science, and the importance of conducting reproducible research, the need for the thoughtful development of biostatistics is growing.

In medical science, the MLFF is widely employed. MLFF is a mathematical or computational model replicating the human brain's learning process. It achieves this by utilizing rules derived from data patterns to construct hidden layers of logic used in the analysis (Fashoto, 2015). MLFF incorporates a perceptual interconnection where data and computations follow a single direction path, moving from input to output data. In a linear regression context, MLFF aims to evaluate independent variables with dependent variables. The high-quality technique is pivotal in data analysis, ensuring accurate and precise outcomes. Therefore, it is essential to emphasize the fundamental aspect of computational precision and accuracy to establish a seamless connection between theoretical concepts and practical programming.

Regression analysis is a powerful statistical tool used to examine the association between variables that exhibit a cause-and-effect relationship. In particular, univariate regression focuses on analysing the connection between a dependent variable and a single independent variable, allowing for the establishment of a linear relationship between them. However, when the model needs to consider multiple independent variables, the approach transitions into multilinear regression, also known as multiple linear regression (MLR). In MLR, the aim is to construct a regression model incorporating one dependent variable while considering several independent variables, enabling a more comprehensive exploration of the relationships and factors influencing the outcome of interest (Uyanik *et al.*, 2013). By extending the analysis beyond a single independent variable, MLR provides a more nuanced understanding of the complex interactions and influences involved in the studied phenomena.

The bootstrap technique involves repeatedly generating random samples, with replacement, from the original dataset to construct samples of equal size to the original one. These individual samples are known as bootstrap samples, and each contributes a value for the parameter of interest, like the mean. The expression “with replacement” signifies that any data point can be selected multiple times within each bootstrap sample. This approach is essential because sampling without replacement would lead to a random rotation of the initial data, leaving various statistics like the mean ($\bar{x}$) unaffected. While the standard error is computed using the standard deviation of the statistics derived from the bootstrap samples, conducting the procedure more extensively provides crucial insights into the variability of the estimator (Walters & Campbell, 2005).

HBA1C, also called glycated haemoglobin, plays a vital role as a biomarker in diabetes management. It offers valuable information about an individual’s average blood sugar levels over a specific duration. It is a critical tool for monitoring glycaemic control and evaluating the efficacy of diabetes treatment strategies. Stratton *et al.* (2000) conducted a study emphasizing the importance of maintaining proper blood sugar levels to reduce the likelihood of complications associated with diabetes. The research also presented evidence supporting the reliability of HBA1C as a long-term indicator of glycaemic control and its ability to predict unfavorable outcomes.

Measuring HBA1C and fasting blood sugar (FBS) levels is essential in assessing and managing glycaemic control, particularly in individuals with diabetes. HBA1C indicates long-term blood sugar levels, while FBS provides valuable insights into the immediate state of fasting glucose. Amelia & Luhulima (2020) conducted a study investigating the relationship between HBA1C and FBS levels in individuals diagnosed with prediabetes. The researchers gathered data on HBA1C levels and FBS measurements from a cohort of type 1 diabetes patients and examined the correlation between these two variables. The study findings indicated a positive association between HBA1C and FBS levels in individuals with prediabetes, demonstrating that higher HBA1C levels were associated with elevated FBS levels. Another pilot study by Chung *et al.* (2017) examines the correlation between HBA1C and FBS levels in individuals with prediabetes. The study aimed to comprehensively understand the relationship between HBA1C and FBS for the early detection of diabetes. The pilot study highlighted the association between these two biomarkers, offering valuable insights into the effectiveness of HBA1C as an indicator of glucose control in individuals at risk of developing diabetes.

The measurement of HBA1C, FBS, and creatinine levels plays a significant role in the comprehensive assessment of glycaemic control and kidney function in individuals with diabetes. These three biomarkers provide a more comprehensive understanding of diabetes management and the risk of renal complications (George *et al.*, 2020). Another study by Farasat *et al.* (2015) explored the relationship between HBA1C, FBS, and serum creatinine levels, aiming to identify the association that could indicate the risk of chronic kidney disease in individuals with impaired glucose tolerant. The study findings suggest that by considering HBA1C, FBS, and creatinine collectively, healthcare professionals can assess the likelihood of chronic kidney disease in impaired glucose tolerant patients. This comprehensive evaluation allows for early detection and intervention, facilitating the prevention or management of diabetic nephropathy.

Next, combining HBA1C with FBS, urea, and creatinine levels provides comprehensive insights for assessing glycaemic control and renal function in individuals with diabetes (Ekun *et al.*, 2022). This study also discusses the clinical implications of these associations, suggesting that incorporating all four biomarkers in routine assessments can provide valuable information for optimizing diabetes management and identifying individuals at higher risk for renal complications. In another separate study, Faheem *et al.* (2019) investigated the correlation between HBA1C and FBS, urea, and creatinine levels in patients with type 2 diabetes. The study revealed significant correlations between HBA1C and the abovementioned variables, suggesting potential kidney dysfunction or impaired renal function.

Nowadays, statistical methodologies play a crucial role in analysing and interpreting data, making predictions, and drawing meaningful conclusions. With the advancement in data analysis techniques, researchers have started exploring combining multiple statistical methodologies to develop hybrid models that can leverage the strengths of different approaches. A previous study conducted by Eğrioğlu & Fildes (2022) proposed a hybrid model that combines bootstrap resampling with MLFF for forecasting tasks. The authors aim to improve forecasting performance and robustness by integrating the bootstrap technique with neural networks. In the study, the authors also compare the performance of their hybrid model with individual neural networks and other classification algorithms. The results show that the hybrid model achieves better classification results, showcasing the benefits of combining bootstrap resampling with MLFF. Another study by Ahmad *et al.* (2016) proposes a hybrid model combining bootstrap resampling with MLR to enhance statistical inference in linear regression analysis. The study demonstrates the effectiveness of the proposed bootstrap approach in improving the performance of MLR models, particularly in terms of robustness and accuracy of parameter estimates and hypothesis testing.

Integrating statistical methodologies in a hybrid model allows researchers to use diverse techniques. By combining these various statistical methodologies, researchers can enhance their ability to model complex relationships, handle different data types, and make more accurate predictions. Another separate study conducted by Greene (2007) suggests that method integration approach enables researchers to address method limitations, leverage method strengths, and mitigate method biases. According to Creswell & Plano Clark (2011), using this approach enhances comprehension more effectively than relying solely on one research method in specific studies. However, it is still a necessary to conduct further research and development in hybrid methodologies. Therefore, the current study aims to integrate the MLFF with MLR while taking bootstrapping into account for biometry modeling. This study is expected to contribute to biometry modeling by offering a hybrid methodology with enhanced accuracy and efficiency compared to traditional standalone approaches.

1.3 Problem Statement

The complexity of medical diagnosis work has prevented the full realization of a completely automated, computer-based medical diagnostic system. However, recent advancements in intelligent systems have opened up greater possibilities for employing computers with Artificial Intelligence (AI) techniques in medical diagnostics. Given the current availability of affordable, high-speed, and efficient computers, a thoughtfully developed intelligent diagnostic system can cater to numerous patients. Despite the application of artificial intelligence in medical diagnosis, there is an ongoing quest for improved diagnostic systems (Djam & Kimbi, 2011).

Multiple factors contribute to why healthcare professionals have not completely adopted contemporary decision-support systems. One key factor is the inherent complexity and challenges within the medical field, emphasizing the need for a high degree of precision. Hence, healthcare professionals express concerns regarding the safety of decision-making tools. Additionally, many decision support systems are designed in a way that makes them challenging for users to navigate or understand, leading to a lack of user engagement (Khan *et al.*, 2000). Furthermore, healthcare professionals often believe they possess a deeper comprehension of medical concepts and are hesitant to accept guidance from computer-based decision-support systems.

However, previous studies have predominantly focused on individual model techniques, which possess limited capabilities in addressing the challenges associated with biometry modeling studies. This limitation arises from the infeasibility of relying solely on a single approach to cater to the diverse complexities encountered. Integrating hybrid techniques into the statistical models makes it possible to achieve more precise estimations for the modeling purposes at hand. Furthermore, this study adheres to the principle of parsimony, aiming to adopt the most straightforward assumptions. This research primarily explores MLR models combined with MLFF while incorporating bootstrapping techniques. The rationale behind this focus lies in the potential applicability of this integrated model for a diverse medical case study.

This research will employ a hybrid model to examine the relationship between the parameters of interest. The utilisation of multiple statistical techniques within a single model has been relatively unexplored, particularly in the context of biometry modeling studies conducted in Malaysia. The present study aims to address this gap by developing a hybrid model that combines different methodologies, resulting in a more comprehensive and versatile approach. Various metrics will be employed to evaluate the developed hybrid model's performance. These include the assessment of the Mean Square Error for Neural Network (MSE.net), the Mean Square Error for Linear Model (MSE.lm), and the accuracy testing, which involves comparing the actual values with the predicted values generated by the model.

1.4 Significance of the Study

In biometry modeling, there is a lack of specific research methodologies that focus on integrating different methods and procedures. The first study objective is to address this gap by proposing two distinct processes: integrating MLFF and MLR with case resampling and integrating MLFF and MLR without case resampling (as illustrated in Figure 1.1). The second study objective involves conducting a validation procedure on the constructed model in order to assess its precision and accuracy. Finally, as for the third objective, a comparison will be made between the two produced models, and the model exhibiting higher precision and accuracy will be considered as the superior model. Previous studies have not explored the integration of bootstrapping, MLFF, and MLR to create a new integrated model, which inspired the development of this novel methodology. Theoretically, these new statistical modeling approaches can significantly improve predictions' accuracy and precision. Moreover, this study offers a means of comparing the accuracy of the proposed models, enabling the identification of the most effective model for future applications.

1.5 Conceptual Framework of the Study

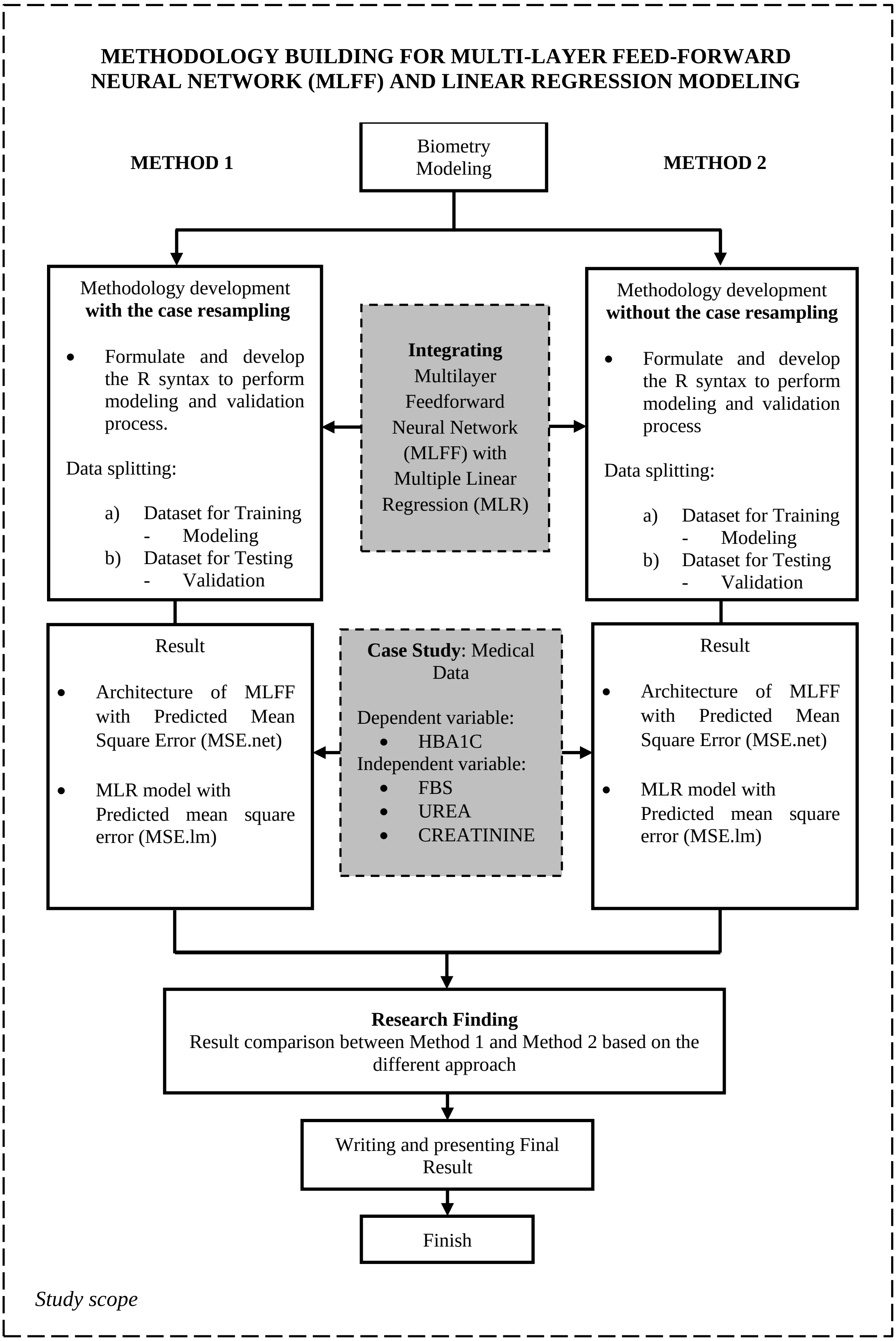
METHODOLOGY BUILDING FOR MULTI-LAYER FEED-FORWARD NEURAL NETWORK (MLFF) AND LINEAR REGRESSION MODELING
METHOD 1
Biometry Modeling
METHOD 2
Methodology development with the case resampling
Formulate and develop the R syntax to perform modeling and validation process.
Data splitting:
a) Dataset for Training
- Modeling
b) Dataset for Testing
- Validation
Integrating Multilayer Feedforward Neural Network (MLFF) with Multiple Linear Regression (MLR)
Methodology development without the case resampling
Formulate and develop the R syntax to perform modeling and validation process
Data splitting:
a) Dataset for Training
- Modeling
b) Dataset for Testing
- Validation
Result
Architecture of MLFF with Predicted Mean Square Error (MSE.net)
MLR model with Predicted mean square error (MSE.lm)
Case Study: Medical Data
Dependent variable:
HBA1C
Independent variable:
FBS
UREA
CREATININE
Result
Architecture of MLFF with Predicted Mean Square Error (MSE.net)
MLR model with Predicted mean square error (MSE.lm)
Research Finding
Result comparison between Method 1 and Method 2 based on the different approach
Writing and presenting Final Result
Finish
Study scope

Figure 1.1 Conceptual framework of the study

Figure 1.1 above shows the conceptual framework of this study. The overall process can be divided into three main phases. The first phase focuses on the development of the hybrid model. This phase also involves entering the data and naming the variables in the dataset. Next, this first phase also consists of the data-splitting process. The data are divided into train and test data which comprises 70% and 30% of the booted data, respectively. Train data is used to develop the model, while test data is used to validate the model. The second phase mainly focuses on the validation of the developed model. R syntax for MLR is applied to construct the regression model, followed by R syntax for estimating the developed regression model's mean square error (MSE). The final phase of the analysis focuses on model comparison, which starts with accessing the value of MSE.net, MSE.lm, and the actual versus predicted value. The importance of those components in Method 1 will be compared with Method 2 to determine which method provides a more accurate result. The normalised data were then divided again into train and test data, where train data is used to build the neural network's architecture. In contrast, test data is used to obtain the MSE of the MLFF model.

1.6 Research Hypotheses

There is no hypothesis involved in this study. The fundamental of this study is to focus on methodology building.

1.7 Study Objectives

1.7.1 General Objective

To develop a combination of methodology building for multi-layer feed-forward neural networks and linear modeling in biometry modeling.

1.7.2 Specific Objective

1. To develop and elucidate a new integration linear model methodology that combines Bootstrap, Multi-Layer Feed-Forward Neural Network, and Multiple Linear Regression for biometry modeling.
2. To validate the model obtained in (1).
3. To compare the accuracy and precision of the newly integrated hybrid method with and without case resampling procedure.

1.8 Scope of the Study

The main scope of this study focuses on integrating three components; data bootstrapping, MLFF, and MLR. These components serve a fundamental role in the development of the model. Given the existing gap in research regarding the utilisation of a combination of MLFF and MLR, along with the incorporation of the bootstrap technique in the medical field, this study intends to introduce a new perspective to elucidate biometry modeling using the proposed statistical approach.

In this study, the proposed approach will be implemented using two distinct methods; Method 1 and Method 2. Method 1 incorporates the bootstrapping or case resampling technique, while Method 2 focuses solely on MLFF and MLR without including the case resampling process. The evaluation of both approaches will be based on the obtained results, explicitly focusing on their accuracy and performance.

1.9 Study Contribution

This study has the potential to offer further valuable contributions to the field. One potential contribution is in the aspect of practical applications. By developing an integrated methodology applicable to biometry modeling, this study can provide valuable tools for future predictions in these domains. This can assist professionals in making informed decisions, improving investigative procedures, and enhancing overall outcomes in these critical areas.

Another potential contribution lies in the advancement of statistical techniques and methodologies. By exploring and implementing the combination of bootstrap, MLFF, and MLR methods, this study can contribute to developing and refining statistical approaches tailored explicitly for medical cases. This can pave the way for more accurate and reliable predictions, facilitating evidence-based decision-making and fostering advancements in the respective fields.

Furthermore, this research can contribute to the existing body of scientific literature by filling gaps in knowledge and understanding. Addressing the methodological challenges and complexities inherent in this study, it can serve as a reference for future researchers, providing insights and guidance in designing new methodologies or conducting further investigations. This can encourage a continuous cycle of research and innovation, leading to advancements in the biometry modeling field.

Lastly, this study's contribution extends to the scientific community and society. Proposing a superior modeling method with high predictability and accuracy can benefit various sectors and industries that rely on data analysis and predictive modeling. This can have implications in healthcare, law enforcement, and policy-making, where accurate predictions and informed decision-making are paramount.

Overall, this study can significantly contribute to the respective fields and catalyze further research and advancements through its contributions to practical applications, statistical methodologies, scientific literature, and societal impact.

1.10 Limitation of the Study

Although better methods are desired, especially in creating the combination of methodology for biometry modeling purposes, there are always limitations. This study's objective is to propose a hybrid method that can be used to model quantitative variables, testing, and validation precisely. Therefore, a limitation of the hybrid model lies in its reliance on just three core statistical techniques: bootstrapping, MLR involving quantitative variables, and the construction of MLFF. As mentioned earlier, the model's outcome will be evaluated for accuracy. Therefore, the dataset used to test the methodology will be a secondary data type. The dataset being used in this study is about health science. In other words, the dataset will not belong to various fields, which can also be considered another study limitation. Another limitation of this study is that the dataset included only one outcome variable and three predictor variables. Last but not least, the type of software used is also one of this study's limitations. All the techniques employed, assessed, or confirmed in this study will solely rely on the R-programming software. The current study aims to fill the knowledge gap to address the limited exploration of employing multiple statistical methods within the same model, particularly in the context of biometry modeling in Malaysia.

1.11 Thesis Organisation

This thesis is comprised of five comprehensive chapters and is meticulously organised to ensure coherence and clarity. The first chapter serves as a comprehensive introduction designed to provide a fundamental for the study by explaining its background and the contextual landscape within which it is conducted. In this introductory chapter, particular attention is devoted to elucidating the study's objectives, rationale, scope and methodology, thereby offering a comprehensive understanding of the research framework. Furthermore, the chapter discusses the significance of the study concerning the existing body of knowledge while also acknowledging and addressing the inherent limitations that may influence the outcomes and generalizability of the findings.

Within the second chapter of this thesis, an extensive literature review is presented, focusing on the statistical methods that will be seamlessly integrated into the study. This chapter is a valuable resource as it meticulously explores and examines the previous statistical approaches employed in biometry modeling. By thoroughly reviewing these methods, the chapter offers valuable insights into their applicability, effectiveness, and potential limitations. The comprehensive analysis and synthesis of the literature enhance the understanding of these statistical methods and provide a critical foundation for the subsequent chapters, enabling a thorough evaluation of their sustainability and efficacy in the current research context.

The third chapter of this thesis is dedicated to the methodology section, which provides an in-depth and detailed explanation of the procedures and statistical models utilised in the study. Specifically, it covers implementing Bootstrap, MLFF, and MLR models. Concisely, this chapter elucidates various aspects, such as the study design and the study's geographical location. Additionally, a visually informative flow chart depicting the proposed hybrid modeling methodology is included in this chapter, offering a clear overview of the study's sequential steps and process.

In chapter four, the discussion centers around the outcomes of the analysis conducted in this study. It begins with explaining the data preparation process, encompassing tasks such as identifying missing values and normalizing the data. Subsequently, the chapter also elaborates on the findings derived from the two developed biometry models: the first implemented with bootstrapping and the second without bootstrapping. This chapter also outlines the model evaluation procedure employed to assess the efficiency of each model. Lastly, this chapter also compares the performance of each model to determine the model that yields the most accurate results.

In chapter five, a comprehensive elaboration of the analysis outcomes is presented, encompassing the insights derived from both models developed within this study. This chapter also engages in a discussion to determine the superior model, guided by the model evaluation criteria. Meanwhile, chapter six also summarizes this study based on the findings from the analysis. This chapter emphasizes the study recommendation and future direction of this research and discusses the consideration for future research to improve the understanding of biometry modeling.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter gives an insight into the previous study applying statistical modeling methods in analysing dyslipidemia and diabetic disease. The second section discusses the history of statistics. The third section in this chapter provides an overview of bootstrap, covering its history and the underlying concept. The fourth section briefly explains the overview of regression, including its development throughout history. In the fifth section, past studies that implement neural network methodology and its application in biostatistics are highlighted. The sixth section of this chapter explains the rationale for deploying a hybrid method over a singular modeling technique. The seventh section highlights the type of relation between statistical models and how each model is combined into a single hybrid model. Section eight provides a general overview of hybrid techniques, specifically the combination of the three main statistical methods employed in this study. Meanwhile, section nine explores the utilization of statistical methodology in the medical domain, providing insight into the relevant factors to this study. Additionally, section ten delves into the validation while comparison of statistical models is being explained in the eleventh section of this chapter. Section twelve focuses on the methodology review and lastly, section thirteen concludes this chapter by summarizing the key insights gained from the review of literature.

2.2 Early History of Statistics

The concept of statistics emerged during the late 18th century in the German field of nation-state studies known as *Statistik* or *Staatenkunde*, as noted by Porter

(1985). While states had been gathering data since ancient times, it was in the early 19th century that a growing fascination with numerical data became prominent in Continental Europe. Hacking (1990) describes this period as marked by a significant rise in the publication of numerical data and the creation of detailed tables, enabled by the increased use of durable, transferable records. These figures and tabulations revealed a startling consistency in various social statistics: murders, suicides, and even the quantity of dead letters in the main post office in Paris stayed consistent despite all the changes around them.

The British Association for the Advancement of Science did not entirely support the founding of the Statistical Society of London, subsequently known as the Royal Statistical Society, which took place in 1834 (Poovey, 1993). For society members during this formative time, only numbers were considered significant. It was only necessary to gather information; no analysis or inferences were to be made. However, it was unavoidable not to do so. The inscription of numbers, followed by their incorporation into probability calculations within an expanding set of commercial and statist networks, produced worlds to be organised, controlled, and manipulated. There are a few key elements that need to be emphasised here.

First, statistics was a way of thinking that fit the times, given the rise of industrial capitalism and the nation-state. According to Williams (1982), capitalism’s expanding complexity, vastness, and unmanageability was the most significant development. It is remarkable that the statistical model of analysis, a classic method developed in response to the difficulty of understanding modern society through experience, had its exact beginnings in the 19th century. It is because a community growing from the industrial revolution could not be known in any meaningful sense without integrating statistical theory and mechanisms for collecting statistical data (Williams, 1982). Hence, there exists a fundamental aspect associated with capitalist industrialization and its interaction with the state, which acted as the catalyst for network development. Employing statistical records, these capitalists managed to sustain and grow their enterprises. As indicated by Marx (2006), capitalism introduced the concept of double-sided bookkeeping, while the census stands out as one of the state's most notable innovations. The meticulous record-keeping facilitated the transformation of workshops into global manufacturing centers and enabled the British Foreign Office to assert control over half of the globe.

Secondly, a brand-new world that did not exist before is created just by mobilizing large numbers. According to Asad (1994), statistics is not just a method of representing social life but also a method of shaping it. This is true in several distinct ways. First, there is the establishment and management of an institutional framework tasked with gathering and analysing the numbers. In addition, related computation technology has been developed, including conceptual techniques like correlation and regression analysis and practical ones like early calculators and punch cards (Hacking, 1990). Third, new categories for organising the obtained data provide the foundation for comprehending and interacting with the world. Last, there is the new reality of statistical measurements themselves. With the same conceptual resonance as dead birds and wood tables, terms such as the mean, standard deviation, and later correlation coefficient become things in their own right. To sum up, the avalanche of numbers produced a new world in the sense that things could be uttered and words written down that were not only not understood previously but could not even be dreamed of (Hacking, 1992).

Thirdly, the introduction of statistical discourse brought about a shift in how individuals perceived themselves within society. When society is viewed through the lens of statistics, individuals assess themselves and are assessed by others in relation to the norms of the collective. As Hacking (1990) articulates, individuals are considered normal when their characteristics align with the central tendencies of socially derived statistical aggregates. In contrast, those falling at the extremes are deemed abnormal or pathological, a characterization few aspire to. Consequently, most individuals strive to conform to these norms, thereby influencing the definition of what is considered normal. The pivotal moment in this transformation was the recognition of statistical regularities, particularly the application of the probability distribution known as the "error curve" or, in contemporary terms, the normal distribution. This concept, originally developed by Abraham De Moivre and further advanced by Pierre-Simon Laplace, gained prominence through Carl Gauss's work in astronomy and geodetic observations. Belgian astronomer and social statistician Adolphe Quetelet extended this idea to human populations, as exemplified in his analysis of chest measurements from over 5,000 Scottish Highland soldiers. This led to the conception of "*l'homme moyen*" or the 'average man' which went beyond a mere arithmetic mean and acquired moral implications. Quetelet's reliance on the error curve language and its origins in astronomical measurement implied that anyone deviating from this average was, in one way or another, considered an error.

Fourthly, after a large population has achieved normality, there is a clear mandate for administrative oversight and intervention to maintain it. In other words, the importance placed on numbers was merely a surface impact. There were new technologies for cataloging and numbering things and new bureaucracies with the authority and continuity to put those technologies into use (Hacking, 1990). However, this intervention could have repercussions in both directions. At "the bad end" of the distribution, there may be an attempt to regulate the abnormal by removing the most extreme cases or bringing them closer to the norm through reformation and education. Similarly, there may be an effort to shift the standard towards “the good end” of the distribution.

Fifthly, the use of statistics is a compelling piece of rhetoric in the process of gathering allies. Numbers were the connecting factor between the many divisions of government, branches of the state, numerous institutions, and quasi-professional organisations such as the London Statistical Society. It is also fascinating to see how statistics bridges the gap that has traditionally existed between the natural sciences and the social sciences. Mirowski (1994) described this phenomenon as a spiral that oscillates back and forth between historically contingent locations of natural science and social science and wobbles as the poles alter. For instance, the probability distribution that Gauss adopted to assess the precision of astronomical observations was later projected by Quetelet back onto society as the normal distribution. Later on, James Clark Maxwell subsequently adopted to derive his gas laws (Porter, 1981), which Ysidro Edgeworth then applied to describe the reliability of the market in the face of uncertainty, and so forth (Mirowski, 1994). In each instance, something foreign transforms into something recognisable, bringing together new adherents and allies. It is feasible to begin to see the importance of some of the ideas stated by Bloor and Latour by using this story as a starting point. From Bloor’s point of view, the study of statistics does not involve unfolding an innate mathematical logic. On the other hand, statistics developed into a highly relevant field of study as a direct result of the interests of individuals who studied it.

2.3 Overview of Bootstrapping

The bootstrap method, which Efron developed in 1993, uses a resampling approach (Efron & Tibshirani, 1993). The idea behind bootstrap is to utilise a sample as a population to take an example of the model in hand and then create many "case resampling samples", known as bootstrap samples. The bootstrap procedure begins with the original sample drawn from the population of interest. The next step is replicating the original sample multiple times to create a new population while keeping the old one in mind. The bootstrap selects many samples that are then replaced by a random sampling technique, resulting in a fresh sample from the beginning.

Bootstrap is the most simple and direct method because it does not require the complex computations of derivatives and Hessian matrix inversion required by linear methods or the Monte Carlo solutions of integrals needed by the Bayesian approach (Dybowski & Roberts, 2001). The bootstrap method has many applications, including estimating means, confidence intervals, parameter uncertainties, and network design techniques. (Lall & Sharma, 1996). The bootstrap method has also been used to develop artificial neural network models. Abrahart (2003) continuously applied the bootstrap technique to sample the input space in rainfall-runoff modeling and reported that it improved significantly in greater precision and enhanced generalisation. Jia and Culver (2006) used the bootstrap technique to estimate the generalisation errors of neural networks with various structures and build confidence intervals for synthetic flow prediction with a small data sample.

The bootstrap technique builds new distributions and stores new data sets for later analysis. The advantage of bootstrapping lies in its ability to expand the sample size to match the original size, facilitating the retention of specific observations while discarding others. As Chong *et al.* (2011) noted, bootstrapping offers the advantage of not necessitating assumptions about data distribution or specialized statistical expertise. However, some skeptics question the accuracy of estimates produced by resampling methods because they repeatedly employ the same values from the sample. The argument revolves around the limitations imposed by the dataset's size and its representativeness of the population (Chong *et al.*, 2011).

2.3.1 History of Bootstrapping

American statistician Bradley Efron developed bootstrapping in the 20th century (Efron, 1979). This approach assumes that the sample and population have the same relationship to an empirical distribution produced by resampling *N* samples of the same size from the original distribution with replacement. The researcher can assess the precision of the inferences on the population parameter by developing this empirical distribution and contrasting it to the sample statistic. Bootstrapping has grown in popularity over the past forty years and encompasses several variations, including parametric and Bayesian bootstrapping.

Initially, bootstrapping evolved to develop and enhance a previously created technique known as jackknife resampling (Efron, 1979). British statistician Maurice Quenouille first proposed the jackknife method in his paper, “Problems in Plane Sampling”, in 1949. Quenouille provided expressions for the precision of calculating the linear sampling error and sampling error in a systematic and stratified sampling of an area in this paper. In the context of math and science advancement in 1959, when the United States was engaged in a Cold War rivalry with the Soviet Union, John Tukey, an American mathematician, developed these expressions. The phrase was later referred to as the Quenouille-Tukey jackknife. Tukey gave the technique the nickname “jackknife” to allegorise the robustness of the statistical tool, referring to the typical folding knife that men at that time used but did not consider an ideal tool (Champkin, 2010).

With the Quenouille-Tukey jackknife resampling method, the original sample is taken, one of the observations is left out, and a new sample is used to calculate the desired statistic. There would be *n* sample statistics based on a sample size of (*n*-1) after methodically excluding each observation one at a time and computing the statistic. Efron's bootstrap sampling method uses a single sample to generate the entire sampling distribution. The next step of the jackknife method was to find the centre of this sampling distribution by averaging the *n*-sample statistics (Quenouille, 1949). Similar to a point estimate of the parameter, the original sample's statistic can be used to estimate the parameter. However, unlike a point estimate, the jackknife method provides a measure of the accuracy and validity of this estimate by supplying the necessary tools to evaluate these three assumptions. The first assumption is that the parameter's sampling distribution is normal. The second assumption is that the estimated parameter's standard error is close to the estimated parameter's sampling distribution standard deviation. Meanwhile, the third assumption is that the estimated parameter has a slight bias in its estimation.

The jackknife method allows the evaluation of normality assumption either visually or statistically using tests like the Shapiro-Francia normality test, which gives a view of the sampling distribution. By comparing the centre of the sampling distribution, the jackknife method also provides an estimate of the amount of bias in the estimated parameter. In addition, comparing each sample statistic in the generated sampling distribution to the previously calculated mean can determine the standard error of the parameter estimation (Quenouille, 1949). Understanding how well these three assumptions are satisfied, the jackknife method can clarify that a confidence interval for a complicated parameter is valid.

$$\hat{\theta}_{\mu} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{i}$$

$$Bias = (n-1)(\ \hat{\theta}_{\mu} - \hat{\theta})$$

$$SE\ (\hat{\theta}) = \left\{\frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_{i} - \hat{\theta}_{\mu})^{2}\right\}^{\frac{1}{2}}$$

The above method quickly attracted the interest of statisticians due to its capability to support inferences made on parameters, particularly on nonparametric parameters that had previously presented a challenge to statisticians. Rupert Miller, a professor at Stanford University, was one significant statistician who showed interest. Miller investigated the jackknifing method and published numerous papers to identify and address its problems. Bradley Efron, a PhD student of Miller's who developed the bootstrapping techniques, was probably influenced by this research regarding his career. In fact, after earning his doctorate and spending a few years at Stanford, Efron took a leave of absence and travelled to Imperial College, where Miller delivered a lecture centred on his 1964 paper on the technique of jackknifing (Holmes, 2003). Then, with some prodding from a coworker, Efron started researching the jackknife technique. His allusions to Miller's earlier works, particularly "The Jackknife: A Review", which attempted to summarise all of the research and findings on the jackknife method from its inception through 1974, show the influence of Miller on Efron in this early research (Efron & Stein, 1981). Over the subsequent years, Efron worked on creating a technique that would achieve the same result but be less systematic and more randomised. He published a paper on bootstrap methods in January 1979, arguing that they would be more reliable and applicable than jackknife methods (Efron, 1979). Jackknifing is a