

**POLYNOMIALS FEATURE-TRANSFORMED HEAP  
DIMENSIONALITY REDUCTION AND STACKING  
ENSEMBLE FOR SPECTROMETRY DATA  
CLASSIFICATION**



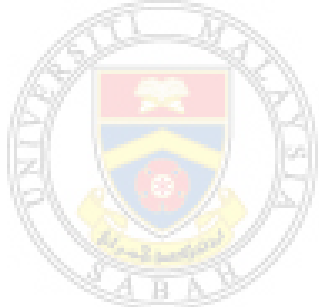
**NUR HASSHIMA BINTI HASBI**

**UMS**  
UNIVERSITI MALAYSIA SABAH

**FACULTY OF SCIENCE AND NATURAL RESOURCES  
UNIVERSITI MALAYSIA SABAH  
2023**

**POLYNOMIALS FEATURE-TRANSFORMED HEAP  
DIMENSIONALITY REDUCTION AND STACKING  
ENSEMBLE FOR SPECTROMETRY DATA  
CLASSIFICATION**

**NUR HASSHIMA BINTI HASBI**



**UMS**  
UNIVERSITI MALAYSIA SABAH

**THESIS SUBMITTED IN FULFILMENT OF  
THE REQUIREMENTS FOR THE DEGREE OF  
MASTER OF SCIENCE**

**FACULTY OF SCIENCE AND NATURAL RESOURCES  
UNIVERSITI MALAYSIA SABAH  
2023**

**UNIVERSITI MALAYSIA SABAH**

**BORANG PENGESAHAN STATUS TESIS**

JUDUL : **POLYNOMIALS FEATURE-TRANSFORMED HEAP DIMENSIONALITY REDUCTION AND STACKING ENSEMBLE FOR SPECTROMETRY DATA**

IJAZAH : **SARJANA SAINS**

BIDANG : **MATEMATIK DENGAN GRAFIK BERKOMPUTER**

Saya **NUR HASSHIMA BINTI HASBI**, Sesi **2020-2023**, mengaku membenarkan tesis Sarjana ini disimpan di Perpustakaan Universiti Malaysia Sabah dengan syarat-syarat kegunaan seperti berikut:-

1. Tesis ini adalah hak milik Universiti Malaysia Sabah
2. Perpustakaan Universiti Malaysia Sabah dibenarkan membuat salinan untuk tujuan pengajian sahaja.
3. Perpustakaan dibenarkan membuat salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. Sila tandakan ( / ):

SULIT

(Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA 1972)

TERHAD


(Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi/badan di mana penyelidikan dijalankan)

TIDAK TERHAD

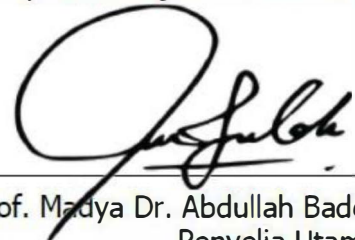
Disahkan Oleh,



**NUR HASSHIMA BINTI HASBI**  
**MS2011014T**

 ANITA BINTI ARSAD  
PUSTAKAWAN KANAN  
UNIVERSITI MALAYSIA SABAH

(Tandatangan Pustakawan)



(Prof. Madya Dr. Abdullah Bade)  
Penyelia Utama

Tarikh : 11 OGOS 2023

## DECLARATION

I hereby declare that the material in this thesis is my own except for quotations, excerpts, equations, summaries, and references, which have been duly acknowledged.

12<sup>th</sup> May 2023



Nur Hashhima binti Hasbi

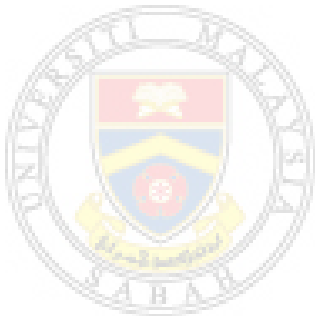
MS2011014T



UMMS  
UNIVERSITI MALAYSIA SABAH

## CERTIFICATION

NAME : **NUR HASSHIMA BINTI HASBI**  
MATRIC NUM. : **MS2011014T**  
TITLE : **POLYNOMIALS FEATURE-TRANSFORMED HEAP  
DIMENSIONALITY REDUCTION AND STACKING  
ENSEMBLE FOR SPECTROMETRY DATA  
CLASSIFICATION**  
DEGREE : **MASTER OF SCIENCE**  
FIELD : **MATHEMATICS WITH COMPUTER GRAPHICS**  
VIVA DATE : **12 MAY 2023**



CERTIFIED BY;

Signature

**1. MAIN SUPERVISOR**

Assoc. Prof. Dr. Abdullah Bade

A handwritten signature in black ink, appearing to read 'Abdullah Bade', is written over a horizontal line.

**2. CO-SUPERVISOR**

Assoc. Prof. Dr. Chee Fuei Pien

A handwritten signature in black ink, appearing to read 'Chee Fuei Pien', is written over a horizontal line.

## **ACKNOWLEDGEMENT**

First and foremost, I would like to begin by expressing my utmost appreciation to both of my supervisors, Assoc. Prof. Dr Abdullah Bade and Assoc. Prof. Dr Chee Fuei Pien. Their unwavering guidance, invaluable insights, and continuous support have played a pivotal role in shaping the successful completion of this project. Their expertise and dedication have been instrumental in steering me in the right direction throughout the research process, enabling me to meet the project's timeline and objectives.

I am also deeply grateful to my family for their unwavering support and encouragement. Their belief in my abilities and their constant encouragement have been a source of motivation and strength throughout my educational journey. Their understanding and sacrifices have allowed me to fully dedicate myself to this endeavor, and for that, I am truly grateful.

Furthermore, I extend my heartfelt appreciation to all my friends who have provided assistance, encouragement, and support along this remarkable journey. Their presence and uplifting words have been a source of inspiration, making the challenging moments more manageable and the achievements more rewarding.

Lastly, I would like to acknowledge the contributions of all the individuals who have directly or indirectly played a role in this research project. Their collaboration, feedback, and assistance have enriched my work and have contributed to its overall success.

Nur Hasshima binti Hasbi

12<sup>th</sup> May 2023

## ABSTRACT

Pattern recognition has emerged as a burgeoning field of study with increasing prominence in light of technological advancements, finding applications across various multidisciplinary domains. An essential part of pattern recognition is classification where it involves the categorization of labelled samples based on their data features. Fourier Transform Infrared (FTIR) spectroscopy, a well-established spectroscopic technique, have long been used to detect organic, polymeric, and even inorganic materials. This research endeavours to develop an accurate and optimal classification framework on FTIR spectra data using a combination of heap dimensionality reduction (DR) technique, polynomial features transformation and a heuristic stacking ensemble technique. The high-dimensionality nature of FTIR data poses a significant challenge for classification. To address this issue, DR techniques are used. However, no DR technique is superior to all others. Depending on the dataset used, one method may produce a better approximation of a dataset than the other techniques. In this study, the high-dimensional data undergo multiple existing DR techniques. The resulting transformed features are consolidated into a heap and subsequently undergo polynomial feature transformation. Then Partial Least Square (PLS-DA) method is applied to obtain the final transformed features. The transformed features are then utilized as input for the stacking ensemble (SE) model, selected through a heuristic SE procedure. Artificial data was employed for the initial two experiments, while the complete framework was tested on the six FTIR datasets for the third experiment to assess its applicability to real-world datasets. The experimental results on these six datasets revealed that the proposed framework was outperformed the other examined models. Notably, an average accuracy, sensitivity, and specificity of up to 99% was achieved for the D06 dataset. As a result, this framework holds potential not only for the classification of FTIR data but also for other high-dimensional data in general.

## **ABSTRAK**

### **PENGURANGAN DIMENSI HIMPUNAN TRANSFORMASI CIRI POLINOMIAL DAN TEKNIK PENYATUAN BERTINGKAT UNTUK PENGELASAN DATA SPEKTROMETRI**

*Sejajar dengan kemajuan teknologi, pengecaman corak telah pesat berkembang sebagai satu keperluan dan kepentingan dalam pelbagai bidang disiplin dan aplikasi. Pengecaman corak adalah pengelasan yang melibatkan pengkategorian sampel yang berlabel berdasarkan ciri-ciri data. Teknik spektroskopi Transformasi Fourier Inframerah (FTIR), merupakan satu teknik spektroskopi yang telah terbukti berkesan dan telah lama digunakan dalam mengesan bahan organik, polimer, dan juga bahan bukan organik. Kajian ini bertujuan untuk membangunkan satu rangka kerja pengelasan yang tepat dan optimum untuk data spektra FTIR dengan menggunakan teknik pengurangan dimensi timbunan, transformasi ciri polinomial dengan teknik penyatuan bertingkat bersifat heuristik. Data FTIR mempunyai dimensi yang tinggi dan telah menjadi satu cabaran dalam pengelasan dan dapat diselesaikan dengan menggunakan teknik pengurangan dimensi. Walau bagaimanapun, tiada teknik pengurangan dimensi yang lebih baik daripada yang lain kerana bergantung pada set data yang digunakan dan satu kaedah mungkin menghasilkan anggaran set data yang lebih baik daripada teknik lain. Dalam kajian ini, satu rangka yang optimum untuk mengelas data berdimensi tinggi terutamanya data FTIR telah didirikan menggunakan teknik pengurangan dimensi timbunan berserta dengan teknik penyatuan bertingkat bersifat heuristik. Data yang dikaji telah menjalani pelbagai teknik pengurangan dimensi sedia ada. Ciri yang diubah hasilnya digabungkan untuk membentuk satu timbunan ciri dan seterusnya diubah kepada ciri polinomial. Kaedah 'Partial Least Square-Discriminant Analysis (PLS-DA)' seterusnya digunakan untuk mendapatkan ciri baru. Setelah mendapatkan ciri yang baru, ciri tersebut digunakan sebagai input pada teknik penyatuan bertingkat (SE), yang dipilih melalui prosedur penyatuan bertingkat bersifat heuristik. Data tiruan yang berdimensi tinggi telah digunakan untuk eksperimen pertama dan kedua, manakala untuk eksperimen ketiga, rangka kerja yang lengkap telah diuji pada enam set data FTIR untuk menentukan kebolegunaan rangka kerja tersebut pada set data dunia nyata. Hasil analisis menunjukkan bahawa rangka kerja yang dicadangkan dapat mengatasi model lain dengan hasil tertinggi purata ketepatan, purata kepekaan dan purata kekhususan sehingga 99% terutamanya untuk dataset D06. Dengan ini, rangka kerja ini memiliki potensi bukan sahaja untuk pengelasan data FTIR, tetapi juga untuk data berdimensi tinggi secara umum.*



# LIST OF CONTENTS

|                                      | Page |
|--------------------------------------|------|
| <b>TITLE</b>                         | i    |
| <b>DECLARATION</b>                   | ii   |
| <b>CERTIFICATION</b>                 | iii  |
| <b>ACKNOWLEDGEMENT</b>               | iv   |
| <b>ABSTRACT</b>                      | v    |
| <b><i>ABSTRAK</i></b>                | vi   |
| <b>LIST OF CONTENTS</b>              | vii  |
| <b>LIST OF TABLES</b>                | xi   |
| <b>LIST OF FIGURES</b>               | xiii |
| <b>LIST OF ABBREVIATIONS</b>         | xv   |
| <b>LIST OF APPENDICES</b>            | xvi  |
| <b>CHAPTER 1: INTRODUCTION</b>       |      |
| 1.1 Overview                         | 1    |
| 1.2 Problem Background               | 3    |
| 1.3 Problem Statement                | 8    |
| 1.4 Aim                              | 9    |
| 1.5 Objective                        | 9    |
| 1.6 Scope and Limitation             | 9    |
| 1.6.1 Type of Data Used              | 9    |
| 1.6.2 Type of Classification Employs | 10   |
| 1.7 Justification                    | 10   |
| 1.8 Thesis Organization              | 12   |
| <b>CHAPTER 2: LITERATURE REVIEW</b>  |      |
| 2.1 Pattern Recognition              | 13   |
| 2.2 Overview of Machine Learning     | 15   |

|       |   |    |
|-------|---|----|
| 2.3   | Machine Learning Paradigm                         | 16 |
| 2.3.1 | Supervised Learning                               | 16 |
| 2.3.2 | Unsupervised Learning                             | 17 |
| 2.4   | Data Resampling and Preprocessing                 | 18 |
| 2.4.1 | Synthetic Minority Oversampling Technique (SMOTE) | 18 |
| 2.4.2 | Cross Validation                                  | 19 |
| 2.4.3 | Polynomials Features Transformation               | 19 |
| 2.5   | Dimensionality Reduction                          | 20 |
| 2.5.1 | Principal Component Analysis                      | 22 |
| 2.5.2 | Partial Least Square – Discriminant Analysis      | 24 |
| 2.6   | Ensemble Learning                                 | 25 |
| 2.6.1 | Bagging   | 26 |
| 2.6.2 | Boosting  | 27 |
| 2.6.3 | Stacking  | 27 |
| 2.7   | Classification Task in Machine Learning           | 28 |
| 2.7.1 | Linear Discriminant Analysis                      | 29 |
| 2.7.2 | Logistic Regression                               | 32 |
| 2.7.3 | Multilayer Perceptron                             | 33 |
| 2.7.4 | Support Vector Machine                            | 34 |
| 2.7.5 | Decision Tree                                     | 35 |
| 2.7.6 | Random Forest                                     | 36 |
| 2.7.7 | Gradient Boost                                    | 38 |
| 2.7.8 | XGBoost   | 38 |
| 2.8   | Fourier Transform Infrared (FTIR)                 | 39 |
| 2.9   | Pattern Recognition for Spectroscopy Data         | 42 |
| 2.10  | Discussion  | 53 |

### **CHAPTER 3: METHODOLOGY**

|       |                                   |    |
|-------|-----------------------------------|----|
| 3.1   | Introduction                      | 54 |
| 3.2   | Study Approach                    | 54 |
| 3.3   | System Architecture               | 56 |
| 3.3.1 | Data Acquisition and Descriptions | 57 |
| 3.3.2 | Data pre-processing and sampling  | 60 |

|       |                              |    |
|-------|------------------------------|----|
| 3.3.3 | Dimensionality Reduction     | 64 |
| 3.3.4 | Model selection and training | 64 |
| 3.3.5 | Classification process       | 65 |
| 3.4   | Evaluation Metrics           | 65 |
| 3.5   | Experiment Settings          | 67 |
| 3.6   | Software                     | 68 |
| 3.7   | Summary                      | 69 |

#### **CHAPTER 4: DIMENSIONALITY REDUCTION WITH HDR-PO T TECHNIQUE**

|       |  |    |
|-------|--|----|
| 4.1   | Introduction   | 70 |
| 4.2   | Components of HDR-PoT technique                              | 70 |
| 4.3   | Experimental Setup   | 74 |
| 4.4   | Result and Discussion  | 74 |
| 4.4.1 | Effect of HDR-PoT Method with Different Number of Dimensions | 75 |
| 4.4.2 | Effect of HDR-PoT Method with Different Sample Size          | 82 |
| 4.5   | Discussion   | 89 |
| 4.6   | Summary  | 90 |

#### **CHAPTER 5: HEURISTIC STACKING ENSEMBLE PROCEDURE**

|     |                           |     |
|-----|---------------------------|-----|
| 5.1 | Introduction              | 92  |
| 5.2 | Component of Heuristic SE | 93  |
| 5.3 | Experimental Setup        | 99  |
| 5.4 | Result and Discussion     | 100 |
| 5.5 | Summary                   | 104 |

#### **CHAPTER 6: FTIR SPECTRA CLASSIFICATION FRAMEWORK**

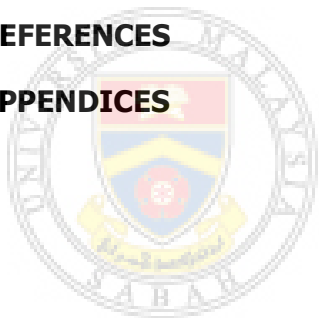
|     |  |     |
|-----|--|-----|
| 6.1 | Introduction   | 105 |
| 6.2 | Accurate and Optimized FTIR Spectra Classification Framework | 105 |
| 6.3 | Data Exploration   | 106 |
| 6.4 | Overall Comparison of Classification Performance             | 111 |
| 6.5 | Summary  | 126 |

## **CHAPTER 7: CONCLUSION**

|       |  |     |
|-------|--|-----|
| 7.1   | Conclusion   | 128 |
| 7.2   | Research Contribution  | 129 |
| 7.2.1 | Introduction of a New DR Technique   | 129 |
| 7.2.2 | Application of Heuristic SE procedure for the best SE model selection in the classification of high-dimension data | 130 |
| 7.2.3 | Establishments of stable FTIR spectra classification frameworks  | 130 |
| 7.3   | Future Works   | 130 |
| 7.3.1 | Varying the Base DR Technique in HDR-PoT   | 130 |
| 7.3.2 | Experimentation on an Imbalanced Dataset   | 131 |
| 7.3.3 | Discovering a More Effective Method to Get the Best Combination for SE Method                                      | 131 |
| 7.3.4 | Determine the Exact Number of Features Required  | 132 |

|                   |     |
|-------------------|-----|
| <b>REFERENCES</b> | 133 |
|-------------------|-----|

|                   |     |
|-------------------|-----|
| <b>APPENDICES</b> | 153 |
|-------------------|-----|



**UMS**  
UNIVERSITI MALAYSIA SABAH

## LIST OF TABLES

|   | Pages |
|---|-------|
| Table 2.1 : Best-known approaches for pattern recognition   | 14    |
| Table 2.2 : Example of kernels used in SVM and its formula  | 35    |
| Table 2.3 : Summary of selected studies on the pattern recognition of FTIR spectra data   | 42    |
| Table 3.1 : Summary of the FTIR spectra dataset   | 58    |
| Table 3.2 : Confusion matrix for binary classification  | 65    |
| Table 4.1 : Average accuracy of different models across various data dimensions   | 79    |
| Table 4.2 : Average sensitivity of different models across various data dimensions  | 80    |
| Table 4.3 : Average specificity of different models across various data dimensions  | 81    |
| Table 4.4 : Average accuracy of different models across various sample sizes  | 86    |
| Table 4.5 : Average sensitivity of different models across various number of samples  | 87    |
| Table 4.6 : Average specificity of different models across various data dimensions  | 88    |
| Table 5.1 : Method of prediction for selected model   | 98    |
| Table 5.2 : Average experimental performance of the base learners   | 100   |
| Table 5.3 : Comparative results on the SE model's average performance with the best-performing meta-learner for each evaluation | 103   |
| Table 6.1 : Comparison of the visual representation for PCA, PLS-DA and LDA in each of their respective transformation subspace | 108   |
| Table 6.2 : Violin plot of average accuracy of stacking models with different meta-learners for all datasets                    | 112   |

|           |   |   |     |
|-----------|---|---|-----|
| Table 6.3 | : | List of base learner and meta-learner used in SE for all datasets | 113 |
| Table 6.4 | : | Average performance of 5-Fold CV for D01                          | 114 |
| Table 6.5 | : | Average Performance of 5-Fold CV for D02                          | 116 |
| Table 6.6 | : | Average Performance of 5-Fold CV for D03                          | 118 |
| Table 6.7 | : | Average Performance of 5-Fold CV for D04                          | 119 |
| Table 6.8 | : | Average Performance of 5-Fold CV for D05                          | 121 |
| Table 6.9 | : | Average Performance of 5-Fold CV for D06                          | 123 |



UMS  
UNIVERSITI MALAYSIA SABAH

## LIST OF FIGURES

|  | Pages |
|--|-------|
| Figure 1.1 : Example of Spectra data   | 2     |
| Figure 1.2 : Curse of Dimensionality   | 4     |
| Figure 2.1 : General framework of pattern recognition on high-dimensional data | 15    |
| Figure 2.2 : Illustration of oversampling of data using SMOTE                  | 19    |
| Figure 2.3 : Feature selection   | 21    |
| Figure 2.4 : Feature Extraction  | 22    |
| Figure 2.5 : Bagging   | 26    |
| Figure 2.6 : Boosting  | 27    |
| Figure 2.7 : Stacked generalization ensemble architecture                      | 28    |
| Figure 2.8 : Before and after applying LDA                                     | 30    |
| Figure 2.9 : Sigmoid function of LR  | 32    |
| Figure 2.10 : Multilayer perceptron  | 33    |
| Figure 2.11 : Classification process of SVM                                    | 34    |
| Figure 2.12 : Decision tree  | 36    |
| Figure 2.13 : Random Forest  | 37    |
| Figure 2.14 : Example of FTIR instrument                                       | 39    |
| Figure 2.15 : IR absorptions of Common Functional Group                        | 41    |
| Figure 3.1 : Research Framework  | 55    |
| Figure 3.2 : Overall System Architecture                                       | 57    |
| Figure 3.3 : Mean absorbance value of spectra for each class for M1 dataset    | 59    |
| Figure 3.4 : Mean absorbance value of spectra for each class for M2 dataset    | 60    |

|            |   |     |
|------------|---|-----|
| Figure 3.5 | : Graph of the number of samples in the dataset after data sampling   | 62  |
| Figure 3.6 | : Stratified K-Fold Cross Validation  | 64  |
| Figure 4.1 | : HDR-PoT method  | 71  |
| Figure 4.2 | : Visual Comparison of the dataset with 50 dimensions   | 76  |
| Figure 4.3 | : Visual Comparison of the dataset with 100 dimensions  | 77  |
| Figure 4.4 | : Visual Comparison of the dataset with 150 dimensions  | 77  |
| Figure 4.5 | : Visual Comparison of the dataset with 200 dimensions  | 78  |
| Figure 4.6 | : Visual Comparison for the dataset with 50 samples   | 83  |
| Figure 4.7 | : Visual Comparison for the dataset with 100 samples  | 84  |
| Figure 4.8 | : Visual Comparison for the dataset with 500 samples  | 84  |
| Figure 4.9 | : Visual Comparison for the dataset with 1000 samples   | 85  |
| Figure 5.1 | : Flowchart for Heuristic SE  | 96  |
| Figure 5.2 | : Visual representation of SE algorithm   | 97  |
| Figure 5.3 | : Visualisation of the generated dataset after applying HDR-PoT   | 99  |
| Figure 5.4 | : Violin plot of the average accuracy for all 84 combinations of base learners with different meta-learners without HDR-PoT | 101 |
| Figure 5.5 | : Violin plot of the average accuracy for all 84 combinations of base learners with different meta-learners with HDR-PoT    | 102 |
| Figure 6.1 | : FTIR spectra Classification Framework   | 106 |



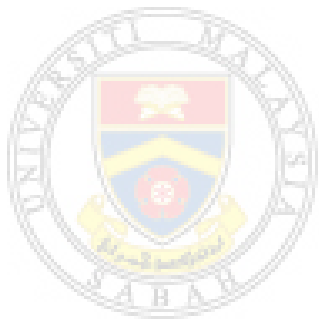
## LIST OF ABBREVIATIONS

|                |   |  |
|----------------|---|--|
| <b>AI</b>      | - | Artificial Intelligence  |
| <b>CART</b>    | - | Classification and Regression Tree                                 |
| <b>DT</b>      | - | Decision Tree  |
| <b>DR</b>      | - | Dimensionality Reduction   |
| <b>FN</b>      | - | False Negative   |
| <b>FP</b>      | - | False Positive   |
| <b>FTIR</b>    | - | Fourier Transform Infrared   |
| <b>GB</b>      | - | Gradient Boosting  |
| <b>HDR-PoT</b> | - | Heap Dimensionality Reduction with<br>Polynomial feature Technique |
| <b>LDA</b>     | - | Linear Discriminant Analysis                                       |
| <b>LR</b>      | - | Logistic Regression  |
| <b>ML</b>      | - | Machine Learning   |
| <b>MLP</b>     | - | Multilayer Perceptron  |
| <b>PCA</b>     | - | Principal Component Analysis                                       |
| <b>PLS-DA</b>  | - | Partial Least Square – Discriminant<br>Analysis                    |
| <b>RBF</b>     | - | Radial Basis Function  |
| <b>RF</b>      | - | Random Forest  |
| <b>SE</b>      | - | Stacking Ensemble  |
| <b>SVM</b>     | - | Support Vector Machine   |
| <b>TN</b>      | - | True Negative  |
| <b>TP</b>      | - | True Positive  |



## LIST OF APPENDICES

|  | Page |
|--|------|
| Appendix A : Full CV results of the dimensionality reduction techniques on different number of dimensions and samples size | 153  |
| Appendix B : Full results for all combination of learners in the SE method   | 167  |
| Appendix C : Full CV results on FTIR spectra data  | 170  |
| Appendix D : List of Published Papers and Conference Attended  | 194  |



UMS  
UNIVERSITI MALAYSIA SABAH

# CHAPTER 1

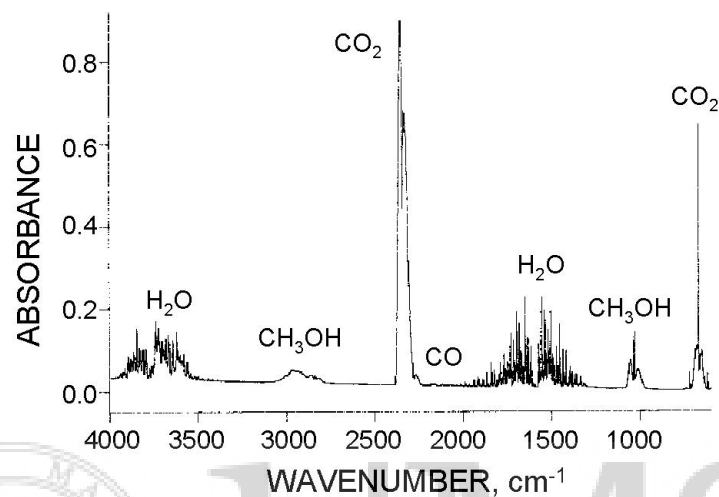
## INTRODUCTION

### 1.1 Overview

Pattern recognition is an interdisciplinary field that involves detecting patterns and relationships in data using computer algorithms. It covers many areas of statistics, engineering, artificial intelligence, computer science, bioinformatics and computer vision. It is aimed to extract patterns as efficiently as possible depending on particular criteria and distinguish one class from the others. Usually, pattern recognition can be sorted according to the type of learning procedure, either supervised or unsupervised. Supervised pattern recognition is a model that classifies based on experimental data where the unknown samples may be assigned to a previously established sample class according to their pattern of measured features. On the other hand, unsupervised pattern recognition arranges data into clusters and then defines it since it lacks a predetermined sample class (Bishop, 2006). Other terms for supervised and unsupervised pattern recognition are classification and clustering, respectively.

Pattern recognition is both the application of machine learning (ML) and statistical data analysis. ML is one of the subdisciplines of Artificial Intelligence (AI) and Computer Science. It uses data and algorithms to imitate humans' learning capabilities and eventually improves its accuracy without being programmed explicitly (Mitchell, 1997; Luxton, 2016). Statistical analysis is a scientific instrument that aids in data collection and analysis. Its purpose is to identify recurring patterns and trends and then turn them into useful information (Kerlinger & Lee, 2000; Ali & Bhaskar, 2016). Simply said, statistical analysis is a technique for data analysis that helps draw

useful conclusions from unstructured and raw data. There are many kinds of data, one of them is spectra data. Due to spectra data frequently comprising of the chemical information about the sample's composition, this kind of data, besides being intricate and multivariate, also has high dimensionality. Figure 1.1 shows an example of spectra data.



**Figure 1.1** : Example of Spectra data

Source : <https://webbook.nist.gov/chemistry/special/spray-combust/baseline-case/ftir/>

Chemometrics is the subdiscipline in the chemical field that employs statistical and mathematical procedures. Its purpose is to design or select the best measurement methodologies and experiments to extract information from chemical systems (Otto, 2017). Spectroscopy is a tool in chemometrics and an example of the analytical procedure for obtaining high-dimensional spectra data. It entails creating, measuring and analysing spectra resulting from infrared radiation's interaction with matter. This approach is informative and widely utilised for quantitative and qualitative assessments (Nielsen, 2017). Spectroscopic techniques that are accessible today come with many different types and functionality. One of them, notably Fourier Transform Infrared (FTIR) spectroscopy, will be the main focus of this research. FTIR spectroscopy has the potential to become an essential routine analytical tool as FTIR analysis can be performed rapidly with minimum sample preparation and without the use of a reagent.

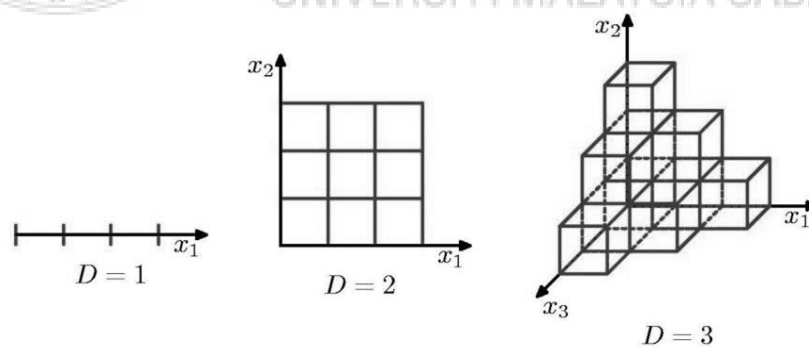
## 1.2 Problem Background

Over the past few decades, technological advancements have made novel discoveries and inventions possible. High-dimensional datasets are datasets in which the number of features  $p$  is larger than the number of observations  $N$ . Infrared spectrum data is an example of high-dimensional vectors that are made up of absorbance values corresponding to different wavenumbers. The characteristic variables of spectral data have become bigger as spectral detection technology become more advanced.

Applications for FTIR spectroscopy can be found in a wide range of disciplines. Two out of the many prominent fields that widely use FTIR spectroscopy are the identification of diseases in the medical profession and also the detection of food fraud – also known as economically motivated adulteration – in the food science and technology field. In the medical field, various studies have been conducted on a wide range of diseases, including neurodegenerative diseases. Neurons, cells of the nervous system, have a particular physiology that makes it possible for them to convey information effectively through an electrochemical process called signal transduction. The fundamental cause of neurodegenerative disorders is the progressive and irreversible loss of neuronal cells found in the tissues of the central and peripheral nervous system as well as brain tissue (Jellinger, 2010). This disease has a significant influence on a person's life and can even result in death. When dealing with living organisms specifically humans and animals, the FTIR spectra of the biological fluid (biofluids) such as blood, urine or sweat are the ones that are being compared. Typically, in the identification of diseases, the spectra data of the diseased samples will be compared with those of the healthy samples.

Aside from biofluids, FTIR spectroscopy is frequently used to characterise the ingredients in food products by performing rapid, non-destructive tests on both natural and artificial materials. In the topic of food adulteration, food-based product manufacturers must balance managing rapid screening of raw materials and ingredients as well as protecting the safety and quality of their products (Cebi *et al.*, 2023). One way to ascertain it is by the use of FTIR spectroscopy. Essentially, the FTIR spectra of the tested ingredient are measured and compared to a collection of known good ingredients.

Several difficulties arise when displaying high-dimensional data or during the training of the ML models. The most common problem is known as the "Curse of Dimensionality" which explains the explosive nature of expanding dimensions of data and the upsurge in computer work necessary for processing and analysis. In other words, the "Curse of Dimensionality" is a phenomenon where the performance improves until the maximum number of features is achieved. When more features are added with the same sample size, the classifier's performance worsens (Duda *et al.*, 1999). The Curse of Dimensionality is a phrase that was first used by Richard E. Bellman to illustrate how the addition of extra dimensions to the field of dynamic programming increased the volume of Euclidean space (Bellman, 1957). In theory, it is considered a blessing where increasing the dimensions might provide more information to the data, eventually boosting its quality (Donoho, 2000). In return, it is also a curse as it increases noise and redundancy during analysis. Thus, the common trend across these issues is that when dimensionality increases, the volume of space expands rapidly, that the accessible data becomes sparse. As a result, the amount of data required to produce a credible result frequently rises exponentially with dimensionality. When high dimensionality is combined with large sample numbers, concerns such as high computing cost and algorithmic instability arise (Jianqing *et al.*, 2014; Genender-Feltheimer, 2018).



**Figure 1.2 : Curse of Dimensionality**

Source : Bishop (2006)

In ML, a trait or attribute that characterises an object is considered as its feature. Each feature represents a dimension, and a set of dimensions forms a data point (Zhu & Goldberg, 2009). In other words, the term "dimension" refers to the number of attributes or features that describe each data point or instance. It

represents the number of variables or measurements associated with each sample in a dataset. The dimensionality of data determines the number of axes required to represent and visualize the data effectively (Guyon & Elisseeff, 2003; Aggarwal, 2015). As dimensionality increases, so does the number of characteristics required to describe the data. In the field of biomedical research, for example, age and gender might be utilised as factors to create some form of prognosis. A feature vector's dimensions are made up of these features. However, additional factors, such as patient history, blood composition and other related features can assist a doctor in assessing the prognosis more accurately. In this scenario, adding features theoretically expands the dimensions of the data. The number of data points required for any ML algorithm to function well increases exponentially as the dimensionality increases. This is because a larger amount of data points is required for each given combination of attributes for any ML model to be viable. Furthermore, traditional data classification strategies rely on recognising parts where objects form groups with similar attributes; however, with high-dimensional data, all objects look sparse and distinct in many respects, making standard data classification strategies inefficient (Altman & Krzywinski, 2018).

Another hurdle that arises when dealing with high-dimensional data is preventing overfitting the training data (Clarke *et al.*, 2008). It is crucial to build a classification model that is capable of generalisation. In machine learning, generalization refers to the ability of a trained model to perform accurately on unseen or new data that it has not encountered during the training phase. The model can capture and learn underlying patterns, trends, and relationships from the training data and apply that knowledge to make accurate predictions or classifications on previously unseen instances (Wang *et al.*, 2021). Generalization ensures that the model can perform well in real-world scenarios beyond the specific instances it was trained on. It is ideal that, in addition to having a good performance on the training set, such a model should do well on an independent testing set. However, when dealing with high-dimensional data, the low samples count usually causes the classification model to overfit the training data. This results in poor generalisation capability for the model (Pappu & Pardalos, 2014). Thus, the dimensionality reduction (DR) approach is a potential solution for the aforementioned problems as its main purpose is to reduce the data dimensions while retaining important information.

For the classification part, the classifier chosen mainly depends on the application domain to produce the best results. There are different kinds of classification problems. Binary classification or two-class classification is one of them. This sort of classification task occurs when only two distinct classes must be classified. Diseases against healthy, spam against non-spam and even strawberry against blueberry are examples of such classification. Linear Discriminant Analysis (LDA), Logistic Regression (LR), Support Vector Machine (SVM) and Decision Tree (DT) are some popular simple techniques that have been widely used, particularly for binary classification. Most classical algorithms have the advantages of being straightforward to apply, having speedy computational time, and performing accurately in a myriad of classification or regression applications.

However, as research in the pattern recognition field expands, the limits of the approaches utilised for classification are increasingly being explored. As a result, it is evident that highly specialised and correctly configured classifiers are quite powerful compared to classical approaches. However, selecting the best classifier for a given problem and properly configuring it is not a simple task. Furthermore, there is no ideal solution for solving all proposed problems. The quality of the classification results will be heavily influenced by the quantity and quality of data samples utilised in the training phase, as well as the feature selection and parameter settings (Kuncheva and Whitaker, 2003). Although some classifiers produce successful solutions on their own, Dietterich's (2000) experimental evaluation demonstrates a decline in quality especially when there are either vast sets of patterns or a considerable number of missing data samples or irrelevant features within the dataset. Thus, the effectiveness of such classifiers to accurately classify the patterns are reduced when presented with a more complicated problem.

Ensemble learning has been presented as a solution to the aforementioned problem to boost the advantages of simple algorithms while reducing their drawbacks. Ensemble learning, such as Bagging, Boosting, and Stacking approaches, is a reliable methodology for improving overall prediction performance. Theoretically, ensemble learning entails creating a series of learning algorithms, feeding each algorithm a set of data, and cleverly integrating the results to achieve high classification rates. In this regard, these computational approaches are