UNIVERSITI TEKNOLOGI MARA

# MACHINE LEARNING AND PENALIZED REGRESSION MODELS FOR HIGH DIMENSIONAL DATA ANALYSIS ON MULTI OMICS BLOOD-BASED BIOMARKERS FOR ALZHEIMER'S DISEASE

**MOHAMMAD NASIR BIN ABDULLAH**

Thesis submitted in fulfilment
of the requirement for degree of
**Doctor of Philosophy**
**(Statistics)**

**Faculty of Computer and Mathematical Sciences**

**February 2022**

# ABSTRACT

Alzheimer's disease (AD) is a neurodegenerative disorder that can be characterised by the gradual progression of memory loss, impairment of cognitive function, and progressive disability. Currently, there are no treatments available for AD. Detection of biomarkers would assist in early prediction of AD. Prevalent technologies of high throughput in genomics have opened a new horizon to achieve this purpose. Transcriptomics, metabolomics, and cytokinomics data are among the multi-omics data used in modern genomics studies. Multi-omics data is high dimensional data where the number of dimensions is larger than the number of sample observations. Additionally, due to the large number of features, there are issues of multicollinearity and complete separation. Therefore, questions have been raised about how to analyse multi-omics high dimensional data and what the best classifiers for the classification of AD are. The main objective of this study was to establish an algorithm for AD classification using multi-omics data and to discover the potential blood-based biomarkers of AD. This study utilized three sets of real omics data (transcriptomics: $n_{AD}: 92, n_{non-AD}: 92, p = 22{,}254$; metabolomics: $n_{AD}: 55, n_{non-AD}: 55, p = 100$; cytokinomics: $n_{AD}: 39, n_{non-AD}: 39, p = 13$). In the first phase of the study, an algorithm was developed to simulate synthetic data with correlated features using transcriptomics real data by controlling experimental and biological variation for different sample sizes ($n$ = 10-50 (increment of 10), $n$ = 100-500 (increment of 100)) and the number of features ($p$ = 100, 200, 300, 400, and 500). Only continuous features were generated, and the target variable is binary type (1,0). The second phase involves the evaluation of machine learning (ML) classifiers (support vector machine with different kernels, random forest, Naïve Bayes, and k-NN) and penalized logistic regression models (lasso, ridge, and elastic net logistic regression) using synthetic data. The classifier performance measures were sensitivity, specificity, accuracy, error rate, and F-measure. Simulation results with 500 replications showed that Naïve Bayes performed well for high dimensional data ($p > n$) while the support vector machine performed well for classifying low dimensional data ($n > p$). The results also showed that lasso and ridge logistic regression performed well for high dimensional data. In the third phase, machine learning classifiers were evaluated for each omics dataset, and the potential blood-based biomarkers from the individual-omics dataset were then identified using the best ML model. This study successfully identified 16 transcriptomics biomarkers, 14 metabolomics biomarkers, and nine (9) cytokinomics biomarkers from the individual-omics dataset. The identified transcriptomics, metabolomics, and cytokinomics biomarkers were then combined to form an integrated multi-omics dataset. In the final phase of this study, an algorithm for AD prediction model was established using principal component analysis and Firth logistic regression based on the integrated multi-omics dataset, which has issues of complete separation and multicollinearity. Based on the established AD prediction model, 18 potential biomarkers (two transcripts, seven metabolites, and nine cytokines) for AD were discovered. This study also produced a guideline for handling high-dimensional datasets that have issues of multicollinearity and complete separation.

# ACKNOWLEDGEMENT

In the name of Allah, the Most Gracious, the Most Merciful.

*"Then High above all be Allah, the True King. And be not in haste with the Qur'an before its revelation is complete to you and say: "My Lord! Increase me in knowledge.""*

*The Holy Qur'an (20:114)*

Praise to Allah S.W.T., the Most Compassionate and Most Merciful, whose blessings have helped me through the entire completion of this thesis. I would like to record my deepest gratitude and appreciation to the following individuals, who have helped me during the preparation of this thesis and in my pursuit of the coveted Doctor of Philosophy (Statistics) at the Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA.

1. My Supervisor, ***Professor Dr Yap Bee Wah*** for her support, encouragement, invaluable advice, comment, suggestion, and contribution of her time throughout the research and completion of this thesis.

2. ***Professor Dato' Dr Abu Bakar Abdul Majeed*** and ***Dr Yuslina Zakaria*** as my co-supervisor for their support, suggestion, comments, and commitment.

3. ***Professor Dr Syed Hatim Noor @ Nyi Nyi Naing***, and ***Associate Professor Dr Kamarul Imran Musa*** for their encouragement, inspiration, and support in teaching and awakening the interest in me for the study of statistics.

4. My deepest gratitude to my lovely parents (***Aishah Seeta Devi Abdullah*** and ***Abdullah Marican***), my wife (***Nor Hazlin Ramli***), my sons (***Amsyar Arsyad Mohammad Nasir***, ***Aydin Aariz Mohammad Nasir,*** *and* ***Afiq Ahsan Mohammad Nasir***), and siblings for their endless patience, tolerance, great supports, and time spared for the success of this thesis.

5. LRGS grant team members who collected the multi-omics dataset, especially ***Dr Ainon Zahariah Samsudin, Che Nor Adlia Enche Ady, Dayana Sazereen Md Hasni, Associate Professor Dr Kalavathy Ramasamy, and Dr Nazif*** from Brain Degeneration and Therapeutics Group, Faculty of Pharmacy, UiTM.

6. All my good friends for their moral support, opinions, and sharing of knowledge throughout this research especially ***Dr Mohamed Imran Mohamed Ariff*** and ***Imran Md Jelas***.

7. Finally, to all others who were directly or indirectly involved in this study.

Pray Allah s.w.t. rewards all of you bountifully for all your support.

# TABLE OF CONTENTS

# CHAPTER ONE
# INTRODUCTION

## 1.1 Background of Study

This section covers some background knowledge about Alzheimer's disease (AD), genomic studies on AD, the issues of the high dimensionality of omics data and classification using machine learning classifiers. Alzheimer's disease (AD) is a neurodegenerative disorder characterized by the gradual progression of memory loss, impairment of cognitive function, and progressive disability (Doecke et al., 2012; Fuente-Ferna´ndez, 2011; Lv et al., 2014; Motta et al., 2007; Tezel et al., 2019). In addition, the foundation of AD is related to a gradual accretion of amyloid-beta (Aβ) peptides in the brain, forming senile plaque and inducing synaptic loss (Hadjichrysanthou et al., 2018). Thus, AD does not appear suddenly; instead, it is a progressive disorder with a continuing asymptomatic phase followed by a symptomatic pre-dementia phase and finally by the dementia stage (Parnetti et al., 2019). AD is the most common disease amongst the elderly aged over 60 years, and it contributes to approximately 60% to 80% of dementia cases. Dementia is a disorder characterised by cognitive decline and impairment in learning, memory, language, attention, perceptual-motor skills, and social cognition (Tezel et al., 2019). Clinical indicators of AD-related to dementia can be more devastating, where the sufferer could experience swallowing disorder, aspiration pneumonia, and behavioural symptoms such as agitation, hallucinations, and aggression. The impact of AD is substantial on global health care due to its significant economic and social repercussions, not limited to the sufferer but also the caregivers and family members (Martí-Juan et al., 2019). All the symptoms might lead to caregiver stress, burnout and medical illness (Reuben et al., 2019). Caregivers are defined as people responsible for taking care of AD patients, providing help with their daily living activities, a role that might require permanent commitment by the caregiver.

In Malaysia, it is estimated that the population will increase from 28.6 million in 2010 to 41.5 million in 2040. Thus, the elderly aged 65 and above will increase to 14.5% of the total population in 2040. In low- and middle-income nations like Malaysia, these figures are a top focus for public health (Kan et al., 2019). The estimated number