Department of Electrical and Computer Engineering

Faculty of Science and Engineering

HYPERSPECTRAL IMAGING FOR DETECTING AUTHENTICITY AND GEOGRAPHICAL ORIGIN OF SARAWAK GROUND BLACK PEPPER

Terence Chia Yi Kai

0000-0002-7553-0298

This thesis is presented for the Degree of
Master of Philosophy
of
Curtin University

March 2021

DECLARATION

To the best of my knowledge and belief this thesis contains no material previously published by any other person except where due acknowledgment has been made.

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university.

Signature: ______________________

Date: **26-4-2021**

ACKNOWLEDGEMENT

First and foremost, I would like to thank my principal supervisor Dr. Agus Saptoro for providing great opportunity to work on research project and endless guidance as well as support to my well-being despite my slow progress. Not just that, he has tremendously helped me in sourcing the black pepper products from various regions of Sarawak. Additionally, I would like to thank my co-supervisor Dr. Garenth Lim King Hann for providing useful lessons and tips on deep learning, convolutional neural network and the model training process, as well as local workstation (local desktop with powerful GPU) to experiment my model rapidly.

After that, I would like to thank Curtin University Malaysia for great student welfare and services. Especially Curtin University Graduate School, I had been awarded fee waiver to proceed with my research without worries. More importantly, I am grateful to Sarawak Multimedia Authority for the provision of grant to fund my research project.

Finally, I would like to thank my family especially my mother who has supported me all the way until I finish my whole research project. Without her, I would not be able to go on further.

LIST OF PUBLICATIONS

The contents of this thesis are going to be published in the following journals:

1. Chia, T.Y.K., Saptoro, A., 2021. Chemical and microbiological analysis of Black Pepper from Five Different Regions in Sarawak, Malaysia, in preparation to be submitted to Food Chemistry (2019 IF: 6.306, Q1). [Corresponds to Chapter 5]
2. Chia, T.Y.K., Saptoro, A., Lim, K.H., 2021. Adulterant Detection of Ground Black Pepper using Hyperspectral Imaging combined with SVM and Deep Learning based Classification Techniques, in preparation to be submitted to LWT – Food Science and Technology (2019 IF: 4.006, Q1). [Corresponds to Chapter 4]
3. Chia, T.Y.K., Saptoro, A., Lim, K.H., 2021. The Use of Hyperspectral Imaging and Various Classification Approaches to Identify Geographical Origin of Sarawak Black Pepper, in preparation to be submitted to Journal of Food Engineering (2019 IF: 4.499, Q1). [Corresponds to Chapter 4]

ABSTRACT

Pepper (*Piper nigrum*), is a popular spice and mainly grown in Sarawak, Malaysia. It has contributed to the agricultural sector in Malaysia for being 6th largest pepper producer in the world. As Sarawak pepper quality is highly valued around the world, black pepper, in particular, is subjected to adulteration by various additives such as papaya seeds, chilli powder and black pepper plant parts so that greater economic profit can be taken advantage of. Among these adulterants, papaya seeds are the most commonly identified adulterant due to their resemblance in appearance. There were several methods to detect these adulterants including microscopic, chromatographic, molecular and spectroscopic methods. However, these techniques have been acknowledged to be expensive and require tedious sample preparation. Additionally, proper specialist training is required to gain the expertise in operating these techniques effectively. In this regard, hyperspectral imaging appeals as a viable, rapid and non-destructive alternative for adulterant detection as the captured hyperspectral images contain rich spatial and spectral information, hence allowing interpretation of most hidden spectral information outside the visible light spectrum.

However, hyperspectral images are associated with large volume and high dimensionality of data. Consequently, they require proper data pre-processing and feature extraction. Chemometrics or multivariate data analysis is then used to extract representative features from these data to develop an interpretable model to determine the authenticity of Sarawak black pepper powder samples. It is also of interest to investigate the applicability of hyperspectral imaging and chemometrics to classify the geographical origin of Sarawak black pepper powder samples. Chemometrical methods such as principal component analysis (PCA), partial least square (PLS) and support vector machine (SVM) or support vector regression (SVR) were assessed on the determination of authenticity and geographical origin of Sarawak black pepper powder samples. At the same time, deep learning (DL) neural network models, such as, convolutional neural network (CNN) and stacked autoencoders (SAE) were explored and compared with current chemometrical methods. All these models were assessed on the full spectra range of mean spectra data from Visible-NIR hyperspectral images (400 – 1000 nm) with various pre-processing methods (SG, SG-SNV, SG-1st and SG-2nd). Results indicated that deep learning SAE model on SG-SNV pre-processed data had the best predictive performance on the determination of authenticity of black pepper powder samples, with R^2, root mean square error (RMSE) and mean absolute percentage error (MAPE) of 0.9010, 0.0143 and 1.17% respectively. While for the classification of geographical origin, all models except PLS with discriminant analysis had the best predictive performance with accuracy of 100%.

Next, Visible-NIR hyperspectral imaging was used to predict the chemical and biological analytical properties of Sarawak black pepper powder samples from the lab analyses. Predictive models were developed and assessed with various data pre-processing methods. The final results demonstrated that all the models had decent predictive performance, with SVR model being the best model on the prediction of chemical analytical properties. While for the prediction of biological analytical properties, DL CNN had the best predictive performance. Despite that, overall predictive performance was at most satisfactory.

Overall, HSI was demonstrated to be sufficient as rapid, non-destructive and affordable detector and estimator of authenticity and geographical origin of Sarawak black pepper powder samples. Further wavelength range of HSI and better quality of lab data is highly desirable to fully assess the capability of HSI in not just the quality assurance of black pepper powder samples, but also other food and agricultural products.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ANN	Artificial Neural Network
BIL	Band-interleaved-by-line
BIP	Band-interleaved-by-pixel
CFU	Colony Forming Unit
CNN	Convolutional Neural Network
CVS	Computer Vision System
DART-MS	Direct Analysis using Real-Time Mass Spectrometry
DL	Deep Learning
FTIR	Fourier-Transform Infrared
GC	Gas Chromatography
HPLC	High Performance Liquid Chromatography
HPTLC	High Performance Thin-Layer Chromatography
HSI	Hyperspectral Imaging
IPC	International Pepper Community
KNN	K-nearest Neighbour
LSTM	Long Short Term Memory
MAPE	Mean Absolute Percentage Error
MIR	Mid-infrared
MPB	Malaysian Pepper Board
MS	Mass Spectrometry
MSC	Multiplicative Scatter Correction
NIR	Near-infrared
PCA	Principal Components Analysis
PLS	Partial Least Square
PLS-DA	Partial Least Square – Discriminant Analysis
RAPD	Random Amplified Polymorphic DNA
ReLU	Rectified Linear Unit
RGB	Red Green Blue
RMSE	Root Mean Square Error
RMSEP	Root Mean Square Error of Prediction

SAE	Stacked Autoencoder
SCAR	Sequence Characterised Amplified Region
SFE	Supercritical Fluid Extraction
SG	Savitzky-Golay
SGD	Stochastic Gradient Descent
SIMCA	Soft Independent Modelling of Class Analogy
SNV	Standard Normal Variate
SVM	Support Vector Machine
SVR	Support Vector Regression
TLC	Thin Layer Chromatography
UV	Ultraviolet

CHAPTER 1
INTRODUCTION

Pepper (*Piper nigrum*) is one of the most consumed spices in the world. It is mainly used as a savoury seasoning, preservative and even medicine (Ravindran and Kallupurackal, 2001). The main constituent of pepper is piperine which is one main part of alkaloids and responsible for its pungency. It is shown to exhibit various physiological and pharmacological properties including antioxidant, anti-inflammatory, antimutagenic, antitumor, antiapoptotic, antigenotoxic, antiarthritic, antifungal, antidepressant, anti-hepatitis B and gastro-protective activities (Embuscado, 2019; Orrillo et al., 2019; Shityakov et al., 2019). Due to these benefits, pepper is highly demanded by consumers and widely grown in tropical countries, for example, the Asia Pacific countries such as India, Malaysia, Indonesia, Thailand, Vietnam, China, Sri Lanka, Cambodia, and some regions in Southern America such as Brazil, Mexico, Guatemala and so on (Ravindran and Kallupurackal, 2001).

Pepper is harvested, threshed, blanched and dried before being further processed depending on the maturity of the harvested pepper fruits (Ravindran and Kallupurackal, 2001). Black and white pepper are the common end products found in the market. They are either made into berry or powder form for sale and export both locally and globally. Malaysia, being a member country of International Pepper Community (IPC), is 5th largest pepper producer in the world, with the production of 32.3 thousand tonnes and 33.9 thousand tonnes in 2018 and 2019 respectively (FAO, 2019; IPC, 2018; Mohd Uzir Mahidin, 2020a). IPC is an organisation comprising of mainly pepper producing countries which include India, Malaysia, Indonesia, Vietnam, Sri Lanka, Philippines and Papua New Guinea (IPC, 2021). It was established in 1972 to promote, co-ordinate and streamline the activities to bolster the pepper economy, such as to encourage research on diseases of pepper plant and development of varieties that exhibit high yield and resistance to common diseases, streamline the exchange of information and provision of statistics on pepper production, consumption, trading and pricing (IPC, 2021). In Malaysia, pepper is one of the important crops owning to its contribution to the agricultural sector. The agricultural sector has helped to contribute 7.3% of the Gross Domestic Product which amounts RM99.5 billion in 2019 in Malaysia, with a positive growth of 2% (Mohd Uzir Mahidin, 2020b). On 2020, over 8.5 thousand tonnes of pepper has been exported to major countries like Japan, China, Vietnam, Taiwan and Singapore with export value of RM120.80 million (MPB, 2021). All of these are owning to Sarawak, being the primary pepper producer among all states in Malaysia, as it accounts for 98% of total pepper production (Entebang et al., 2020; MPB, 2021). Additionally, high quality of Sarawak pepper ensures Malaysia's competitiveness in the global market on the pepper industry, as the prices of Sarawak peppers are comparatively higher than those from other major exporting countries. This is presently shown in IPC website where the price of pepper for Kuching variant has been higher and more stable than that of most peppers from various countries (IPC, 2021). Furthermore, it boosts the Sarawak's economic growth by improving the quality of life and creating more job opportunities for local farmers in rural areas (Entebang et al., 2020; Keong, 2017).

Nowadays, food fraudulence has become a norm, covering various types of food such as herbs and spices, meat, vegetables, dairy products, drugs, oil and so on. To take advantage of greater economic profit, adulterants, which exhibit features of being similar in appearance, smell and visually appealing are intentionally added into the original goods of interest since they are cheaper to be produced (ASTA, 2016; Bansal et al., 2017; Dhanya et al., 2009; Oliveira et al., 2019). Adulterants can also potentially cause adverse effects to human health if consumed excessively (Bansal et al., 2017; Bawden, 2015; Kassie et al., 1999; Kermanshai et al., 2001). Hence, this raises the importance of detector and estimator of adulterants in quality assurance within the food industry. Pepper is one of the highly-priced commodities which is vulnerable to adulteration due to the aforementioned benefits, primarily the papaya seeds are usually added into black pepper products because of their resemblance in appearance with the black pepper (Orrillo et al., 2019). Therefore, stringent regulation of the quality of these pepper products must be enforced at all times.

Currently, in the literature, there were numerous methods of detecting the adulterants of black pepper products qualitatively and quantitatively. They could be divided into several categories. One of them was the physical technique which involved density-based method and visual inspection (Attrey, 2017; Bansal et al., 2017; Curl and Fenwick, 1983; Dhanya et al., 2009; Orrillo et al., 2019; September, 2011; Tremlová, 2001; Vadivel et al., 2018). While they were simple to implement, these only provided qualitative information and created some challenges such as tedious sample preparation and expertise in carrying out the operation (Bansal et al., 2017; Vadivel et al., 2018). The other category was chemical analytical techniques which are commonly used as a detector of adulterants both qualitatively and quantitatively. This involved gas chromatography (Curl and Fenwick, 1983), thin-layer chromatography (Bhattacharjee et al., 2003; Paradkar et al., 2001), high-performance liquid chromatography (HPLC) (Jain et al., 2007; Vadivel et al., 2018), supercritical fluid extraction using carbon dioxide (Bhattacharjee et al., 2003), direct analysis of metabolic profile using real-time mass spectrometry (DART-MS) (Chandra et al., 2014) to detect for adulterants in black pepper samples. Liquid-liquid extraction using ethanol was also used to detect piperine content on adulterated black pepper with papaya seeds (Madan et al., 1996). Detailed qualitative and quantitative results could be obtained with the shortcomings of requiring tedious sample preparation, ample specialist training, high cost, causing destruction to samples and having their operations limited within laboratories (Abbas et al., 2018; Bansal et al., 2017; Orrillo et al., 2019). Additionally, extraction and chromatographic methods produced chemical waste (Manley, 2014). Then, another type is molecular technique, which is a biological method to discriminate the adulterants from the original product and authenticate the product origin using molecular marker (Bansal et al., 2017; Zhang et al., 2019). Examples of molecular techniques include random amplified polymorphic DNA (RAPD) (Khan et al., 2010), sequence characterised amplified region (SCAR) marker (Dhanya et al., 2009), qPCR method (Sousa et al., 2019), and DNA barcoding (Parvathy et al., 2014; Zhang et al., 2019). However, its usage had to be in the hands of specialists and was limited only within laboratories (Oliveira et al., 2019).

Due to the aforementioned drawbacks, the need for rapid, informative, cost-effective, non-contact, non-destructive and reproducible technique to detect for adulterants becomes essential. Vibrational spectroscopic technique, being part of the chemical analytical techniques is another method that provides spectral fingerprint or chemical information of any given sample (Lohumi et al., 2015; Park and Lu, 2015; Rodriguez-Saona et al., 2016). Near-infrared (NIR), mid-infrared (MIR), Fourier Transform infrared (FTIR) spectroscopies were some of the examples that were widely used as adulterants detection tools present in black pepper samples (McGoverin et al., 2012; Orrillo et al., 2019; Vadivel et al., 2018; Wilde et al., 2019). However, they are mostly point-based and can only cover a small area of sample which may not be representative to the sample as a whole (Lohumi et al., 2015). Another method is computer vision system (CVS) which is an optical method that captures digital spatial images of the whole object within the visible light spectrum to detect adulterants present in samples. It has been broadly used in food and agricultural products quality assessment, monitoring, evaluation and assurance (Bhargava and Bansal, 2018; Chen et al., 2019; Di Rosa et al., 2017; Koirala et al., 2019; Patel et al., 2012; Rong et al., 2019; Taheri-Garavand et al., 2019; Vithu and Moses, 2016; Wu and Sun, 2013a; Xu et al., 2017, 2019). This worked well for black pepper samples which are in berry form (P. R. Goswami and K. R. Jain, 2013), but did not perform well if the samples were in powdered form due to stark similarities on appearance (Bhargava and Bansal, 2018). Hyperspectral imaging (HSI) is one such solution that incorporates CVS and vibrational spectroscopy technique, providing digital spatial images over the range of ultraviolet, visible light and infrared regions (Elmasry et al., 2012; Lohumi et al., 2015; Park and Lu, 2015). Researches had been done to utilise HSI to predict essential analytical chemistry parameters such as % sugar content, % moisture content and % volatile oil content using the information from continuous bands of wavelengths through supervised learning (Ke et al., 2020; Shorten et al., 2019).

The size of hyperspectral images is substantial considering the rich spectral information with continuous bands of wavelengths for each pixel. Multivariate data analysis is applied to extract this information to produce a simpler and interpretable mathematical or statistical model. Chemometrics appears as part of multivariate data analysis which is tailored for determination of various chemical analytical properties (Elmasry et al., 2012). Chemometrics had been broadly applied in HSI applications to assess and predict the quality of various food and agricultural products, such as nitrogen content, mineral compositions, moisture content and others (Khan et al., 2020). Particularly, for the case of black pepper, HSI with the help of multivariate data analysis, had been used to detect for adulterants such as papaya seeds, buckwheat and millet (McGoverin et al., 2012; Orrillo et al., 2019). The main drawbacks of most multivariate data analysis methods are requirement of prior domain knowledge and frequent human intervention to effectively extract and utilise the information from the vast amount of data. It was later proven that deep learning is one such solution that allows automatic feature extraction during the processing of hyperspectral images in raw format without too much pre-processing (Al-Sarayreh et al., 2018; Emmert-Streib et al., 2020; LeCun et al., 2015; Li et al., 2017). Additionally, it is mainly powered by data, making it perform better when more data is provided. Deep learning on hyperspectral images had been mainly applied in remote sensing, but it was broadly expanded to various fields, particularly food and agriculture as indicated in recent increasing number of researches on that field (Saha and Manickavasagan, 2021; Signoroni et al., 2019; Yang et al., 2019).

1.1 Motivation

In this research, the main adulterant of black pepper products to be investigated is papaya seeds because of prevalent coverage in literature regarding its usage. Currently, black pepper products quality assessment and adulteration detection are mostly performed using laboratory chemical analytical techniques. While they provide detailed analyses, since most quality assurance and control activities demand the analysis to be done on-site and in real-time, the chemical analytical techniques usually require tedious sample preparation and proper specialist training, thus increasing time taken to retrieve the results. CVS is ineffective in detecting the adulterants with resembling appearance, even though it is affordable and rapid. Most spectroscopic techniques are point-based and thus time-consuming. Hence, HSI is introduced as a rapid alternative that combines computer vision and spectroscopy to assess the quality of black pepper products. However, HSI is not widely adopted in most on-site quality assurance and control of food and agricultural products in Malaysia. Most research on this field were still bounded within lab scale (Saha and Manickavasagan, 2021). This research serves to assess the applicability of HSI in determining the quality of particularly black pepper powder products compared to chemical analytical techniques, so that HSI can be included as a potential alternative in most industrial real-time quality assurance applications.

Hyperspectral images by nature have high dimensionality and a large volume of data, thus they require proper data pre-processing and effective feature extraction so that an interpretable model can be developed to verify the presence of adulteration in black pepper products and determine their quantity. Most multivariate data analysis techniques require prior knowledge and human intervention to effectively extract features and fit the model. Deep learning, on the other hand, allows automatic feature extraction and processing of these data in raw format without compromising the accuracy of its outputs. There were numerous deep learning methods which were mainly tasked for detecting food spice powder products, and the most commonly found ones were convolutional neural network and stacked autoencoders. To date, no previous work is found for utilising deep learning models in detecting and quantifying the adulterants present in black pepper powder samples, thus it is addressed in this research project. Furthermore, only a few handfuls of resources could be found in identifying the geographical origin of black pepper samples using HSI. It is thus addressed in this project as the geographical origin of black pepper samples has considerable influence on the pricing of black pepper and understanding it ensures appropriate actions can be made to maintain the excellence of quality of black pepper for that identified region (Liang et al., 2021). It is also of interest to predict the chemical and biological analytical properties using solely HSI inputs. This allows immediate analytical results to be produced from machine learning models while waiting for occasional chemical analyses to complete, hence reducing chemical resources and wastes.

1.2 Objectives

The research objectives are outlined as follows:

1. To determine the characterisation of black pepper powder samples of different purity and geographical origins in Sarawak (Serian, Sungai Tenggang, Pakan, Lachau, Sibu) by representative features extracted from the HSI inputs;
2. To assess and compare the predictive performance of various machine learning models which includes PLS, SVM and deep learning using HSI inputs in determining the purity of black pepper powder products and their geographical origins using RMSE (comparing predicted and true purity values of black pepper) and accuracy (comparing predicted and true geographical origins of black pepper) respectively;
3. To evaluate the predictive performance of various machine learning models using HSI inputs on the prediction of chemical and biological analytical properties of black

pepper powder samples and compare them with true results from detailed chemistry analyses

1.3 Significance

This research seeks to bring about the contributions as follows:

1. Since chemical analyses which are currently used to carry out quality control of black pepper require tedious sample preparation, high cost and expertise in operating and maintaining the equipment, HSI with functional machine learning model can be a viable, affordable and user-friendly alternative on lab scale and industrial on-line quality assessment;
2. Development of such model ensures the consistent high quality of Sarawak black pepper powder products, hence bolstering the confidence among consumers on the food safety of the black pepper powder products;
3. It is envisaged that the introduction of such technologies will enable Sarawak and subsequently Malaysia to achieve the goal of boosting the agricultural industry by streamlining the quality assurance process of not just black pepper products, but also various food and agricultural products, as well as digital economy indirectly by empowering the locals with the knowledge and understanding of cutting edge technologies and tools (SMA, 2019);
4. Dataset containing qualitative and quantitative information of Sarawak black pepper from HSI can be provided, which subsequently delivers valuable comparisons and benchmark of machine learning models regarding the adulteration detection, classification of geographical origins and prediction of internal quality of black pepper products for future refinements

1.4 Thesis Outline

The contents of this thesis can be outlined as follows:

Chapter 2 provides a detailed research background and literature review on adulteration issues for black pepper powder products, current adulterants detection methods, HSI and various machine learning models including deep learning before moving on to the core research.

Chapter 3 presents the methodology of the research project. This includes materials preparation such as black pepper powder samples and its adulterant, chemical analyses to be carried out, equipment and apparatus to be used, data pre-processing and processing, design of training pipeline for various machine learning models and optimisations to be made.

The results of the research project are divided into two chapters:

Chapter 4 presents the results for the adulteration detection and classification of geographical origin of black pepper powder samples. This includes an overview of preliminary data exploration, data pre-processing and results from various machine learning models including PLS, SVM and deep learning. The deep learning model training is also accompanied by a series of model parameters tuning. A critical and detailed analysis is made to assess if HSI qualify to be rapid, affordable and major alternative for chemical analytical techniques in detection of adulterants present in black pepper samples or not.

Chapter 5 shows the results for the prediction of chemical and biological analytical properties based on solely HSI inputs. Comparisons on the predictive performance from various machine learning models including PLS, SVM and deep learning are made and analysed to assess the suitability of HSI inputs in predicting chemical and biological analytical properties accurately or not.

Chapter 6 concludes the results and analyses of the research that shall fulfil the research objectives. Future outlook and recommendations on this research project are outlined as well.

CHAPTER 2
RESEARCH BACKGROUND

Food fraudulence has become a prevalent issue, covering various types of food such as herbs and spices, meat, vegetables, dairy products, drugs, oil and so on. Adulterants, which exhibit appealing appearance and similarity in appearance and smell with the original products of interest are intentionally added since they are cheaper and affordable to be produced to take advantage of greater economic profit (ASTA, 2016; Attrey, 2017; Bansal et al., 2017; Dhanya et al., 2009; Galvin-King et al., 2018; Oliveira et al., 2019; Sørensen et al., 2016). Adulterants can also potentially cause adverse effects to human health. For example, peanuts and almonds, which can be fatal for those who are allergic to them, are usually used as adulterants in cumin and paprika (Bawden, 2015), coffee powder adulterated with date seed powder or tamarind induces diarrhoea (Bansal et al., 2017), papaya seeds present in black pepper may cause damage to liver and DNA when consumed in excessive amount due to presence of benzyl isothiocyanate in papaya seeds (Bansal et al., 2017; Kassie et al., 1999; Kermanshai et al., 2001). Hence, this raises the importance of detector and estimator of adulterants in quality assurance within the food industry.

Pepper is one of the valuable spices and commodities in the world. On 2020, over 588 thousand tonnes of pepper had been produced around the globe and Malaysia was the 5th largest pepper producer with 30.8 thousand tonnes of pepper was produced (IPC, 2018; MPB, 2021). In terms of export and import of pepper in Malaysia, 8.5 thousand and 2.2 thousand tonnes of pepper were made respectively. Although other countries such as Vietnam and India had higher production rates than in Malaysia, Malaysia has the highest pepper price among all the pepper producing countries (Entebang et al., 2020; IPC, 2021). Because of its high economical and aforementioned nutritional value, pepper is particularly vulnerable to adulteration. The target of this research is black pepper is due to the fact that the adulteration issue on black pepper, both in berry and powdered forms, receives more attention than that of white pepper judging from the larger volume of literature covering such this issue. The adulterants of black pepper found in the literature were commonly papaya (*Carica papaya L*) seeds (Bhattacharjee et al., 2003; Curl and Fenwick, 1983; Dhanya et al., 2009; Govindarajan and Stahl, 1977; McGoverin et al., 2012; Orrillo et al., 2019; Paradkar et al., 2001; Sousa et al., 2019; Vadivel et al., 2018; Wilde et al., 2019), and the others including chilli powder (Parvathy et al., 2014), other pepper species (Chandra et al., 2014; Sousa et al., 2019), buckwheat and millet (McGoverin et al., 2012; September, 2011), maize (Sousa et al., 2019), black pepper plant parts such as husk, pinheads and spent (Wilde et al., 2019), cassava starch and corn flour (Lima et al., 2020).

2.1 Detection of Adulteration of Black Pepper

To effectively detect for adulterants present in black pepper products qualitatively and quantitatively, numerous methods had been extensively researched in the past. They were divided into physical technique, chemical analytical technique which included vibrational spectroscopy, molecular technique, and imaging system including computer vision and hyperspectral imaging. Their functionality along with pros and cons are presented in Table 2.1. Most current adulteration detection methods of black pepper products are reviewed and outlined as tabulated in Table 2.2. The details of each method are subsequently elucidated in the following sections.

2.1.1 Physical Techniques

Physical technique is a simple technique that serves to provide macroscopic and microscopic visual inspection and analysis, and analyse various physical parameters of the samples of interest such as morphology, texture, solubility, bulk density, integrity and so on (Bansal et al., 2017). The oldest method involving physical technique to detect adulterants in black pepper samples was the use of alcohol to detect for floating papaya seeds in black pepper samples (Curl and Fenwick, 1983; Dhanya et al., 2009; September, 2011). Since that method was prone to human error, optical microscopic examination was used instead (Bansal et al., 2017; Orrillo et al., 2019; Vadivel et al., 2018). From the examination as illustrated in Figure 2.1, it revealed the presence of exclusive components found in papaya seed powder such as criss-cross fibres, fatty oil, oil globules, endosperm, long tubular cells, testa cells, aleurone grains, tracheids, vessel with spiral thickening and thick-walled parenchyma (Vadivel et al., 2018). Other common components found in both black pepper and papaya seed powders were brown content, simple and compound starch grains, calcium oxalate crystals, fibres, parenchyma cells, sclerenchyma cells, sclereids and trichomes (Vadivel et al., 2018). One can identify the difference between black pepper and papaya seed powders through careful observation, sufficient expertise and good quality equipment. However, tedious sample preparation and availability of only qualitative assessment are main drawbacks of the physical detection methods since quantitative information is essential to fulfil industrial requirements and provide tangible information on the severity of adulteration (Bansal et al., 2017; Vadivel et al., 2018).

Table 2.1: Black pepper adulteration detection methods and their advantages and disadvantages

Detection Method Type	Feature	Advantages	Disadvantages
Physical	Detect components physically using visual observation	• Easy to implement and execute • Rapid result fetching	• Only qualitative results are available • Tedious sample preparation
Chemical Analytical	Detect and measure individual components using chemicals	• Provide detailed qualitative and quantitative analysis • Analysis results are highly accurate	• Tedious sample preparation • Not rapid • Samples are destroyed after analysis • Produce chemical waste
Spectroscopic	Use vibrational spectroscopy to detect spectral footprint of the sample	• Provide non-destructive qualitative and quantitative analysis • Ease of sample preparation	• Measurement is only done on one point at one time, hence require averaging • Not rapid
Molecular	Utilise molecular markers to identify DNA fingerprints of individual components present in the sample	• Provide detailed qualitative and quantitative analysis • Rapid result fetching	• Tedious sample preparation • Samples are destroyed after analysis • Operations limited within laboratories
Computer Vision	Use imaging device to capture images of samples within visible light region and perform detection	• Provide non-destructive qualitative and quantitative analysis • Ease of sample preparation	• Does not work well for samples with high visual similarity • Not as detailed as chemical analytical technique
Hyperspectral Imaging	Use hyperspectral imaging device to capture hyperspectral datacubes of samples and perform detection and estimation	• Provide non-destructive qualitative and quantitative analysis • Ease of sample preparation • Has additional spectral dimension to provide informative analysis	• Has large amount of data with high number of dimensions, causing data management issues • Not as detailed as chemical analytical technique

Table 2.2: Various black pepper adulteration detection methods found in the literature

Reference	Detection Methods	Detection Method Type	Adulterants	Remarks
Lima et al. (2020)	NIR spectroscopy	Spectroscopic	Starch cassava and corn flour	Using NIR with range of 1100 – 2500 nm, a model based on PLS to differentiate type of adulterants in black pepper and cumin samples yielded correlation coefficient of above 0.9 and RMSE ranging from 2.2 to 7, and based on 13 commercial black pepper powder samples, 62% of them were found to be adulterated with cassava starch or corn flour
Wilde et al. (2019)	NIR and FTIR spectroscopy	Spectroscopic	Papaya seeds, chilli powder and black pepper spent parts	NIR and FTIR binary classification model of whether the black pepper samples were adulterated or not had a measure of fit R2 of 0.93 and 0.83, and predictive ability Q2 of 0.98 and 0.97, respectively
Sousa et al. (2019)	DNA barcoding	Molecular	Papaya seeds, maize and cayenne pepper	qPCR-based method on plant DNA barcodes (*trnL* and *psbA-trnH*), detected 41% of all 29 samples were adulterated
Orrillo et al. (2019)	NIR hyperspectral imaging	Hyperspectral Imaging	Papaya seeds	Classification using principal components analysis and soft independent modelling of class analogy yielded accuracy of 100% for whole samples and 90% for ground samples; PLS regression pre-processed by standard normal variate (SNV) and 2nd derivate yielded RMSEP of 2.51 and coefficient of determination of 0.93
Gul et al. (2018)	GC-mass spectrometry (MS)	Chemical Analytical	Papaya seeds	Adulteration of papaya seeds on black pepper samples was able to be detected using the 500- and 750-bp-sized SCAR markers to amplify their specific SCAR primer sets. Additionally,

	Sequence Characterised Amplified Region (SCAR) marker	Molecular		metabolic profiling using GC-MS on the black pepper and papaya seeds powder samples with different adulteration ratios was able to identify respective specific metabolites as low as 20 mg/g
Vadivel et al. (2018)	Microscopic	Physical	Papaya seeds	NIR spectroscopy was more efficient and rapid than microscopic, phytochemical techniques, HPTLC and GC-MS in black pepper adulterants detection
	Phytochemical techniques, high performance thin-layer chromatography (HPTLC), GC-MS	Chemical Analytical		
	NIR spectroscopy	Spectroscopic		
Dissanayake et al. (2016)	DNA barcoding (*psbA-trnH*)	Molecular	Papaya seeds and chilli powder	Used DNA barcoding primer pair, *psbA-trnH*, to amplify DNA; black pepper DNA produced 200 bp bands, chilli and papaya DNA produced 450 bp bands, while mixture of black pepper with chilli and papaya seeds produced both 200 bp and 450 bp bands
Parvathy et al. (2014)	DNA barcoding	Molecular	Chilli powder	Three barcoding loci *psbA-trnH*, *rbcL*, *rpoC1* were used to detect adulteration in black pepper, *psbA-trnH* was the best in detecting chilli powder present in black pepper sample
Chandra et al. (2014)	Direct analysis using real-time mass spectrometry with multivariate analysis (DART-MS)	Chemical Analytical	Other pepper species (Indian tipali, Bangla tipali)	Metabolic profiling of peppers using DART-MS then performed discrimination using principal components analysis, alkaloids and amides of black pepper and other pepper species were found
P. R. Goswami and K. R. Jain (2013)	Computer vision and image processing	Computer Vision	Papaya seeds and other visually similar foreign objects	Used Canny edge detection on RGB colours converted to grayscale image containing sparsely scattered black pepper with papaya seeds and other foreign objects

McGoverin et al. (2012)	NIR and MIR hyperspectral imaging	Hyperspectral Imaging	Papaya seeds	With recorded hyperspectral images of spatial resolution of 300 µm × 300 µm with a range of 1000–2500 nm at 6.3 nm intervals, root mean square error of prediction (RMSEP) was 2.7% and ratio of standard error of prediction to standard deviation was 11.14
Khan et al. (2010)	Random Amplified Polymorphic DNA (RAPD)	Molecular	Papaya seeds	Five out of eight decamer oligonucleotide primers indicated clear discrimination on black pepper and papaya seeds
Dhanya et al. (2009)	SCAR marker	Molecular	Papaya seeds	Detected the presence of papaya seeds in one of five branded black pepper samples using SCAR molecular marker
Jain et al. (2007)	Fluorescence and high performance liquid chromatography (HPLC)	Chemical Analytical	Papaya seeds	Detected papaya seeds adulteration in black pepper from HPLC; HPLC profiles of black pepper market samples had adulterated peaks compared to pure genuine black pepper samples at different retention times
Bhattacharjee et al. (2003)	Supercritical fluid extraction (SFE) using carbon dioxide	Chemical Analytical	Papaya seeds	Thin-layer chromatography analysis on SFE showed fluorescent band at 366 nm at R_f 0.172 proving the presence of papaya seed in black pepper, gas chromatography identified aldehydes *n*-nonanal, *n*-decanal, *n*-dodecanal proving the same

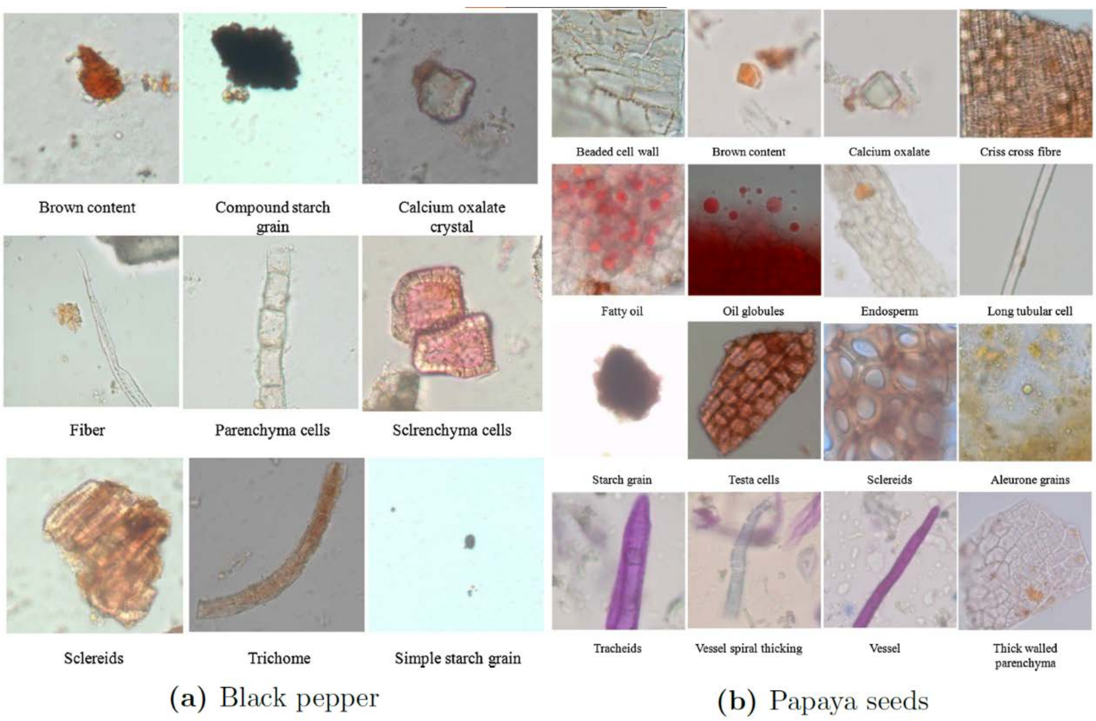


Figure 2.1: Microscopic characteristics of black pepper and papaya seeds (Vadivel et al., 2018)

2.1.2 Chemical Analytical Techniques

Chemical analytical techniques have been one of the most commonly used and well established qualitative and quantitative methods to perform quality assessments on various food and agricultural products with remarkable accuracy and precision. In assessing the degree of adulteration or authenticity of black pepper products, various literature indicated that most assessments involved chromatographic methods and vibrational spectroscopic techniques.

2.1.1.1 Chromatography

Chromatography is a technique that separates a mixture into basic components through the movement of molecules between stationary and mobile phases driven by affinity among the molecular weights and characteristics of the basic components (Coskun, 2016). There are various types of chromatography, such as gas, column, ion-exchange, thin-layer, high-performance liquid, affinity chromatography and others. Among them, the chromatographic techniques that were recently used in detecting for adulterants present in black pepper samples are as follows:

- Gas chromatography (GC):

Gul et al. (2018) utilised GC-MS to perform metabolic profiling on the black pepper and papaya seeds powder samples in different adulteration ratios. GC-MS was able to detect 44 and 33 metabolites out of 84 and 61 chemically diverse metabolites in black pepper and papaya seeds respectively. The detection performance could achieve detection as low as 20 mg/g of metabolites.

- Thin-layer chromatography (TLC):

Bhattacharjee et al. (2003) performed supercritical fluid extraction (SFE) using carbon dioxide on various proportions of black pepper and papaya seed samples and then to be sent to thin-layer chromatography analysis using ethylene dichloride as solvent at 366 nm. The results indicated that SFE extracts of papaya seeds yielded an additional green fluorescent band at R_f 0.172, thus proving the ability to detect the presence of adulteration of papaya seeds in black pepper samples, qualitatively, however.

- High-performance liquid chromatography (HPLC):

Jain et al. (2007) revealed that by using fluorescence markers and HPLC fingerprints, black pepper samples displayed lemon yellow fluorescence while papaya seeds displayed blue fluorescence under UV irradiation at 365 nm. Then, HPLC analysis showed that black pepper market samples had adulterated peaks compared to pure genuine black pepper samples at different retention times, distinguishing the difference between pure and adulterated black pepper products qualitatively.

- High-performance thin-layer chromatography (HPTLC):

Vadivel et al. (2018) prepared extracts of black pepper and papaya seeds using methanol which was then sent to HPTLC analysis to be viewed under UV irradiation at 254 and 366 nm. The results revealed that papaya seed extract had different band (R_f 0.35) than those of black pepper extracts (R_f 0.59) and standard piperine (R_f 0.58), allowing the detection of adulterants present in black pepper products qualitatively.

2.1.1.2 Vibrational Spectroscopy

Chromatographic techniques required high equipment and operation cost, expertise as well as their operations to be performed within the vicinity of laboratories. Furthermore, they produce environmental waste after the analysis (Manley, 2014). Vibrational spectroscopic technique, being another part of the chemical analytical technique is an optical method that evaluates the spectral fingerprint or the unique chemical information of a certain material (Elmasry et al., 2012; Lohumi et al., 2015; Park and Lu, 2015). The spectral fingerprint of a material refers to reflection, scattering, absorption and emission of electromagnetic energy in unique patterns at certain wavelengths or frequencies because of their intrinsic physical structure and chemical composition (Danezis et al., 2016; Elmasry et al., 2012). Some examples of vibrational spectroscopic techniques include near-infrared (NIR), mid-infrared (MIR), Fourier Transform infrared (FTIR), Raman spectroscopies and others (Lohumi et al., 2015; Park and Lu, 2015). A brief schematic of how these techniques work is depicted in Figure 2.2. Among these techniques, NIR, MIR and FTIR were used in detecting the adulterants present in black pepper samples.

NIR covers the wavelength range between 780 – 2500 nm where overtones and combinations of fundamental vibrations of molecules occur. Examples of this include $C-H$, $O-H$, $N-H$ chemical bonds that have high vibrational frequency (Jha, 2016; Lohumi et al., 2015; Oliveira et al., 2019; Park and Lu, 2015). On the other hand, mid-infrared covers the wavelength range between 2500 – 25000 nm which is comprised of functional group (usually 2500 – 6667 nm) and fingerprint (usually 6667 – 20000 nm) regions. Examples of functional groups in MIR are $O-H$ and $N-H$ stretching between 2703 – 4000 nm, $C-H$ stretching between 3030 – 3571 nm, triple bonded functional groups ($C\equiv C$, $C\equiv N$) between 3704 – 5405 nm and double bonded functional groups ($C=C$, $C=N$, $C=O$) between 5128 – 6897 nm (Jha, 2016; Lohumi et al., 2015; Oliveira et al., 2019). Further details about various spectra structure residing in NIR can be referred to Figure A.1 in Appendix A. Most spectroscopic analyses are processed using multivariate data analysis or chemometrics, which will be discussed in Section 2.3 as the spectroscopic results contain redundancies and are thus difficult to interpret and yield meaningful results. FTIR spectroscopy measures fundamental vibrations

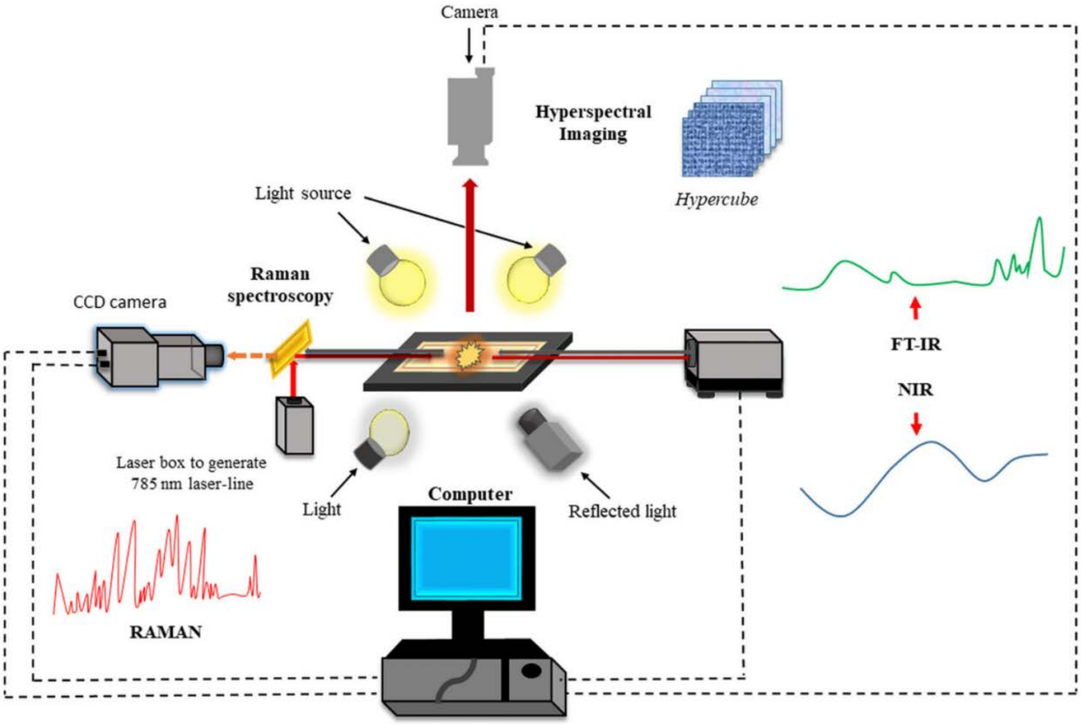


Figure 2.2: Brief introduction of most vibrational spectroscopic techniques (Oliveira et al., 2019)

instead of overtones and combinations through various measurement modes which include attenuated total reflectance, diffuse reflectance, high-throughput transmission and transmission cell (Jha, 2016; Lohumi et al., 2015). Attenuated total reflectance appeared as the most widely used mode in FTIR spectroscopy due to minimal effort in sample preparation for its qualitative and quantitative analysis. Vibrational spectroscopy techniques are widely used as qualitative and quantitative analysis for various food, condiments and agricultural products on their authenticity and adulteration levels (Kucharska-Ambrożej and Karpinska, 2020; Lohumi et al., 2015; Reinholds et al., 2015). McGoverin et al. (2012), Wilde et al. (2019) and Lima et al. (2020) utilised vibrational spectroscopy techniques in the detection of adulterants present in the black pepper samples.

McGoverin et al. (2012) used NIR hyperspectral imaging, which has better functionality than usual NIR spectroscopy and will be discussed further in Section 2.1.5, to detect for adulterants present in ground black pepper samples which were buckwheat and millet. Various proportions of buckwheat and millet were ground and mixed with ground black pepper samples, then sent to NIR hyperspectral imaging and FTIR spectrometer for MIR measurements. All the hyperspectral images were taken with spatial resolution of 300 µm × 300 µm and spectral range of 1000 – 2500 nm at 6.3 nm intervals. For MIR spectral measurements, they were recorded within 550 – 3999 cm^{-1} with spectral resolution of 4 cm^{-1}, with an average of 32 scans for each measured spectrum. Partial least square (PLS) regression modelling, which is part of the multivariate data analysis that will be discussed in Section 2.3.2 was then performed using the NIR and MIR along with testing of various pre-processing methods such as standard normal variate, multiplicative scatter correcting and Savitzky-Golay derivative filtering to determine the amount of adulterants present in ground black pepper samples. The results were by using the pre-processing method of standard normal variate followed by first derivate pre-processing, root mean square error of prediction of 2.7% and ratio of the standard error of prediction to standard deviation of 11.14 for NIR hyperspectral imaging.

Wilde et al. (2019) performed NIR and FTIR spectroscopy to detect for adulterants present in black pepper samples which were papaya seeds, chilli, black pepper husks, pinheads and defatted spent materials. NIR spectra for the samples were acquired within the range of 12000 – 4000 cm^{-1} (833 – 2500 nm) with 32 scans and resolution of 8 cm^{-1}. While for FTIR spectra, they were acquired within the range of 4000 – 400 cm^{-1} (2500 – 25000 nm) with 32 scans and resolution 4 cm^{-1}. A binary classification model using orthogonal PLS discriminant analysis was constructed after a series of pre-processing methods such as derivatives, Savitzky-Golay filter and standard normal variate was made. The results indicated that the receiver operator characteristic curve was 0.98, measure of fit R2 0.93 and 0.83 and prediction ability Q2 0.98 and 0.97 for NIR and FTIR spectroscopy respectively, proving its capability as qualitative and quantitative black pepper adulteration detector.

Lima et al. (2020) utilised NIR spectroscopy to detect and discriminate the adulterants in black pepper and cumin powder samples. The adulterants of interest were starch cassava and corn flour. Targeted models were then constructed based on multiple linear regression and PLS techniques while non-targeted models were based on soft independent modelling of class analogy (SIMCA) and PLS discriminant analysis. The targeted model yielded correlation coefficient of above 0.9 and RMSE ranging from 2.2 to 7. On the other hand, the non-targeted model that utilised SIMCA had high sensitivity in classifying genuine black pepper and cumin powder samples. With that, 13 commercial black pepper samples were tested and 62% of them were found to be adulterated.

Although the detection performance of vibrational spectroscopy techniques is on par with that of the chromatographic technique, vibrational spectroscopy techniques are mostly point-based and can only cover a small area of sample which may not be representative to the sample as a whole. Furthermore, the measurement process could be time-consuming and inefficient considering a large number of samples to be assessed even though it can be performed on all other locations of sample multiple times (Lohumi et al., 2015). Additionally, if the sample was heterogeneous, the measurement of reflectance values would be inconsistent on all locations of the sample (Elmasry et al., 2012; Liu et al., 2017; Lohumi et al., 2015).

2.1.1.3 Other Chemical Analytical Methods

Chandra et al. (2014) applied direct analysis of metabolic profile using real-time mass spectrometry (DART-MS) to detect for adulterants such as Indian *tipali* and Bangla tipali present in black pepper samples and profile for their alkaloids and amides. Through this method, mass spectra, or chemical fingerprints of fruits, leaves and roots for each species were analysed, and it was revealed that there were variations found in the distribution of some common piperamides in these parts for each species in terms of percent ionisation. Additionally, principal components analysis was then applied and it was able to identify distinct clusters belonging to respective species.

Malaysian Pepper Board (MPB), which serves as an authority to aid in the production, development and growth of national pepper industry in Malaysia, had been using chemical analytical techniques in assessing the quality of black pepper products (MPB, 2021). As mentioned above, detailed qualitative and quantitative results could be obtained with the shortcomings of requiring tedious sample preparation, ample specialist training, high cost and having their operations limited within laboratories (Abbas et al., 2018; Bansal et al., 2017; Orrillo et al., 2019). Additionally, extraction and chromatography methods produced chemical waste which may potentially cause environmental damage and hazards (Manley, 2014).

2.1.3 Molecular Techniques

Another type is molecular technique or sequencing-based technique, which is a biological method to discriminate the adulterants from the original product and authenticate the product origin using molecular markers (Bansal et al., 2017; Zhang et al., 2019). Examples of molecular techniques were molecular techniques included:

- Random amplified polymorphic DNA (RAPD):

Khan et al. (2010) displayed that through RAPD, five out of eight decamer oligonucleotide primers yielded species-specific reproducible unique amplicons, which indicated clear discrimination on black pepper and papaya seeds.

- Sequence characterised amplified region (SCAR) marker:

Dhanya et al. (2009) detected the presence of papaya seeds in one of five branded black pepper powder samples using SCAR molecular marker using polymerase chain reaction (PCR) amplification.

Gul et al. (2018) were able to discriminate the papaya seeds and black pepper based on their DNA fingerprints using SCAR markers. By using their respective specific SCAR primer sets, 500- and 750-bp-sized SCAR markers of black pepper and papaya seeds were amplified.

- DNA barcoding:

Sousa et al. (2019) proposed qPCR-based method on plant DNA barcodes using *trnL* and *psbA-trnH* primers, and successfully detected 41% of all 29 black pepper powder samples that were adulterated with papaya seeds, cayenne pepper and maize flour.

Zhang et al. (2019) made use of DNA barcoding on 6 white pepper samples, although unrelated to black pepper but still relevant to adulteration detection, using *ITS2* and *psbA-trnH* sequences and compared them with those in the barcode database. 2 out of 6 samples were found to be adulterated with grass (*Setaria*) and cumin (*Cuminum cyminum*).

Dissanayake et al. (2016) applied DNA barcoding primer pair, *psbA-trnH*, to amplify DNA of the black pepper powder samples. Pure black pepper DNA produced 200 bp bands, while pure chilli and papaya DNA produced 450 bp bands. If the black pepper powder samples were adulterated, their DNA produced 200 bp and 450 bp bands, thus enabling its use as a qualitative adulteration detection tool for black pepper samples.

Parvathy et al. (2014) used three DNA barcoding loci *psbA-trnH*, *rbcL*, *rpoC1* to detect adulteration in black pepper, with *psbA-trnH* was the best in detecting chilli powder present in black pepper samples. 2 out of 9 market black pepper powder samples were found to contain chilli powder adulteration through this method.

Molecular techniques are particularly useful for the detection of genetically modified foods in various food samples and microbial contaminants (Bansal et al., 2017). Although its detection is rapid and low in cost, it requires a strictly controlled environment, specialised training and has its operations limited only within laboratories, making it highly unsuitable to be used in real time (Bansal et al., 2017; Oliveira et al., 2019).

2.1.4 Computer Vision

Another method is computer vision system (CVS) which is an optical method that captures digital images of the whole object within the visible light spectrum to detect adulterants present in samples. Computer vision system has been broadly used in food and agricultural products quality assessment, monitoring, evaluation and assurance (Bhargava and Bansal, 2018; Chen et al., 2019; Di Rosa et al., 2017; Koirala et al., 2019; Patel et al., 2012; Rong et al., 2019; Taheri-Garavand et al., 2019; Vithu and Moses, 2016; Wu and Sun, 2013a; Xu et al., 2017, 2019). CVS is one such method to provide rapid, reliable, objective, cost-effective and highly available adulteration detection as well as reproducibility of data (Taheri-Garavand et al., 2019).

CVS setup is usually made up of charged-coupled device or complementary metal-oxide semiconductor camera, lighting system, background screen, a sample of interest and computer to perform data exploratory, image processing and analysis (Bhargava and Bansal, 2018; Vithu and Moses, 2016). A good lighting system includes proper selection of lighting source, arrangement and geometry as well as appropriate background screen selection are crucial in order to ensure consistency and homogeneity of illuminance over the sample for better image quality (Patel et al., 2012; Vithu and Moses, 2016). The digital images taken are usually within the visible light spectrum similar to how human eyes perceive and can have different colour spaces depending on the nature of the analysis carried out. Colour spaces found in the literature are comprised of red, green and blue (RGB), normalised RGB, XYZ (defined by International Commission on Illumination), HSV, HSL, L*a*b*, L*u*v*, YCrCb, YUV, TSL and I1I2I3 (García-Mateos et al., 2015; Ohta et al., 1980; Shih and Liu, 2005;

Terrillon et al., 2000). After that, the images undergo several levels of processing, namely low-level processing (pre-processing to remove unwanted noise), intermediate-level processing (segmentation to locate and extract the boundaries of objects in images as well as feature extraction) and high-level processing (classification and recognition) (Bhargava and Bansal, 2018; Patel et al., 2012; Taheri-Garavand et al., 2019).

However, one of the main drawbacks of CVS is it has difficulties in detecting the adulterants which are similar in appearance with the original item of interest (Bhargava and Bansal, 2018; Modupalli et al., 2021). Particularly, the papaya seeds powder has similar dark brown colour as black pepper powder. Additionally, this worked well for black pepper samples which are in berry form (P. R. Goswami and K. R. Jain, 2013), but did not perform well if the samples were in powdered form due to stark similarities in terms of colour, texture and size (Bhargava and Bansal, 2018; Modupalli et al., 2021). Another drawback of CVS is in order to evaluate the chemical composition and internal quality characteristics of the sample, spectral information that is outside the visible light spectrum is essential such as certain wavelength range refers to certain functional groups as mentioned in Section 2.1.2, hence leading to the adoption of hyperspectral imaging (Bhargava and Bansal, 2018; Di Rosa et al., 2017; Elmasry et al., 2012; Gowen et al., 2007; Liu et al., 2017; Park and Lu, 2015; Wu and Sun, 2013b).

2.1.5 Hyperspectral Imaging

Hyperspectral imaging (HSI) is one such solution that incorporates computer vision system and vibrational spectroscopy technique, with its setup as shown in Figure 2.2. HSI generates a hypercube or spectral cube which consists of a stack of two-dimensional (x, y) images with varying continuous wavelengths as another dimension (λ). Hypercubes are then stored as band-interleaved-by-pixel (BIP) format which is mostly for microscopic imaging, or band-interleaved-by-line (BIL) format which is the most common mode due to its flexibility and practicability for most industrial applications (Lohumi et al., 2015; Park and Lu, 2015; Wu and Sun, 2013b). The captured hypercube typically has a spectral range covering visible (400 – 780 nm) and near-infrared (780 – 1700 nm) regions as shown in Figure 2.3 (Elmasry et al., 2012; Lohumi et al., 2015; Park and Lu, 2015; Xu et al., 2017). Another reason the HSI is generally favourable for industrial applications is it can be operated on-line, thus accelerating the data acquisition and quality assurance processes (Wu and Sun, 2013c). The use of HSI, which was primarily used in remote sensing, is now prevalent in various fields, notably agricultural (Elmasry et al., 2012; Mahesh et al., 2015; Steinbrener et al., 2019), food quality assurance (Elmasry et al., 2012; Gowen et al., 2007; Qin et al., 2020; Wu and Sun, 2013c), environment (Veraverbeke et al., 2018) and even document analysis (Qureshi et al., 2019).

Vibrational spectroscopy techniques are usually point-based scanning techniques, where all the measured points require averaging for consistency. If the surface of samples is complex and heterogeneous, the measured values are inconsistent and require further time-consuming pre-processing. HSI, on the other hand, can cover the whole sample with push-broom (or line) or area imaging (Gowen et al., 2007; Liu et al., 2017; Lohumi et al., 2015). Hypercubes are taken in either reflectance, transmittance or interactance modes, as shown in Figure 2.4 (Elmasry et al., 2012; Jha, 2016; Lohumi et al., 2015; Wu and Sun, 2013b). Reflectance mode, being the most common mode, is where the light source is reflected to the detector from the surface of the sample. Transmittance mode is where the light source passes through the sample and to the detector, while interactance mode is a combination of reflectance and transmittance, but requires a light barrier to eliminate the interference due to specular reflection (Lohumi et al., 2015; Wu and Sun, 2013b). Reflectance mode is greatly favoured particularly in quality control of food and agricultural products due to its compatibility with external quality features of food samples (Lohumi et al., 2015; Wu and Sun, 2013b).

The size of hypercubes is substantial considering the rich spectral information with continuous bands of wavelengths they carry for each pixel. More importantly, due to high dimensionality (up to hundreds of dimensions, λ), the need for data to undergo further processing to produce an interpretable model with meaningful qualitative and quantitative

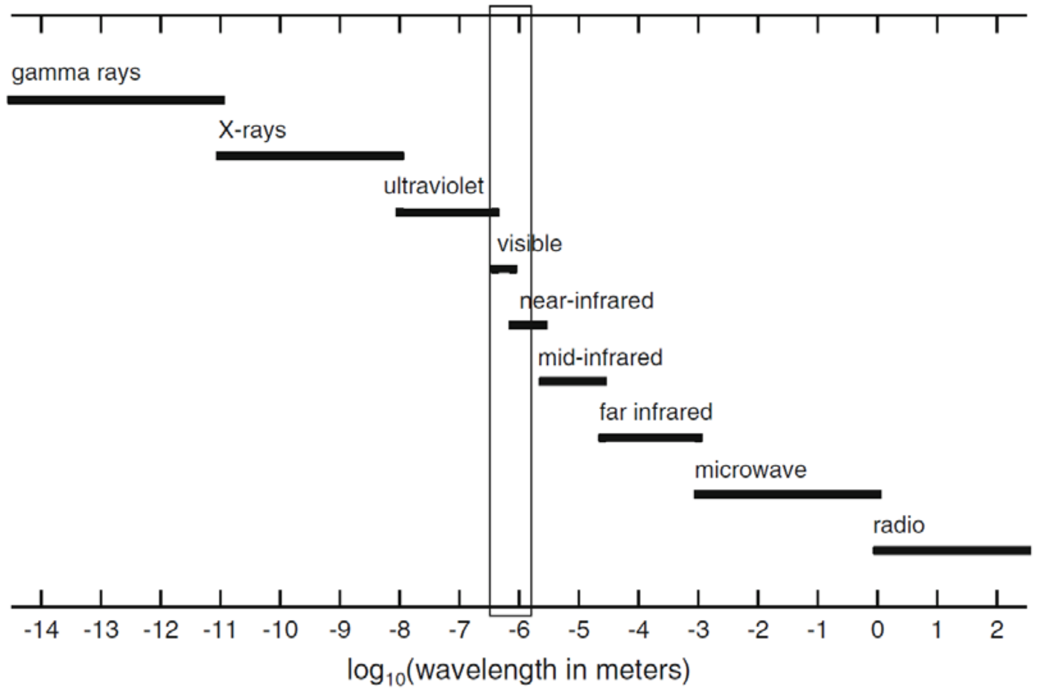


Figure 2.3: Electromagnetic spectrum, with the box region indicating the range used for hyperspectral imaging (400 – 1700 nm) (Park and Lu, 2015)

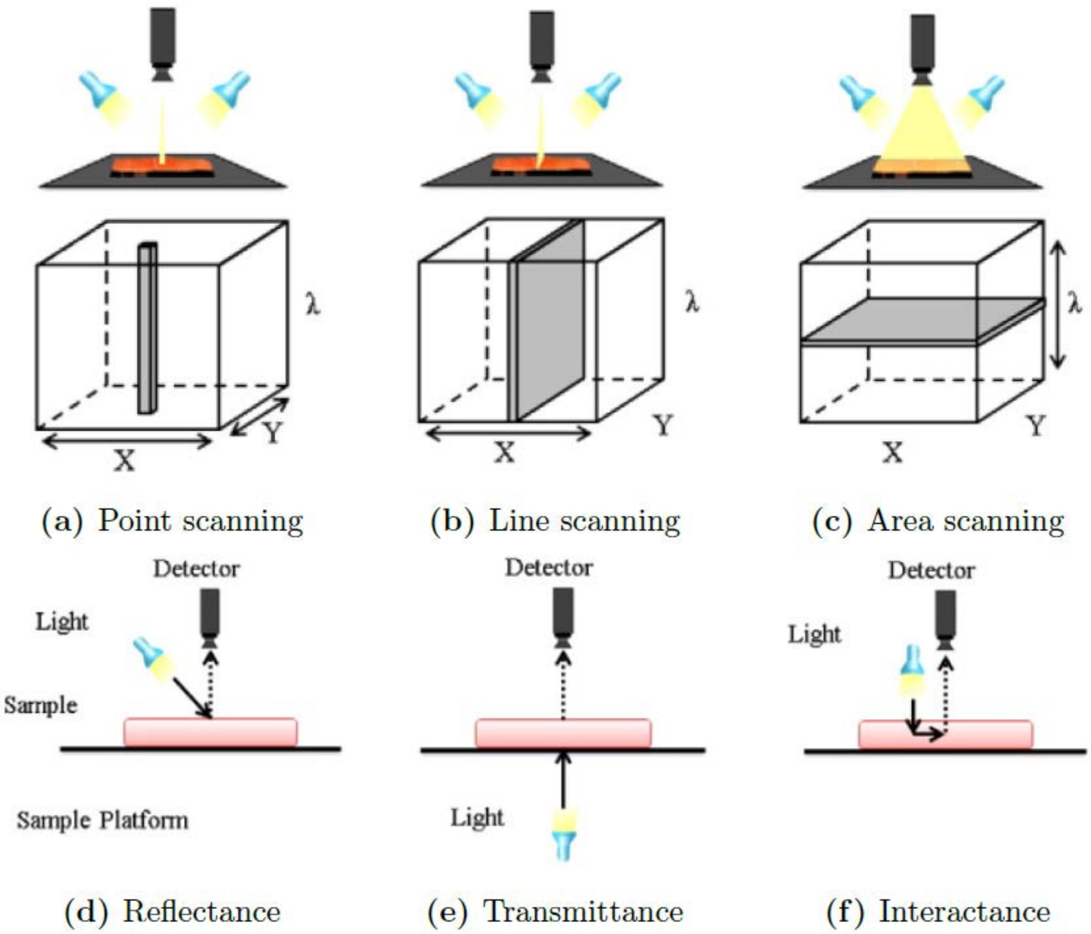


Figure 2.4: Different approaches and modes of scanning methods (Wu and Sun, 2013b)

results necessitates the use of multivariate data analysis or chemometrics, which will be discussed in Section 2.3 (Elmasry et al., 2012; Lohumi et al., 2015). In most analyses, the model uses mean spectra data from the hypercubes since they are rapid, convenient and informative enough to yield meaningful outputs. Pre-processing is usually performed on these spectral data sets. This is to ensure that the random noise from the ambient environment can be eliminated, the amount of variation in the data can be reduced for better generalisation of the model and, baseline and scattering effects can be scaled and corrected properly (Elmasry et al., 2012; Goodfellow et al., 2016; Yang et al., 2019). Examples of pre-processing methods include multiplicative scatter correction, standard normal variate (SNV), normalisation and Savitzky-Golay (SG) derivative filtering (Dhanoa et al., 1994; Lohumi et al., 2015; Wu and Sun, 2013b).

There were previous works on detecting the adulterants in black pepper samples using HSI, including the variation of different pre-processing methods on the detection capability. For example, as mentioned in Section 2.1.2, McGoverin et al. (2012) utilised NIR HSI and spectral pre-processing methods such as SNV, multiplicative scatter correcting and SG derivative filtering to detect for adulterants present in ground black pepper samples which were buckwheat and millet. Orrillo et al. (2019) used HSI system with the spectral range of 900-1710 nm and 5 nm intervals which yields a total of 159 bands to detect different proportions of adulteration of papaya seed powder present in black pepper samples. Similar common spectral pre-processing methods such as multiplicative scatter correction, SNV, SG derivatives were performed to reduce random noise and correct variations in intensity signal from the detector. Principal components analysis (PCA) and soft independent modelling of class analogy (SIMCA) were used in the classification of black pepper and papaya seed samples. From there, 90% prediction accuracy was attained. Furthermore, PLS regression was applied for the quantitative determination of adulteration of papaya seeds in black pepper samples. It attained predictive RMSE of 2.51 and coefficient of determination of 0.93.

2.2 Geographical Origin Classification

Black pepper products from various countries of origin have different compositions such as volatile oil, non-volatile oil, piperine and oleoresin contents (Ravindran and Kallupurackal, 2001). Black pepper from most pepper producing countries contained around 4.6% piperine, 10.3% oleoresin and 3.8% volatile oil (Ravindran and Kallupurackal, 2001). It was reported that Sarawak black pepper contains approximately 3.5% piperine, 11% oleoresin, 2.8% volatile oil and 7.9% non-volatile oil (Johny et al., 2020). It was stated that different geographical origins of black pepper products brought considerable influence on their composition, bioactivities, quality and subsequently pricing (Li et al., 2020; Liang et al., 2021; Oliveira et al., 2019). Hence, it is essential to trace the geographical origin of black pepper products so that appropriate actions can be taken to ensure the excellence of quality of black pepper for that identified region. In identifying the geographical origin of black pepper, several handful methods were found in the literature.

Zhang et al. (2015) utilised GC to profile the black pepper extracts and classify their geographical origins which are Vietnam, Brazil, Indonesia and India using PLS discriminant analysis, resulting in 100% classification rate using the best combinations of polynomial order on baseline correction and degree of the root for logarithmic data transformation.

Hu et al. (2018) tested a total of 150 pure black pepper powder samples from different origins such as India, China, Malaysia and Vietnam using diffuse reactance mid-infrared Fourier Transform spectroscopy. They were adulterated with proportions of sorghum and Sichuan pepper and then detected using Fourier transform infrared spectrometer with a deuterated triglycine sulfate detector. Two chemometrical methods on identifying the geographical origins, namely genetic algorithm optimised SVM and PLS discriminant analysis, achieved 100% recognition rate for pure black pepper powder samples, and 96% recognition rate for adulterated black pepper powder samples.

Mercer et al. (2019), on the other hand, utilised high-resolution gas chromatography mass spectrometry to record the profiles of volatile organic compounds present in 252 black pepper samples from Malaysia and India, and classify their geographical origins using PCA and fold change analysis. The results indicated that different compounds that originated from Malaysia or India were found, for example, gibbenellic acid was only found in Malaysia black pepper powder samples. Furthermore, two distinct clusters were identified from the PCA score plot and fold change analysis plot.

Li et al. (2020) applied various chemical analytical techniques such as gas chromatography mass spectrometry, systematic cluster analysis, chemical assaying and principal components analysis to investigate if different origins of black and white peppers influence chemical and biochemical activities or not. It was demonstrated that visually, essential oil in Hainan white pepper is colourless, Guangdong black pepper is dark green and other regions are yellow-green. There was also a difference in essential oil compositions, antioxidant and antifungal activities for pepper samples of different regions.

Liang et al. (2021) applied GC-MS, liquid chromatography-MS, thermal desorption DART-MS on seventeen black pepper powder samples to classify their geographical origins which were Indonesia, Brazil, India and Vietnam. Partial least square – discriminant analysis was used to classify the origins, followed by evaluation of total ion mass spectrum data profiles to assess the classification rates. It was demonstrated that by using cubic root data transformation during data pre-processing stage, the classification rate was best achieved by thermal desorption DART-MS method with 97.0 ± 0.3% correctly classified origins.

Rivera-Pérez et al. (2021) made use of metabolomics approach based on GC-Orbitrap-high resolution MS along with chemometrical methods which included principal component analysis (PCA) and orthogonal partial least square – discriminant analysis (OPLS-DA) to identify the geographical origins of black pepper powder samples. PCA and OPLS-DA models were built using 60 black pepper powder fingerprint data, where 80% of data went to training set while 20% went to prediction set. For PCA model, 91% of total variance of data was explained and predictive ability of model, Q^2 was 0.818. While for OPLS-DA mode, 100% correct classification rate was achieved on the prediction set along with Q^2 of 0.977.

There were previous works on identifying the geographical origins of various food and agricultural products using HSI with multivariate data analysis (Choi et al., 2020; Guo et al., 2013; Kamruzzaman et al., 2014; Ke et al., 2020; Wang et al., 2019). However, limited works could be found on classifying the geographical origins of black pepper using HSI, which will be one of the core objectives of this research project.

2.3 Multivariate Data Analysis

Vibrational spectroscopy techniques, including HSI, carry an immense volume of information but at the cost of having inherently high dimensionality and redundancies. It is therefore essential to reduce the dimensionality while at the same time trying to extract as much informative data or features as possible. Multivariate data analysis serves to perform the aforementioned spectral processing so that the data can be decomposed to yield interpretable model by establishing the relationship between the data and the desired attributes of test data (Amigo et al., 2013; Liu et al., 2014; Wu and Sun, 2013b). Chemometrics is part of the multivariate data analysis solutions which is geared for determination of chemical analytical properties (Khan et al., 2020). In most vibrational spectroscopy applications, chemometrics is broadly applied to assess and predict the quality of most food and agricultural products such as moisture content, mineral content, volatile oil % and so on (Elmasry et al., 2012; Khan et al., 2020).

Qualitative classification and quantitative regression can be performed on the data or hypercubes. In both cases, a model is created and learns from the data so that its parameters can be tuned and optimised to establish the best possible weightings that can explain the aforementioned relationship, which is known as machine learning. Some examples of multivariate data analysis techniques were multiple linear regression, principal components analysis (PCA), k-means clustering, partial least square (PLS), support vector machine (SVM), support vector regression (SVR), k-nearest neighbour (KNN), artificial neural network (ANN) (Amigo et al., 2013; Kucharska-Ambrożej and Karpinska, 2020; Wu and Sun, 2013b).

Qualitative classification serves to provide prediction where the results are belonging to what kind of class. The model is either created in a supervised or unsupervised manner where the former learns the data with appropriate labelling of their respective known classes while the latter learns the patterns present in the data (Wu and Sun, 2013b). From there, patterns of results were observed and analysed to Examples of methods used in qualitative classification are PCA, SVM, k-means clustering to separate the dataset into *k* number of clusters where each data point belongs to the respective cluster with the minimum distance to the cluster centroid, linear discriminant analysis to classify objects by finding the optimal boundary that discriminates the classes with maximum between-class variance and minimum within-class variance (Elmasry et al., 2012; Wu and Sun, 2013b).

On the other hand, quantitative regression provides numeric prediction from a linear or non-linear relationship between the data and desired attributes. A linear relationship is usually derived from the commonly used methods, such as multiple linear regression, principal component regression, SVR and PLS (Elmasry et al., 2012; Saha and Manickavasagan, 2021; Wu and Sun, 2013b). While for non-linear relationship, commonly used methods are SVR with radial basis or polynomial kernel function and ANN (Kucharska-Ambrożej and Karpinska, 2020; Saha and Manickavasagan, 2021; Yang et al., 2019). Particularly, ANN is given the most attention in recent machine learning research due to its promising breakthrough in its predictive capabilities and application in various fields. ANN paves the way to another branch of modelling known as deep neural network or deep learning which will be discussed in Section 2.4.

The performance criterion for classification is usually accuracy, precision, recall, F1 score and others such as receiver operating characteristic which is suitable for binary classification (Elmasry et al., 2012; Goodfellow et al., 2016). Accuracy is a measure of number of correctly predicted class for all samples over total number of samples, and simply calculated as follows:

$$Accuracy\ (\%) = \frac{1}{n}\sum_{i}^{n} 1(\hat{y} = y)_i \times 100 \tag{2.1}$$

where for n input data or total number of samples, $1(\hat{y} = y)_i$ is the indicator function which compares the predicted label, $\hat{y}$ of the sample i with the ground truth, y label, and hence returns 1 if true and 0 if false for each sample i.

For classification of multiple classes, confusion matrix is used which is a measure of accuracy on determining how much correctly predicted classes of samples are made based on the actual classes of those samples. Confusion matrix allows visualisation and insight on how well the classification of all classes is performed.

The performance criterion for regression are usually root mean squared error (RMSE) which compares predicted and true values of a dependent variable, with lower RMSE value indicating the predicted value is closer to the true value, and coefficient of determination (R^2), which determines the goodness of fit of the model on the data (Elmasry et al., 2012). RMSE and R^2 are calculated as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i}^{n}(\hat{y} - y)_i^2} \tag{2.2}$$

$$R^2 = 1 - \frac{\sum_{i}^{n}(\hat{y}_i - y_i)^2}{\sum_{i}^{n}(\hat{y}_i - \bar{y})^2} \tag{2.3}$$

where the predicted value, $\hat{y}$ is compared to the ground truth, y value in each sample i for n input data.

Mean absolute percentage error (MAPE) is another performance criterion which calculates how much the absolute error of predicted values deviates from the true values in average, and calculated as follows:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{\hat{y}_i}\right| \times 100 \tag{2.4}$$

By using these metrics above, predictions from both classification and regression can be optimised and subsequently, meaningful predictive results can be obtained. The multivariate data analysis techniques which are more relevant to the research objectives are PCA, PLS, SVM and deep learning. The following sections elaborate these techniques in details and review them thoroughly.

2.3.1 Principal Components Analysis

Principal components analysis (PCA) is one of the most encountered methods in most multivariate data analysis. It performs dimensionality reduction and feature selection by decomposing the data into user-defined principal components that explains the highest variability of the data. PCA mainly linearly transforms the data with correlated variables into variables or principal components which are mutually uncorrelated (Goodfellow et al., 2016). PCA is mainly driven by singular value decomposition (SVD), where it factorised any matrix into singular vectors and singular values. SVD decomposes any matrix $\boldsymbol{A}$ of $m \times n$ as follows:

$$\boldsymbol{A} = U\Sigma V^{\mathrm{T}} \tag{2.5}$$

where U and V are orthogonal matrices of $m \times m$ and $n \times n$ respectively, Σ = diagonal matrix of $m \times n$. U constitutes eigenvectors of $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ while V has eigenvectors of $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$. The diagonal matrix, Σ is the covariance matrix containing all the singular values so that data with highest variances can be extracted. For PCA, m mainly refers to the number of samples in the data, n refers to the number of features or variables in the data. It chooses the most significant terms or variables (known as principal components) based on SVD where the eigenvectors of first few highest singular values are selected. In summary, PCA separates the data into score, U $(m \times k)$ and loading, V $(n \times k)$ matrices based on principal components where k refers to the number of principal components as the goal is to decompose the data and reconstruct it using principal components which are fewer in number (Goodfellow et al., 2016; Kherif and Latypova, 2020). The loading and score matrices contain the maximum variability of data to each other and their corresponding coefficients respectively, as well as the residual matrix (Amigo et al., 2013; Elmasry et al., 2012; Goodfellow et al., 2016; Jha, 2016).

PCA is used to identify possible outliers present in the data as well as possible patterns or clusters which categorise the data into respective classes, for example clusters of black pepper samples belonging to respective geographical origins, using score and loading plots. Explained variance in PCA serves as an indicator to understand how much variance of data is accounted by the principal components while the remaining data corresponds to noise (Kherif and Latypova, 2020). For the first few principal components, the more the explained variance, the more these principal components can represent the distribution of the data, hence the less noisy the data is.

2.3.2 Partial Least Square

PLS is commonly found in most regression models. It projects the independent and dependent variables into a latent structure with orthogonal factors or latent variables so that the covariance between independent and dependent variables can be maximised, which solves the main drawback of PCA (Elmasry et al., 2012; Orrillo et al., 2019). The transformation is performed according to the equations as follows (Ge and Song, 2010):

$$X = \boldsymbol{T}^{\mathrm{T}} P + \mathrm{E} \tag{2.6}$$

$$y = \boldsymbol{T}\mathrm{q} + \mathrm{f} \tag{2.7}$$

Given a dataset with independent variables, X and dependent variable, y, PLS decomposes them into score matrix, $\boldsymbol{T}$, loading matrix and vector, P and q, weight matrix, $\boldsymbol{W}$ along with residual matrix and vector, E and f. The weight matrix is used to predict the new value of the output dependent variable using a new data point, $\mathrm{x}_{\mathrm{new}}$ in the following equation:

$$\hat{y}_{new} = \mathrm{x}_{\mathrm{new}} \boldsymbol{W} (\boldsymbol{P}^{\mathrm{T}} \boldsymbol{W})^{-1} \mathrm{q} \tag{2.8}$$

PLS is generally more favourable than PCA because PLS considers relationship with response variable while PCA considers the variability of only the input variables (Elmasry et al., 2012; Orrillo et al., 2019). Although PLS is mostly used in regression, it can be applied to classification with usage of another technique known as discriminant analysis. Based on the threshold of predicted values for each sample, the class is assigned accordingly and compared with true label of respective samples.

While PLS provides sufficient solution in the modelling for HSI data, there exists a possibility that the data exhibits non-linearity where PLS may fail to approximate the data to yield meaningful predictions (Li et al., 2019; Petersson et al., 2016; Yang et al., 2019). Non-linear modelling provides a better and more accurate analysis than linear models. Support vector machine and artificial neural network are notable examples of non-linear modelling.

2.3.3 Support Vector Machine

Support vector machine (SVM) is another commonly used supervised learning model which serves to find the optimal decision boundary to distinguish between clusters of two classes (Cortes and Vapnik, 1995; Elmasry et al., 2012; Goodfellow et al., 2016; Saha and Manickavasagan, 2021; Suthaharan, 2016). SVM uses kernel trick to evaluate the following output for prediction (Goodfellow et al., 2016):

$$y = \sum_i \alpha_i K\left(\boldsymbol{x}, \boldsymbol{x}^{(i)}\right) + b \tag{2.9}$$

where $K\left(\boldsymbol{x}, \boldsymbol{x}^{(i)}\right)$ is the kernel function which transforms non-linear data into linear feature space, $\boldsymbol{x}$ and $\boldsymbol{x}^{(i)}$ are input training matrix and input training vector for sample i, α_i is vector of coefficients and b is the intercept of the model. Examples of kernel function are radial basis function or Gaussian kernel, polynomial kernel, linear kernel, sigmoid kernel (Saha and Manickavasagan, 2021). The widely used kernel function is Gaussian kernel function, which can be expressed as follows (Scholkopf et al., 1997):

$$K\left(\boldsymbol{x}, \boldsymbol{x}^{(i)}\right) = \exp\left(-\frac{\left\|\boldsymbol{x}-\boldsymbol{x}^{(i)}\right\|^2}{2\sigma^2}\right) \tag{2.10}$$

where σ^2 is variance, and the expression inside exponential function is calculation of Euclidean distance between $\boldsymbol{x}$ and $\boldsymbol{x}^{(i)}$.

The main objective of SVM is to find the optimal decision boundary in the form of hyperplane (for linear data) or feature space (for non-linear data) which yields the maximum margin from vectors supported by kernel function shown in Equation (2.10) or support vectors between the feature space and the data belonging to certain class or label. SVM is mainly utilised in classification cases, while for regression cases, support vector regression (SVR) is applied instead. SVR works similar to SVM with the difference being SVR finds the feature space or hyperplane with permissible error margin, where this error margin is a parameter to be optimised, and hence providing continuous range of values in its prediction instead of discrete classes.

2.4 Deep Learning

Multivariate data analysis has been a primary solution in developing an interpretable model to present functional predictions. Data pre-processing becomes essential as aforementioned to eliminate random noise or artefacts for robustness and reliability of the model (Elmasry et al., 2012). Feature selection can then proceed but is usually performed manually or through human intervention. This requires prior knowledge to enable modelling to function properly and is especially prone to errors in losing informative data (Li et al., 2017; Yang et al., 2019). Additionally, it becomes drastically harder to manage the staggeringly high number of variables of input data as general modelling functions are insufficient to explain that input data statistically (Goodfellow et al., 2016).

Deep learning is one such solution that overcomes some of these shortcomings present in the most multivariate data analysis solutions. What makes deep learning appealing in modern modelling solutions is because deep learning serves like a black box model that relies solely on data, easing the whole modelling process. Hence, the more the data is provided, the better the predictive performance of the model (Yang et al., 2019). In recent years, deep learning has become a sensation in machine learning and artificial intelligence fields recently due to abundance of data and availability of hardware with high computational power (LeCun et al., 2015; Shrestha and Mahmood, 2019). To understand how deep learning performs extremely well, it is important to understand how it functions beneath the hood.

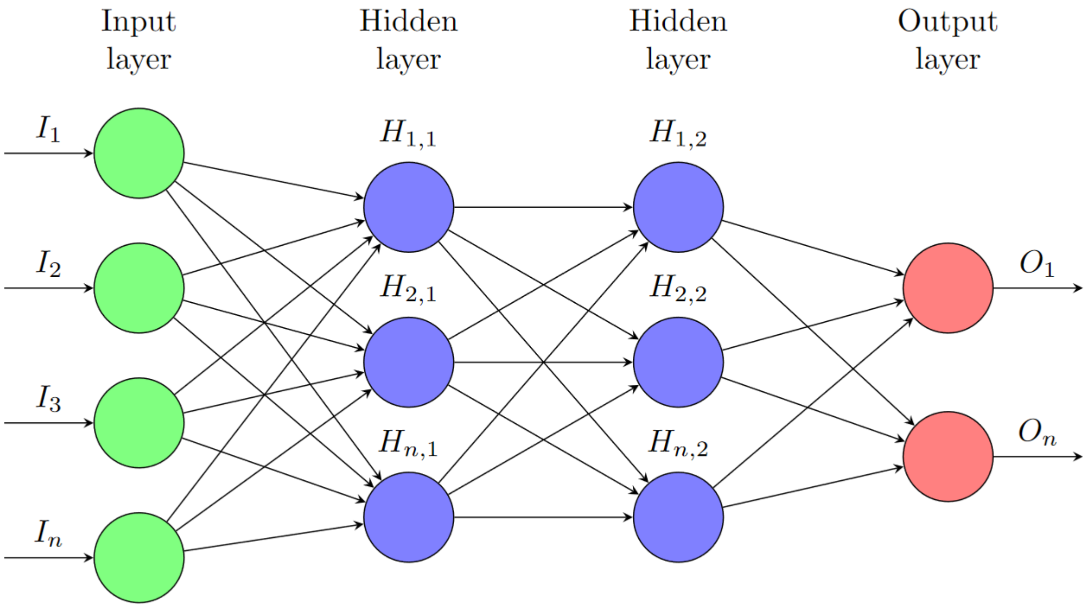


Figure 2.5: Feedforward neural network

2.4.1 Introduction to Deep Learning

Deep learning, or deep neural network, is based on artificial neural network (ANN), which loosely resembles human brain and nervous system that simulates its behaviour for learning and prediction purposes (Emmert-Streib et al., 2020; Goodfellow et al., 2016; Khan et al., 2019; Shrestha and Mahmood, 2019; Wu and Sun, 2013b). A typical neural network is made up of three main layers filled with artificial neurons, namely input, hidden, and output layers. The input feed flows from the input layer through the hidden layer to the output layer in one direction. Because no loopbacks of information are involved, this is known as feedforward neural network, which can be represented in Figure 2.5 (Emmert-Streib et al., 2020; Goodfellow et al., 2016; Shrestha and Mahmood, 2019).

Neurons are simple computational elements that utilise connectionism approach to process information (Goodfellow et al., 2016). The neurons in the hidden layer mainly serve to compute the features of the input data, which is known as representation learning. Each neuron is expressed with the following mathematical expression:

$$y_{out} = f(\sum_{i=1}^{n} w_i \mathrm{x}_i + b) \quad (2.11)$$

For each output neuron y_{out}, there is a ranking of importance for all input neurons, x_i which is represented by their respective weights w_i. The output neuron is fully connected to all n inputs of x of the previous layer as well as a bias constant, b and transforms the sum using an activation function. Activation function is usually non-linear in the neural network since linear functions have limitations in describing the complex relationship of any input data (Emmert-Streib et al., 2020; Goodfellow et al., 2016). The commonly encountered activation functions include hyperbolic tangent, sigmoid and rectified linear unit (ReLU). Figure 2.6 shows sigmoid and ReLU activation functions.

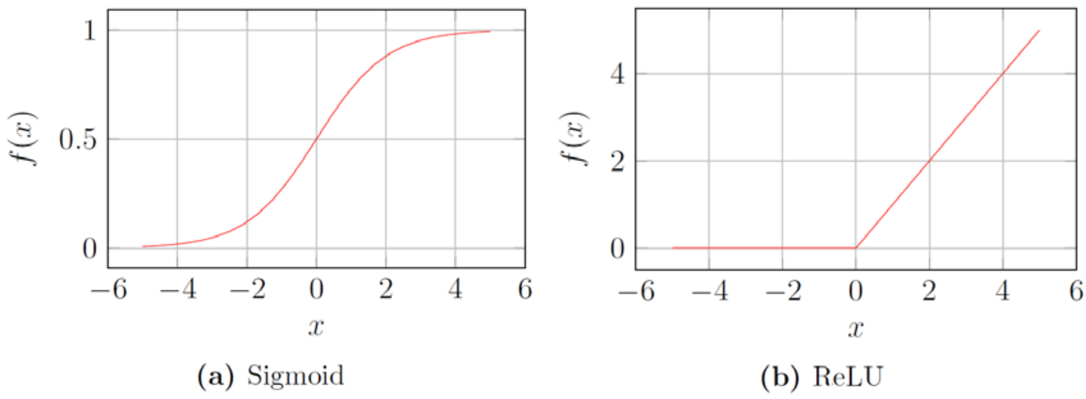


Figure 2.6: Common activation functions used in deep learning

Sigmoid function is expressed as follows:

$$f(x) = \frac{1}{1+e^{-x}}\ , 0 \leq f(x) \leq 1 \tag{2.12}$$

While for ReLU function:

$$f(x) = \max\{0, x\} \tag{2.13}$$

Tweaking the parameters which are weights and biases is the core objective of training a neural network to approximate a function that is suitable in explaining the relationship between target and input variables when given input data. To verify if the parameters are tweaked properly, cost or loss function serving as a performance metric is required to evaluate the predictive performance of the neural network. Examples of loss functions include RMSE and cross-entropy loss (Goodfellow et al., 2016). RMSE can be referred to Equation (2.2), while the cross-entropy loss which is used to compare the probabilities of predicted and true class mainly in the classification of multiple classes, can be defined as follows:

$$L(\text{cross entropy}) = \frac{1}{n}\left(-\sum_{i}^{n} y_i \cdot \log(\hat{y}_i)\right) \tag{2.14}$$

where the predicted value, $\hat{y}$ is compared to the ground truth, y value in each sample i for n input data.

The derivatives of loss function with respect to weights and biases are the targets to be optimised in order to attain the minimum loss function. Back-propagation is hence performed to compute these derivatives from output layer back to input layer using chain rule differentiation due to its high efficiency to be executed in various machines (Goodfellow et al., 2016; LeCun et al., 2015). After these derivatives are computed, optimisation algorithms are then performed to enable the neural network to learn. The learning rate is another parameter that shall be tweaked to optimise the weights and biases (Goodfellow et al., 2016; Shrestha and Mahmood, 2019). Stochastic gradient descent (SGD) is one of the most commonly used optimisers to adjust learning rate so that the global minimum of the loss function can be attained (Emmert-Streib et al., 2020; Goodfellow et al., 2016). To get new weight, w (applicable to bias too), it can be done by adjusting the learning rate, η according to the derivative of loss function with respect to current weight, $\frac{\partial L}{\partial w}$ as follows:

$$w \leftarrow w - \eta \frac{\partial L}{\partial w} \tag{2.15}$$

Normal gradient descent takes only one sample to update gradient at one time, which is slow and may be unable to attain the minimum cost function. SGD instead takes a minibatch of samples to optimise weights and biases, and thus speeds up the learning process of neural network. One notable drawback for SGD is its fixed learning rate may increase the time taken for the neural network to finish learning, especially if the learning rate is set too small or large. To mitigate this issue, various optimisers that apply adaptive learning rate are introduced in the past, for instance, AdaGrad, RMSProp and Adam. Adam optimiser appears to be a more popular choice on most occasions due to its faster convergence at a cost of more parameters to tune (Goodfellow et al., 2016).

The main data set to be fed for the neural network to learn is known as the training data set. Generalisation is another main criterion to dictate if the model can predict new unseen input data correctly or not. While the neural network is trained using training data set, a validation data set is often introduced to verify if the new weights and biases result in good generalisation ability of neural network or not. High training accuracy or low training RMSE does not necessarily indicate the model has better predictive performance on validation data set. The whole process is repeated until all the weights and biases are stabilised or convergence is reached. After that, testing data set which is not involved in the training of neural network is used to ultimately decide the generalisation ability of the trained neural network. A neural network with poor generalisation ability tends to underfit (where model still has a high error on training data set) or overfit (where model performs well on training data set but worse on testing data set) the data (Goodfellow et al., 2016; Shrestha and Mahmood, 2019).

This whole process can be simplified as follows:

1. Feed batches of input data using training data set into neural network and forward to output prediction
2. Compare the prediction value with ground truth value
3. Perform back-propagation to compute gradients of that compared result with respect to all weights and biases of all neurons, followed by optimisation algorithms such as SGD or Adam optimiser to obtain new weights and biases
4. Use validation data set to determine if new weights and biases yield good prediction value or not
5. Repeat the steps 1 - 4 until convergence is reached or the predicted value using validation data set is nearly identical with ground truth value using tolerance

The neural network with one hidden layer is typically known as shallow network, while deep neural network generally refers to the neural network with more than two hidden layers (Emmert-Streib et al., 2020). Such this deep neural network is also known as multilayer perceptron. It was observed that the higher the depth of neural network or the number of hidden layers in the neural network, the more complexity of the features it can capture from the input data (Fan et al., 2019; Goodfellow et al., 2016). Hence, deep neural network hugely benefits the more data is provided. The recent advancement and such success of deep learning can be attributed to the utilisation of graphical processing units and abundance of data as mentioned above (Krizhevsky et al., 2012; Shrestha and Mahmood, 2019).

Deep neural networks still have some limitations. They require a certain amount of data to be functional, which may take a longer period to finish training than most conventional machine learning algorithms (Saha and Manickavasagan, 2021; Shrestha and Mahmood, 2019). Next, it is noted that they tend to have extremely low or high gradients known as vanishing or exploding gradients where convergence cannot be achieved, prompting the use of proper activation functions such as ReLU instead of sigmoid because sigmoid is only sensitive to input when it approaches 0 (Goodfellow et al., 2016; Shrestha and Mahmood, 2019). If the input data has a huge dimension size, it will result in an exorbitant number of parameters of deep neural networks to be tuned, further increasing the time taken to train the model. Hence, variants of deep neural networks are introduced to improve the training process.

2.4.2 Types of Deep Neural Networks

There are various types or architectures of deep neural network: convolution neural network (CNN), recurrent neural network, autoencoder, restricted Boltzmann machine, long short-term memory, deep belief networks and so on (Fan et al., 2019; Goodfellow et al., 2016; Modi, 2018; Shrestha and Mahmood, 2019; Signoroni et al., 2019; Yang et al., 2019). In most spectral analysis cases that are targeting at the food and agriculture, the commonly encountered types of deep neural networks are CNN and autoencoders (Saha and Manickavasagan, 2021; Yang et al., 2019; L. Zhou et al., 2019). Recurrent neural network and long short-term memory structures are not widely encountered compared to other structures due to presence of memory component which is more appropriate for speech or video data (Emmert-Streib et al., 2020).

2.4.2.1 Convolutional Neural Network

Convolutional neural network (CNN) is another variant of feedforward neural network with the difference being the additional layers that perform convolution and pooling (Emmert-Streib et al., 2020; Khan et al., 2019; Modi, 2018; Saha and Manickavasagan, 2021; Shrestha and Mahmood, 2019; Yang et al., 2019). Figure 2.7 depicts the typical structure of CNN. CNN became famous thanks to the creation of AlexNet (Krizhevsky et al., 2012; LeCun et al., 2015). Its usage is currently prevalent in computer vision, face and video recognition, natural language processing, object recognition and so on (Khan et al., 2019; Shrestha and Mahmood, 2019). There were variations of implementation of CNN architectures other than AlexNet, which included VGG, ResNet, GoogLeNet, Inception and many more (Khan et al., 2019; Shrestha and Mahmood, 2019; Yang et al., 2019).

Convolution operation in CNN extracts locally correlated information from the inputs using convolutional filter. The extracted features are considered spatially invariant, making it suitable for spatial data such as images. Assuming a two-dimensional image I input, its convolution output or feature map S using a two-dimensional kernel K will be as follows (Goodfellow et al., 2016):

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(m,n)\, K(i-m, j-n) \quad (2.16)$$

where i and j are the current coordinates being referred at a time, m and n are the range of values within i and j respectively bounded by the kernel size.

Convolutional filter is usually made up of several convolutional kernels. Convolutional kernel contains values to transform localised input within the kernel window of defined size into feature value. The kernel is later slid throughout the whole input to output a feature map. The kernel size determines how much the details to be extracted by the kernel, the smaller the kernel size, the finer the details can be extracted by the kernel. Stride controls the step size of kernel sliding the input and influences the feature map dimensions. Additionally, padding can be necessary if one wants the feature map dimensions to be as same as input dimensions, where the sides of input are filled with zero values. The feature map output dimension, O is thus as follows:

$$O = \frac{I+2P-K}{S} + 1 \quad (2.17)$$

where I = input size, P = padding size, K = kernel size and S = stride.

Next, pooling layer transforms localised input within specified size into another value. The commonly encountered pooling types are max pooling and average pooling (Emmert-Streib et al., 2020; Goodfellow et al., 2016). Max pooling chooses the maximum value among all the localised input values while average pooling obtains the average of all the localised input values. Pooling also serves as dimensionality reducing tool, thus drastically reduces the

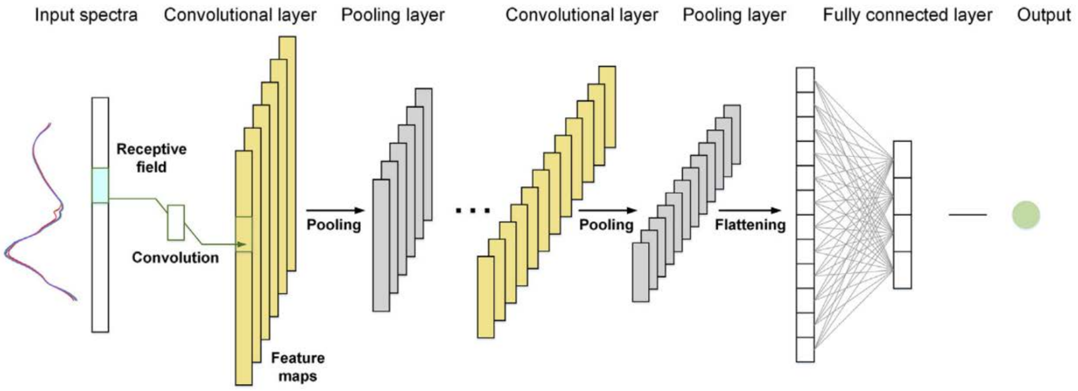


Figure 2.7: Convolutional neural network (Goodfellow et al., 2016)

number of parameters, subsequently the network size and time taken to learn the input data (Emmert-Streib et al., 2020; Goodfellow et al., 2016; Khan et al., 2019; Saha and Manickavasagan, 2021; Shrestha and Mahmood, 2019; Yang et al., 2019).

Dropout is a method that deactivates some neurons of targeted hidden layer in a probabilistic manner. This decreases the likelihood of the neural network to overfit the data. Dropouts are often used in CNN since the extracted features from convolution and pooling operations are fully captured by the fully connected hidden layers, causing the network to overfit easily (Goodfellow et al., 2016; Saha and Manickavasagan, 2021).

CNN is more effective in dealing with multi-dimensional data than typical deep feedforward neural network due to its capability to extract local features which are invariant in their locations, hence the extracted features are still similar no matter how the variations of data are made (Al-Sarayreh et al., 2018; Emmert-Streib et al., 2020; Goodfellow et al., 2016). Subsequently, this reduces the dimensionality, the number of parameters, network size and time taken to train the network. However, CNN requires large amount of data to extract sizeable number of features and its pooling operation may lose valuable information due to dimensionality reduction (Goodfellow et al., 2016).

2.4.2.2 Autoencoders

Autoencoder is another part of deep neural network which has a structure similar to the feedforward neural network. The main purpose of the autoencoder is to reconstruct the input data using the representative features from dimensionality reduction (Goodfellow et al., 2016; Shrestha and Mahmood, 2019). They primarily contain two components, namely encoder and decoder. The encoder reduces the dimensionality similar to PCA and outputs the non-linear representative features instead of linear. On the other hand, the decoder reconstructs the input data using these reduced features (Shrestha and Mahmood, 2019; Yang et al., 2019). Autoencoder usually performs unsupervised learning as they serve to reconstruct the input data just by extracting and learning the features presented by the input data without labels or ground truth values (Shrestha and Mahmood, 2019). Figure 2.8 illustrates the typical structure of stacked autoencoders, where the structure contains multiple layers of autoencoders. It is desirable to have higher depth because the nodes in each layer can approximate the inputs or feature outputs well while reducing their computational cost (Goodfellow et al., 2016). As shown in Figure 2.8, the left half of the structure is the encoder part which typically has several layers with decreasing number of nodes for each layer. The number of nodes is not set to be higher as autoencoder is expected to extract unique representative features and reconstruct the input, not copy the inputs perfectly. On the other hand, the decoder part constitutes the right

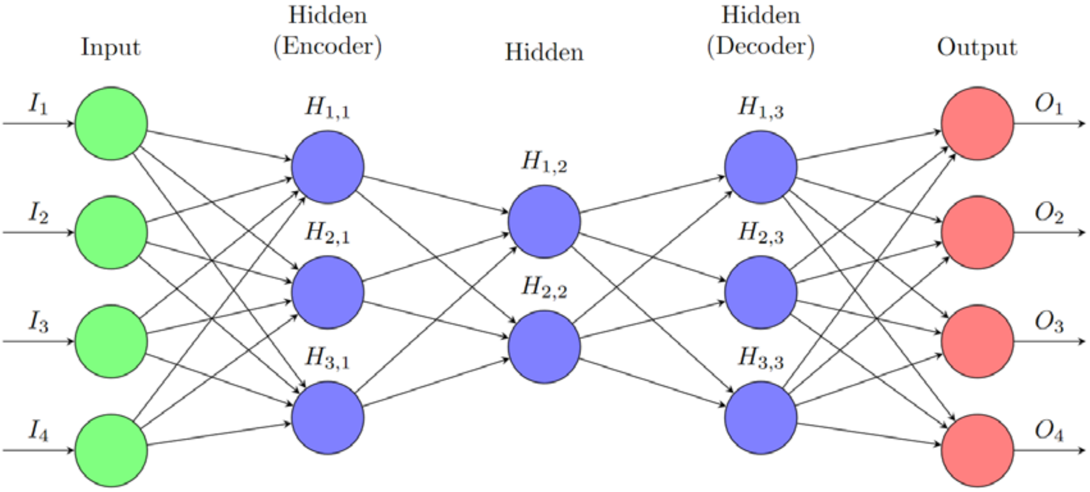


Figure 2.8: Autoencoder

half of the structure, which is typically a mirror image of the left half of the structure. It has numerous layers with an increasing number of nodes until the final output containing the number of nodes as same as the number of input nodes.

Most previous works that applied autoencoders for HSI in food and agricultural applications removed the decoder part to perform classification or regression. The intermediate layer with the least number of nodes in the encoder part is connected to another fully-connected layer where the nodes in fully-connected layer learn the representative features. For example, Yu et al. (2018) pre-trained 7-layer stacked autoencoders to obtain 10 nodes in the 4th layer, removed the decoder part (5th - 7th layers) and subsequently connected the 4th layer to another fully-connected layer to complete the training process and predict the nitrogen concentration in oilseed rape leaf. C. Zhang et al. (2020) also removed decoder part and replaced it with conventional machine learning methods which were PLS and SVM in the deep autoencoder neural network to determine the chemical composition in dry black goji berries.

For autoencoders, they are generally simpler to construct than CNN as the parameters of autoencoders are generally number of encoding layers and their respective number of nodes while CNN has more parameters to consider, such as number of convolution layers, kernel size, number of kernels in convolutional filter, stride step size and padding size. However, some notable issues of autoencoders are they have more number of parameters to be trained compared to CNN if they have large input dimension and high number of nodes in hidden layers and they may end up learning more irrelevant information if the data contains a few relevant components, which can deteriorate the predictive performance of the model.

Table 2.3: Summary of deep learning architectures for the spectral analysis in this research

Deep Learning Architecture	Feature	Advantages	Disadvantages
Feedforward Neural Network/Multilayer Perceptron	Learn through mapping of activation functions and optimization of weights (Emmert-Streib et al., 2020)	Simplest neural network structure to begin and train	Large number of parameters or nodes to be trained, exacerbated by large input dimension
Convolutional Neural Network	Extract locally correlated information from the data and store in kernels (Goodfellow et al., 2016)	Locally correlated information makes it invariant to any location, in addition to reduced number of trainable parameters	Various parameters to be optimised and large number of data
Autoencoder	Unsupervised learning to reduce dimensionality and extract features through mapping of activation functions (Emmert-Streib et al., 2020)	Easier to implement with fewer parameters to train	Large number of parameters or nodes to be trained and

2.5 Outlook of Deep Learning based Spectral Analysis

Table 2.3 summarised all the deep learning architectures described in the above sections that are to be used in this research. In the literature, deep learning for HSI was in prevalent research and use in remote sensing fields before it is applied to other fields such as biomedical, food and agriculture, document and forensic analyses (Signoroni et al., 2019; Yang et al., 2019). It is due to HSI was mainly introduced for remote sensing, hence that explains the abundance of the literature of application of deep learning for HSI in remote sensing fields. From there, deep learning was primarily used in the classification of crops and landmarks, segmentation of regions, anomaly detection as well as denoising the images to produce high-resolution images (Signoroni et al., 2019). Despite that, there was noticeably more research on applying deep learning for HSI in other fields recently. Table 2.4 outlines the application of deep learning for HSI technologies in food and agricultural applications.

The input hypercubes were often multi-dimensional, and feature extraction and engineering were manually performed with the help of PCA. Deep learning allowed feature extraction during the learning process (Al-Sarayreh et al., 2018; Li et al., 2017; Saha and Manickavasagan, 2021). As mentioned earlier, CNN and autoencoders were among the most used deep learning architectures in HSI. The input data with either spectral (1-dimensional), spatial (2-dimensional) or both spectral-spatial (3-dimensional) features were used for CNN layers to handle. Because of its proven capabilities to process a tremendous amount of data and dimensionality as well as extract valuable features on raw data without a lot of pre-processing, deep learning is highly sought after for aforementioned applications. However, there is currently no literature in surveying the degree of adulteration present in black pepper powder samples using deep learning based HSI, thus deep learning will be assessed as part of main research and compared with other multivariate data analysis techniques.

Table 2.4: Use of most deep learning based hyperspectral imaging in food and agriculture applications

Reference/Authors	Objective	Approach	Remarks
Al-Sarayreh et al. (2018)	Detection of red-meat adulteration	Deep 1D and 3D CNN	CNN model was performed with input data of 1D mean spectrum and 3D visible-NIR hypercube to classify 75 red meat products (lamb, beef, pork and fat products). The model was able to perform self-extraction of spectral and spatial features, outputted F1 score and accuracy of 94.3% and 94.4% respectively, which were better than SVM model that used manually extracted spectral and spatial features.
Jin et al. (2018)	Classification of healthy and diseased wheat heads	2D CNN	Two-dimensional convolutional bidirectional gated recurrent unit neural network was used to classify healthy and diseased wheat heads using visible-NIR HSI after pre-processing methods of mean removal, PCA whitening and normalisation were performed. It managed to achieve F1 score and accuracy of 0.75 and 74.3% respectively.
Nagasubramanian et al. (2018)	Classification of healthy and diseased soybean crops	Deep 3D CNN	3D deep CNN was used to test 539 soybean crop samples and managed to yield 95.73%, 0.92 and 0.87 classification accuracy, recall and F1 score respectively. Saliency mapping was also applied to visualise the diseased regions of soybean crops and revealed the most sensitive wavelength regions through pixels of maximum magnitude of saliency gradient.
Yu et al. (2018)	Prediction of N concentration in oilseed rape leaf	Autoencoders	Visible-NIR HSI was used to capture oilseed rape leaves and these images were split to 128 for training set and 64 for prediction set. Various regression models such as PLS and SVM were applied and compared with the deep learning model that was based on stacked autoencoders. The best stacked autoencoders model had a structure of 512-220-100-10-100-220-512, where each number separated by dash is number of nodes in each layer that resulted in the best predictive performance to predict nitrogen concentration, with R_p^2 of 0.903 and RMSEP of 0.307%.

Jiang et al. (2019)	Detection of postharvest apple pesticide residues	2D CNN	Out of 18,432 apple samples for four types of pesticide residues using NIR HSI, 12,288 were used as training data set while the rest of them as testing data set to be fed into AlexNet CNN. Single image detection yielded 95.35% and testing data set detection yielded 99.09% classification accuracy, compared to KNN detection yielding only 43.75%.
Yu et al. (2019)	Prediction of TVB-N content in Pacific white shrimp	Autoencoders	Total of 240 images of samples of Pacific white shrimps were captured using NIR HSI. Stacked autoencoders model and successive projection algorithm were applied and compared with each other in the extraction of spectral features. The processed data was split into 120 for training set and 120 for prediction set, and then sent to PLS, SVM and multiple linear regression models for prediction of TVB-N concentration. The best prediction results were attained from the SVM model that used spectral features extracted by stacked autoencoders of 215-100-50-15-50-100-215 structure, which were R_p^2 of 0.921 and RMSEP of 6.22 mg N [100 g]$^{-1}$ on prediction set.
Al-Sarayreh et al. (2020)	Classification of species of red-meat products	Deep 3D CNN	With the help of novel graph-based post-processing method (connections of superpixels), 3D CNN was built using 105 red-meat samples as training set to classify new 79 red-meat samples using line scanning and snapshot HSI, and performed better than PLS discriminant analysis and SVM, which resulted in overall classification of accuracy of 98.6%, 96.9% and 97.1% on line scanning, NIR and visible snapshots respectively.
C. Zhang et al. (2020)	Determination of chemical compositions in dry black goji berries	1D CNN and autoencoders	100 dry black goji berries were sampled using NIR HSI. Determination of total anthocyanins, avonoids and phenolics compositions were performed using 1D CNN on data with various feature selection methods, all resulted in lower RMSEP than that using PLS and SVM. PLS and SVM models performed better when the features of inputs were extracted using CNN and stacked autoencoders.

Gao et al. (2020)	Classification of ripe and early ripe strawberries	2D CNN	Using portable snapshot visible-NIR HSI, input images of ripe and early ripe strawberries were captured in both field and laboratory conditions. PCA was performed on these hyperspectral images using 3 principal components for spatial feature extraction while sequential feature selection was performed for spectral feature wavelength selection. 336 and 144 of them were used as training and validation data set respectively for training AlexNet CNN, and as a result, 98.6% classification accuracy was obtained.
L. Zhang et al. (2020)	Classification of frost-damaged rice seeds	Deep forest model	1800 NIR hyperspectral images of rice seeds with different degrees of frost damage were assessed, pre-processed using MSC, and modelled using decision tree, KNN, SVM, deep forest model. With only a small amount of data set (around 50 samples), deep forest model was able to achieve more than 80% classification accuracy easily than other models. Visualisation on classification of frost-damaged rice seeds by deep forest model was also made and the model correctly classified 298 out of 300 (99.33%) rice seeds samples.
Ma et al. (2020)	Classification of seed viability	2D CNN	A 2D CNN model was used using PCA and SVM mapping results from the NIR hyperspectral images of Japanese mustard spinach seeds as inputs on the wavelength range of 1002 – 2300 nm to classify the viability of those seeds. The result on the testing set containing 240 seeds showed that combination of PCA and SVM mapping inputs after Savitzky-Golay pre-processing yielded the best classification accuracy of approximately 90% (87.5% true positive, 91.1% true negative).
Weng et al. (2020)	Classification of rice variety	Variant of CNN (PCANet)	Visible-NIR HSI was used to capture images of 10 types of rice which were then pre-processed using MSC, SNV, Savitzky-Golay filtering. The data, containing 3240 samples for training set and 1080 for prediction set, was later processed using PCA on spectroscopic and texture features and sent to deep learning model using PCANet (principal components

			analysis network). It was compared with KNN and random forest models, and resulted in the best classification accuracy of 98.57%.
Xin et al. (2020)	Prediction of cadmium residue in lettuce leaves	Autoencoders	Line-scanning visible-NIR HSI was used to acquire 1120 hyperspectral images of lettuce leaves with 7 different concentrations of cadmium chloride. Using the best spectral pre-processing method of Savitzky-Golay with 1st derivative, the best R_p^2 of 0.9487 and RMSEP of 0.01049 could be obtained from stacked autoencoders of model scale 477-240-111-81 (number of nodes for each layer, separated by dash) with least square SVM regression.
Han et al. (2021)	Predict quality level of nuts by peroxide values	2D CNN	The quality of unbalanced kernels of *Canarium indicum* nuts was estimated through Visible-NIR HSI (388.9 – 1005.33 nm) using CNN model. The inputs were 60 2D sub-images from raw hyperspectral images. The classification (good, medium or poor quality) results were best achieved using inputs after dimensionality reduction by PCA, with testing accuracy of 93.48%. The resulting CNN model consisted of 4 convolution layers of 1×1 convolution kernels and a SVM classifier. Regression was also made on predicting the peroxide values of nuts, but the model managed to produce R^2 of only 0.67.
Hong et al. (2021)	Classification of storage years of black tea samples	CNN, long short term memory (LSTM) and CNN-LSTM	NIR hyperspectral images of black tea samples of different storage years using wavelength range of 874 – 1734 nm were captured. Their mean spectra were pre-processed with wavelet transform and processed using PCA so that obtained PCA loadings could be used as inputs of classification models. Common machine learning methods (linear regression, SVM) and deep learning (CNN, LSTM, and CNN-LSTM) were used to fit full range and optimal wavelengths spectra into models and then compared. Deep learning models had better classification performance, where the LSTM model using full range spectra had 83.601% accuracy, while the CNN-LSTM using optimal wavelengths spectra had 81.029% accuracy.

2.6 Summary

Most black pepper powder samples were adulterated with papaya seeds powder due to resemblance in colour. To effectively detect and estimate these adulterants present in black pepper powder samples, a rapid, non-destructive and reliable detection method is required. Chemical analytical and molecular methods provide detailed qualitative and quantitative results, however they require high operating cost, tedious sample preparation, long processing time, controlled environment, and prior knowledge and expertise in operating the equipment and apparatus to be effective. In addition to that, chemical waste was produced as a result. As these methods could not be applied into real-time industrial quality assessment, CVS and vibrational spectroscopy were typical potential solutions for effective quality assessment of black pepper powder samples. However, CVS did not function well on powdered samples and there were various hidden chemical information and internal quality characteristics that could only be found outside the operating domain of computer vision. Vibrational spectroscopy lacked representativeness which could be inconsistent on every measurement of samples. Thus, HSI was introduced to address both the shortcomings.

Due to exorbitant amount of data and high dimensionality in HSI, multivariate data analysis was used to extract representative features from the hyperspectral inputs. That extracted data would be used to create a model to predict the quantity of adulterants as well as most chemical analytical properties. To assess the model predictive performance, various multivariate data analysis techniques such as PCA, PLS and SVM were applied and compared with deep learning, which was demonstrated in literature that it outperformed the aforementioned techniques and was capable of handling raw data without deterioration of predictive performance. Yet, little to no reference could be found for black pepper powder adulteration detection using HSI and deep learning. The same goes to geographical origin classification, which is also crucial in understanding the state of black pepper powder samples in Sarawak. Hence, this research seeks to achieve these objectives and prove that HSI is fully capable to be an alternative to detailed analytical chemistry techniques as an effective detector and estimator of adulterants in Sarawak black pepper powder samples as well as classifier of geographical origin.

CHAPTER 3
MATERIAL PREPARATION AND MODELLING METHODOLOGY

This research ultimately seeks to investigate the effectiveness of hyperspectral imaging (HSI) in detecting the adulteration, assessing the authenticity and classifying the geographical origin of Sarawak black pepper powder samples. There are two cases of research to be investigated: (a) the authenticity of Sarawak black pepper powder samples are assessed by quantifying their degree of adulteration or purity of black pepper and their geographical origin is classified; (b) HSI is then assessed in predicting chemical and biological analytical properties of Sarawak black pepper powder samples. In both cases, the HSI inputs are pre-processed using data pre-processing methods including Savitzky-Golay (SG), standard normal variate (SNV), SG-1st derivative and SG-2nd derivative, undergone data exploration using PCA and assessed on the prediction of intended targets using PLS, SVM, SVR and deep learning.

3.1 Sample Preparation

Pure black pepper berries from five regions of Sarawak (Serian, Sungai Tenggang, Pakan, Lachau and Sibu) and papaya seeds are acquired from local suppliers. The papaya seeds are dried in a natural convective oven at 45°C for 5 – 7 days to ensure low moisture content. All the black pepper and papaya seeds are then milled into powder and packaged in polyethylene sealed bags. Two sets of black pepper powder samples were prepared, which were the training set and testing set. Different amount of samples in both sets are adjusted for two parts of research.

For the determination of authenticity and origin of black pepper powder samples, the training set consisted of the black pepper powder samples adulterated with papaya seeds powder in 20 defined proportions of 0 - 15% (with 1% interval), 17.5%, 20%, 25% and 30%. Since there are 5 regions to be tested, this results in total of 5 × 20 = 100 black pepper powder samples to be analysed with HSI. An example of a training set containing pure and adulterated black pepper powder samples from Serian is shown in Figure 3.1. While for the testing set, it is made up of 29 black pepper powder samples with randomly defined and labelled proportions of papaya seeds powder.

While for the prediction of chemical and biological analytical properties of black pepper powder samples, the samples are made up of black pepper powder samples with 11 defined proportions of 0 – 10% (with 1% interval). This results in total of 11 × 5 = 55 samples to be analysed. External third-party chemical and biological lab analysis is performed on these

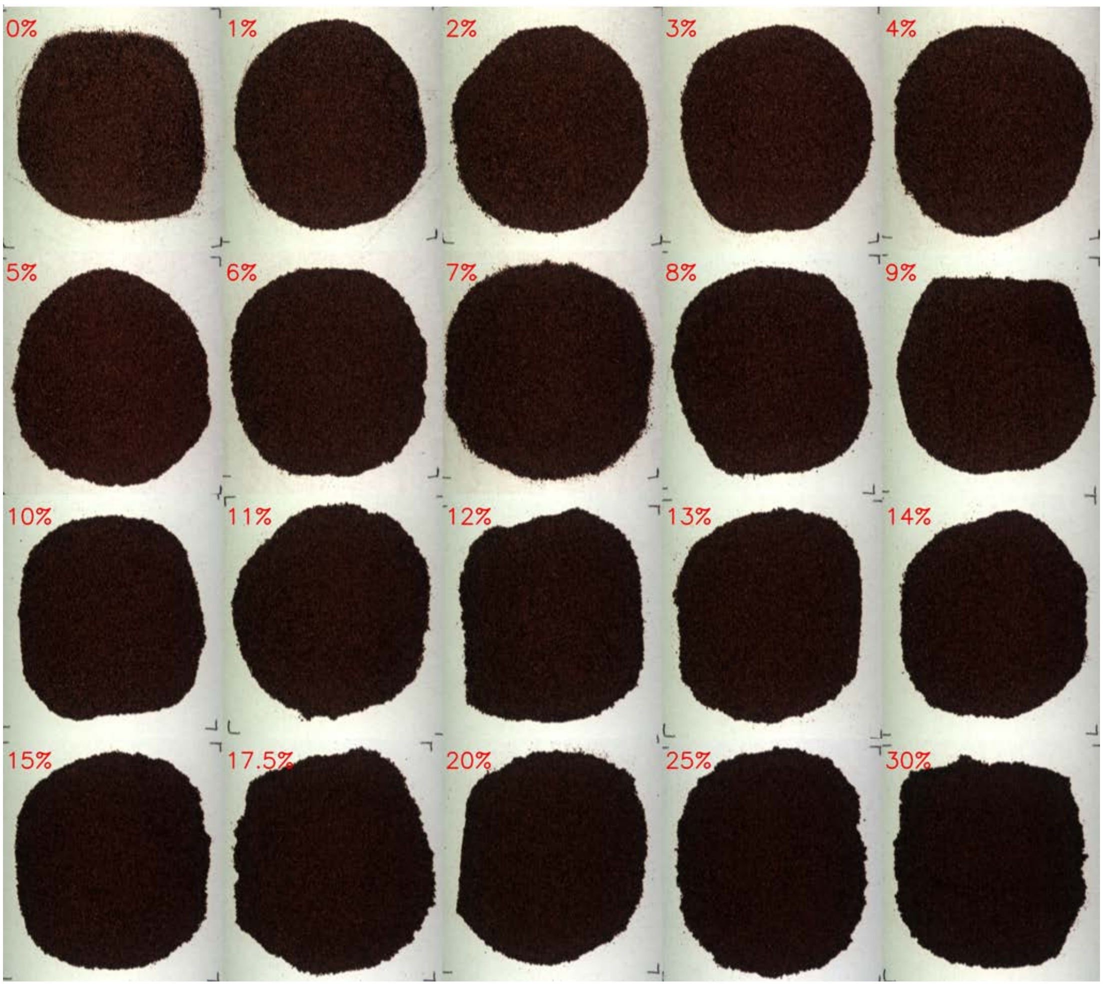


Figure 3.1: RGB image of Serian black pepper powder samples of different degree of adulteration of papaya seeds powder

samples. The scope of lab analysis may include the determination of moisture content, total ash content, volatile oil and non-volatile ether extract contents, and microbiological analysis such as checking total count of yeast, *Escherichia coli* and *Salmonella spp* (MPB, 2021).

3.2 Data Acquisition

The HSI equipment to be used in this research is Visible-NIR HSI equipment (Resonon Pika L) which has a spectral range of 400 – 1000 nm with spectral resolution of 2.1 nm and 300 spectral bands. The frame rate and shutter time were set to be 50 frames/sec and 17.446 ms respectively. Samples were placed on the centre of the stage illuminated by the halogen lamps and scanned subsequently line by line with speed of 4.5 mm/sec. The distance between the camera and the sample was approximately 0.2 m. The captured HSI data is in BIL format and can be processed using Spectral Python library.

The reflectance of captured HSI data or hypercubes is then calibrated in comparison with spectra of standard reference materials when appropriate according to Equation (3.1),

$$R_c = \frac{R_0 - D}{W - D} \times 100 \tag{3.1}$$

where R_c is corrected reflectance, using reflectance from raw hypercube, R_0 with the reflectance values of both reference, W and dark, D images (Elmasry et al., 2012; Orrillo et al., 2019; Wu and Sun, 2013b).

From the centre of hyperspectral image of each sample, they are segmented into 81 sub-samples where each sub-sample will have spatial dimension of 50 × 50, as illustrated in Figure 3.2. Sub-sampling increases the availability of the data, thus it allows machine learning models to learn and attain better predictive performance. However, more sub-samples will reduce the spatial dimension of hyperspectral images, hence the models will have less details and features to learn and eventually deteriorated predictive performance due to reduced fidelity of images. After that, the **raw mean spectra** of all these sub-samples (the black pepper powder samples from these segments) are obtained. As a result,

(1) For determination of authenticity and origin of black pepper powder samples:
   - Training set = 100 × 81 = 8100 sub-samples
   - Testing set = 29 × 81 = 2349 sub-samples

(2) For prediction of chemical and biological analytical parameters, training set from (1) has some unrelated samples (i.e. percentage of black pepper not covered in the lab analysis) removed, thus:
   - Total number of sub-samples = 8100 – 45 × 81 = 4455 sub-samples
   - Training set = 3113 sub-samples
   - Testing set = 1342 sub-samples

3.3 Data Pre-processing

Effects of various data pre-processing methods are assessed as data pre-processing enhances the predictive performance of machine learning models on the prediction of the aforementioned intended targets. It is of interest to investigate the effects of data pre-processing on the predictive performance of various machine learning models. In HSI, data pre-processing is usually applied on primarily spectra data set, targeting the spectral dimension. It is especially helpful in removing redundancy and undesirable effects such as light scattering effects due to irregularity of surface and particle size of powder samples and random noise (Barnes et al., 1989; Orrillo et al., 2019; Saha and Manickavasagan, 2021; Wu and Sun, 2013b).

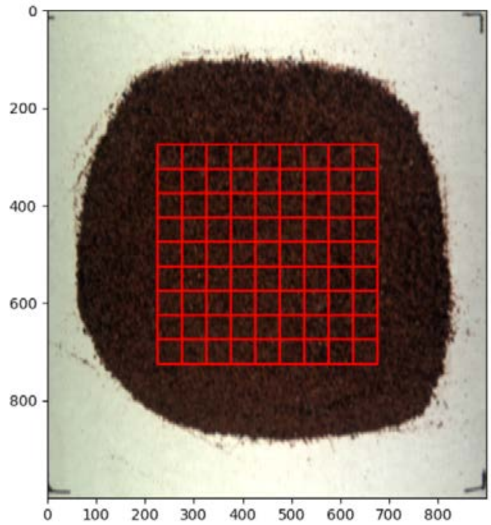


Figure 3.2: Example of 81 sub-samples segmented from the center of Serian pure black pepper powder sample

In this research, pre-processing is mostly applicable for mean spectra data set and some methods to be used are standard normal variate (SNV) and Savitzky-Golay (SG) filtering. SNV normalises the data by centring and scaling them by following the normal distribution (Barnes et al., 1989; Wu and Sun, 2013b). SNV scaled samples can be obtained using the following equation:

$$x_{SNV} = (x - \bar{x}) / \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}} \quad (3.2)$$

where x is the value of a sample, $\bar{x}$ is mean of samples, and n refers to the total number of samples.

SG filtering is usually used to smoothen the signal or spectra, and its derivative helps to correct the baseline effects present in the spectra (Elmasry et al., 2012; Wu and Sun, 2013b). The main factor of SG filtering is the frame length which allows the spectrum to be fitted into polynomial within frame of one measurement point (Savitzky and Golay, 1964). SG filters are tested using frame length of 15 points and second polynomial order with and without first and second derivatives. All pre-processing methods are implemented using Python Scikit-learn and SciPy libraries (Pedregosa et al., 2011).

3.4 Data Exploration and Analysis

The hypercubes usually contain a high number of variables in spectral dimension carrying redundant information that requires proper processing. Multivariate data analysis is broadly applied to address this issue by extracting the essential features in the spectral dimension. The data with extracted features contain information that explains the data and is subsequently used to fit a model to yield desired outputs. PCA is a widely used chemometrical method to act as preliminary data screening tool on whether discrimination of classes or clusters is possible or not (Galvin-King et al., 2021; Kherif and Latypova, 2020). It is also used to reduce the dimensionality and output a model with a certain number of principal components that capture the maximum variability of the data (Orrillo et al., 2019). As the spectral data usually has a high number of dimensions, PCA is used to observe whether the spectral data with reduced dimensionality has variability that explains the correlation (Wilde et al., 2019). The data of interest to be analysed is the mean spectra of the HSI data. PCA transforms the variables in the spectral dimension of the mean spectra into principal components, ranked by the amount of explained variability. The number of principal components is set to be 10. Both raw and pre-processed data that use SG filtering, SNV and SG with 1st and 2nd derivatives are explored using PCA.

3.5 Model Development and Training

Machine learning models for the two aforementioned cases of this research are built, developed and assessed. The models are trained using supervised learning, where the data with designated labels are trained. The main models of choice in this research are PLS, SVM, SVR and deep learning (DL). PLS and SVR are used for regression while PLS with discriminant analysis (DA) and SVM are chosen for classification instead. DL is used in both cases and its model architectures of choice are primarily CNN and stacked autoencoders. The whole implementation is performed using Python, where PLS and SVM models are based on functions from Scikit-learn and SciPy libraries, while DL is from Tensorflow Keras library. The performance criteria for regression are RMSE, R^2 and MAPE from Equations (2.2) to (2.4). Parity plots are then used to review the regression results. For classification, accuracy from Equation (2.1) is usually used and confusion matrix is used to visualise and review the classification results. The details on implementation of these models for the two cases of this research are explained in the following sections.

3.5.1 Determination of Authenticity and Origin of Sarawak Black Pepper Powder Samples

The determination of authenticity is based on the purity of black pepper in those Sarawak black pepper powder samples, or degree of adulteration of papaya seeds (Orrillo et al., 2019; September, 2011; Wilde et al., 2019). The classification is mainly for identifying the geographical origins of black pepper powder samples, targeting mainly multiple regions in

Sarawak, Malaysia. The training set is further split into training and validation data sets with split ratio of 7:3. Typical split ratio is 8:2, however due to limited amount of data and potential overfitting issue by larger training data proportion, the split ratio is set to 7:3 instead (Goodfellow et al., 2016). This then results in: 8100 × 0.7 = 5670 in the training set, 8100 × 0.3 = 2430 in the validation set. Validation data set is established to initially assess the model predictive performance before being tested in an unknown testing data set.

For PLS, the main parameter to be tweaked is the number of latent variables. Thus, to find the optimal number of latent variables for both PLS models, minimum RMSE for PLS regression model and maximum classification accuracy for PLS-DA model are obtained respectively using a range of integers from 1 to 50 for the number of latent variables during training and validation of both PLS models. After the optimal number of latent variables is found, the predictive performance of both PLS regression and classification models are assessed using testing data set.

For both SVM and SVR, the commonly used kernel function to yield the best decision boundary is the radial basis function kernel. The main parameter to be adjusted is the regularisation strength, C, where it controls the balance between smooth decision boundary and correct classification of trained data as well as γ, which is an important factor for radial basis function kernel in controlling influence of single training data on the model (Saha and Manickavasagan, 2021). The higher the C, the lesser the tendency for model to overfit but lesser number of training data to consider. For γ, higher flexibility or non-linear boundaries are considered rather than linear boundaries when higher γ is used. ε-SVR model is used in regression instead, and the parameters are C and γ. ε is another parameter which controls the tolerance during the training of SVR model and is responsible in finding balance between model predictive performance and model training time. ε is fixed to 0.01 in this case. For each different data pre-processing method, various values of C and γ are tested to find the best combination which yields the best model predictive performance. Possible values of C and γ to test are [1, 10, 100] and [0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000, 100000] respectively. As the classification of origin involves multiple classes, the SVM decision boundary shape is set to “one-vs-one” in the Scikit-learn function.

Next, DL models are compared with PLS and SVM or SVR models. There are two sections in the hidden layers of DL model. The first section acts as a feature extractor while the second section acts as a predictor or classifier. As aforementioned, there are two DL model architectures to be tested: CNN and stacked autoencoders (SAE). The input size is fixed at 280 which corresponds to the number of spectral bands of each raw mean spectrum.

In CNN, the feature extractor section consists of convolution and pooling layers. There are several main parameters in the convolution part of CNN model to be adjusted: (1) number of convolutional layers, (2) size of convolutional kernels in each convolutional layer and (3) the layer configuration or number of convolutional kernels in each layer. In (1), the number of convolutional layers is set to 4. Such this adjustment relates to pooling operations as pooling reduces the dimensionality of inputs of each layer. Max pooling is used and its kernel size is fixed to 2, which divides dimensionality by 2. In (2), possible convolutional kernel sizes to be tested for each layer are [1, 3] in a Cartesian product combinations. While for (3), the number of kernels is initially set with proposed numerical values of base 2. Doing this is based on heuristics as it will be slow and inefficient to optimise based on range of values, for example 1 to 10000. The layer configuration is set to 16-32-64-128. Each layer is always accompanied by ReLU activation function to ensure non-linearity of model. Batch normalisation component may be added between convolution and pooling to allow faster convergence, better generalisation due to standardisation of the inputs similar to SNV (Goodfellow et al., 2016; Yang et al., 2019). Using Tensorflow Keras library on the mean spectra data as input, conv1D function is used for the convolution operation while maxPool1D is used for the max pooling operation. Example of the CNN model was illustrated in Figure 3.3.

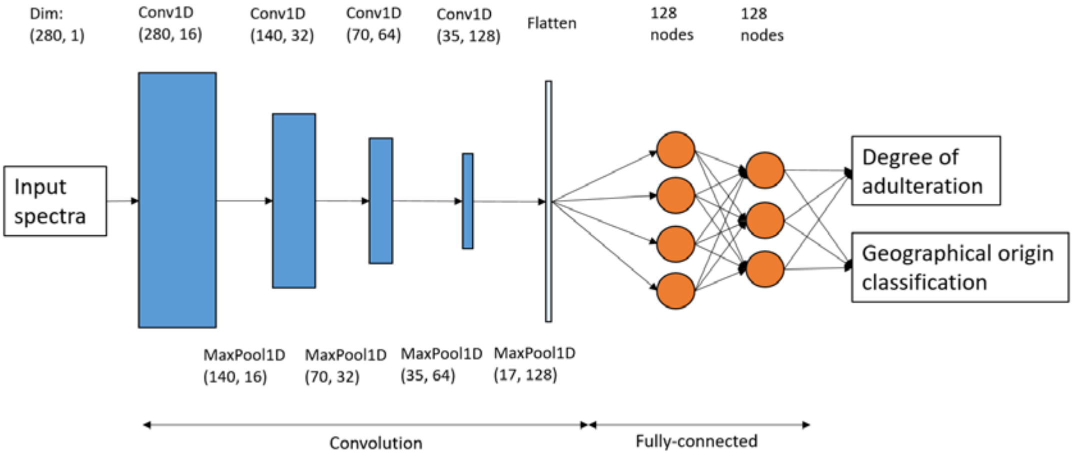


Figure 3.3: Schematic of one example of DL CNN model structure

In SAE, the feature extractor section is instead made up of multiple fully-connected autoencoders. The decoding part of SAE is removed, leaving only encoding part to extract the features. The main parameters to be adjusted are (1) number of SAE encoding layers and (2) the layer configuration. In (1), possible number of SAE encoding layers are [3, 4]. The layer configuration in (2) is proposed to be [192-16, 192-128-16, 192-128-64-16, 192-128-64-32-16, 192-128-96-64-32-16]. Each layer is also accompanied by ReLU activation function. Most SAE layers are implemented using Dense function from Tensorflow Keras library. Example of the SAE model was illustrated in Figure 3.4.

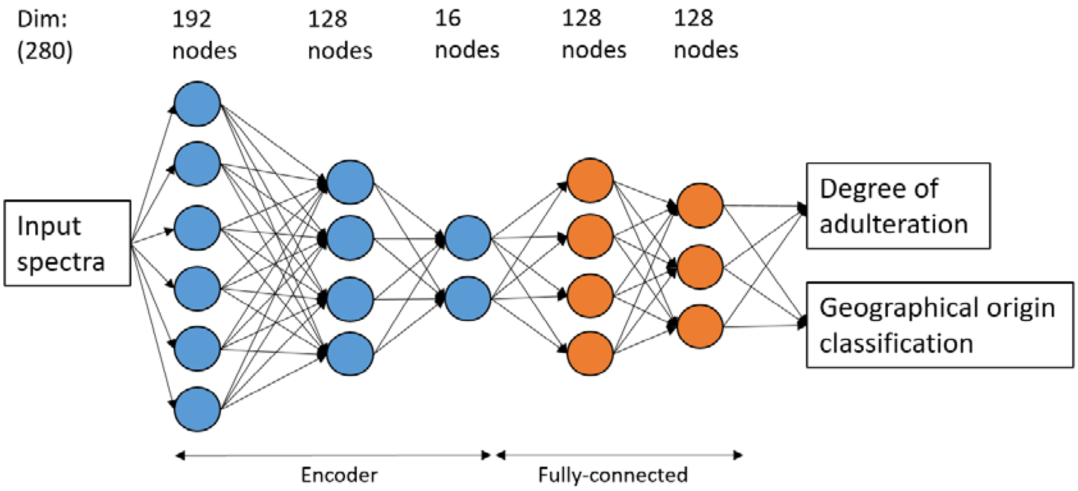


Figure 3.4: Schematic of one example of DL SAE model structure

After that, the predictor section of any model after the feature extraction is fully-connected layers. It is set to 2 layers with layer configuration of 128-128. The reason for that was mainly increasing more layers could make model more computationally heavy and have higher tendency to overfit. These layers are responsible to decide the weighing of extracted features to make appropriate predictions. Each of these layers are accompanied by ReLU activation function as well. The output layer consists of two types of output, which are (1) 1 node using sigmoid activation function from Equation (2.12) to output value of authenticity or purity of black pepper powder samples, $\{x \in \mathbb{R} \mid 0 \leq x \leq 1\}$, and (2) 5 nodes of softmax activation function showing probability of identified geographical origin for 5 regions to be tested, $\{x \in \mathbb{R} \mid 0 \leq x \leq 1\}$. Softmax activation function is defined as follows:

$$\text{softmax}(z_i) = \frac{\exp(z_i)}{\sum_{j=1}^{n} \exp(z_j)} \tag{3.3}$$

Table 3.1: Details of output layer

Output	Type	Activation	Loss function	Metric
Authenticity of black pepper	Regression	Sigmoid	RMSE	RMSE
Geographical origin	Multi-class classification	Softmax	Categorical cross-entropy	Accuracy

To enable model training, Adam optimiser is used, proper loss functions and metrics are set according to Table 3.1. Before the training of DL model, hyperparameters are set externally by the user and they cannot be tweaked by machines easily. Hyperparameters such as the batch size which controls the number of input samples to be fed at one time, number of epochs which is the number of times the training process to be repeated and, decay and learning rates for the Adam optimiser which controls the step size adaptively during each step of parameters tweaking of model are set. The batch size is set to 64 and number of epochs to 100. The learning rate is set to default 0.001 as recommended by the authors of Adam optimiser while the decay rate is set according to formula as follows (Kingma and Ba, 2017):

$$Decay\ rate = \frac{global\ learning\ rate}{epochs} \tag{3.4}$$

Reproducibility is an important factor which affects the data splitting and selection. It is to ensure the consistency during comparison of predicted results from various models. The random seed number is fixed to 30.

3.5.2 Prediction of Chemical and Biological Analytical Properties of Sarawak Black Pepper Powder Samples

Because there is a lack of validation data set, cross validation of 10 folds is used instead. In this case, most chemical and biological analytical properties to be predicted are continuous real values, thus regression is only considered. The difference compared to the modelling done in Section 3.5.1 is the models are required to account for multiple outputs. There are total of 8 output variables to be trained for, where 6 of them belonging to chemical while the remaining 2 are biological properties. The details of these properties are outlined in Chapter 5. Both of these categories have to be trained in separate models due to difference in values.

PLS and SVM models undergo similar assessments as in Section 3.5.1. The possible range for parameter of PLS which is number of latent variables is from 1 to 50. While for the parameters of SVM which are C and γ, their possible values are [1, 10, 100] and [0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000, 100000] respectively.

For DL models, similar models (CNN and SAE), settings and comparison as in Section 3.5.1 are made. Some exceptions are the predictor section is instead made up of two separate fully-connected layers leading to respective outputs of chemical and biological analytical properties. The outputs use ReLU activation function to ensure non-negativity, $\{x \in \mathbb{R} \mid x \geq 0\}$. All the required hyperparameters such as number of epochs, learning rate, decay rate for the optimiser and random seed number are retained. Examples of the CNN and SAE models could be seen in Figure 3.5 and Figure 3.6.

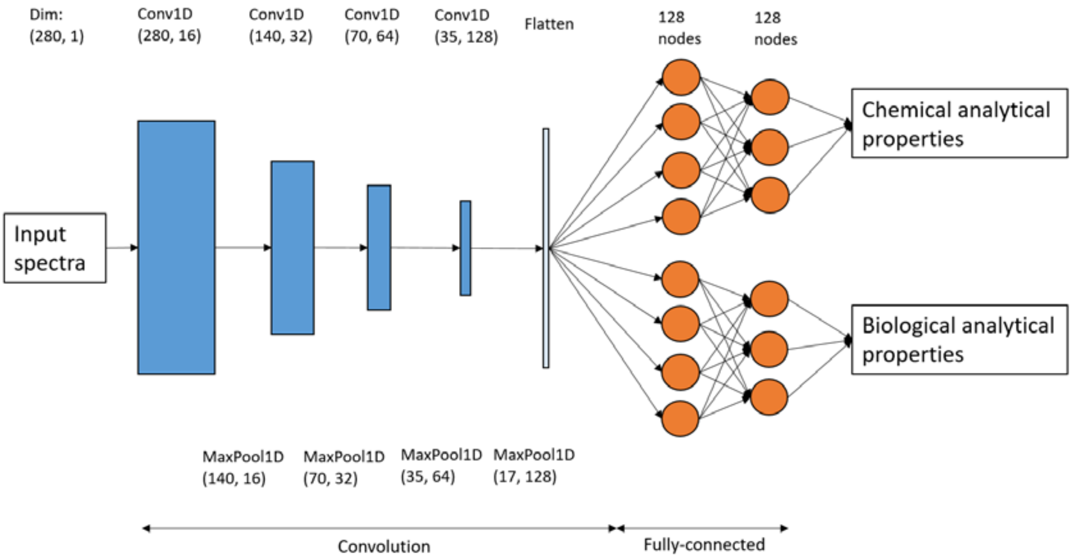


Figure 3.5: Schematic of DL CNN model structure for prediction of chemical and biological analytical properties

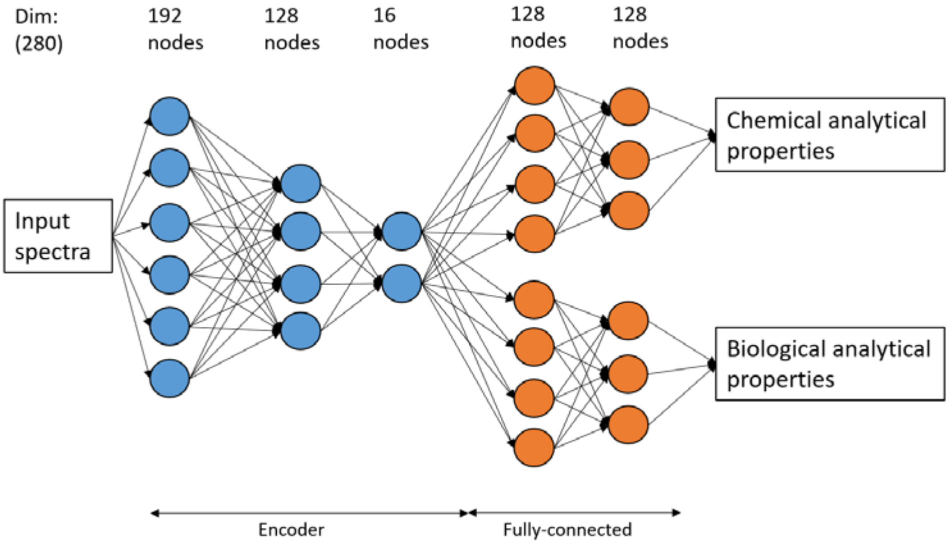


Figure 3.6: Schematic of DL SAE model structure for prediction of chemical and biological analytical properties

3.6 Summary

The whole research workflow is summarised as illustrated in Figure 3.7. The determination of authenticity and geographical origin of Sarawak black pepper powder samples using HSI along with data pre-processing and various machine learning techniques are assessed. After that, prediction of internal quality (chemical and biological analytical properties) of Sarawak black pepper powder samples is evaluated as well. All of these will be discussed in details in the following chapters.

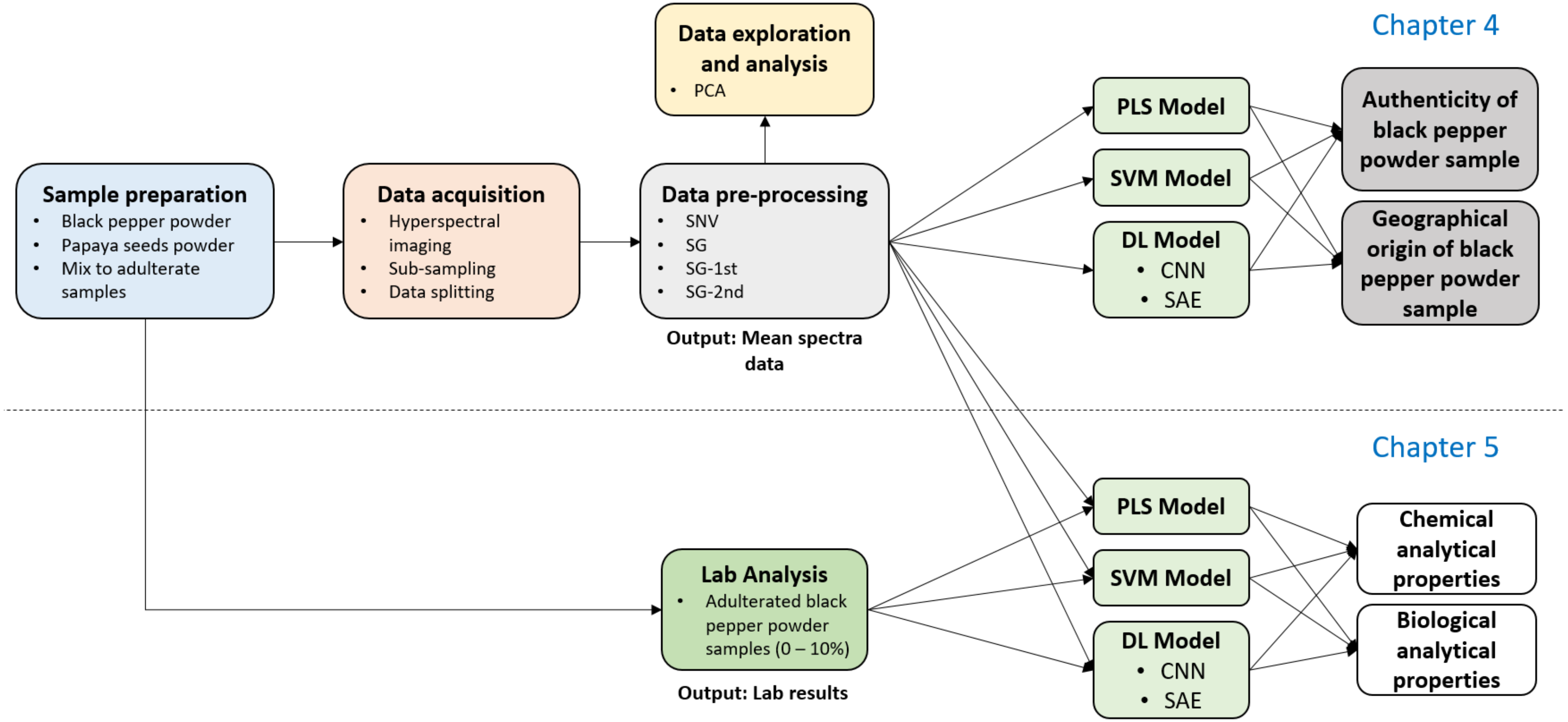


Figure 3.7: Research workflow

CHAPTER 4

REGRESSION OF DEGREE OF ADULTERATION AND CLASSIFICATION OF GEOGRAPHICAL ORIGIN

To determine the degree of adulteration or authenticity and geographical origin of Sarawak black pepper powder samples, visible-NIR hyperspectral imaging (HSI) is mainly used to capture hyperspectral images of Sarawak black pepper powder samples before these HSI data are sent for further pre-processing and modelling. As the HSI equipment used in this research covers the wavelength range from visible light to NIR, it is of interest to observe the optical effects of adulteration under visible light. Referring back to Figure 3.1, an example of RGB image of pure and adulterated Serian black pepper powder samples from HSI equipment was shown. From there, it is apparent that the more adulterated the black pepper powder samples, the darker they appear due to darker colour of papaya seeds powder than black pepper powder.

To prove this trend numerically, as shown in Figure 4.1, the reflectance measured by HSI for all Serian black pepper powder samples showed that the more adulterated the black pepper powder samples, the lower the reflectance as the darker colour of the samples reflects lesser light. In visible light region between 450 – 700 nm, the effects of adulteration on the difference in reflectance among samples were apparent, but they became less noticeable as the wavelength increases. On the NIR region beyond 700 nm, the difference in reflectance was still apparent from 700 – 800 nm, but it becomes spectrally similar after 800 nm. It could be

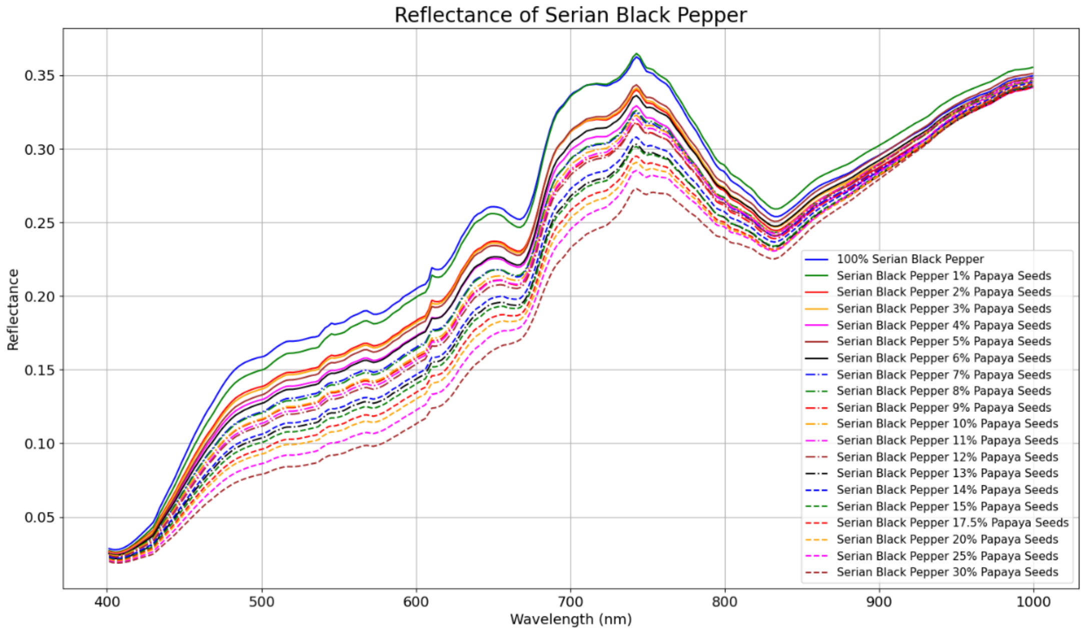


Figure 4.1: Reflectance of raw mean spectra of pure (100%) and adulterated Serian black pepper powder samples

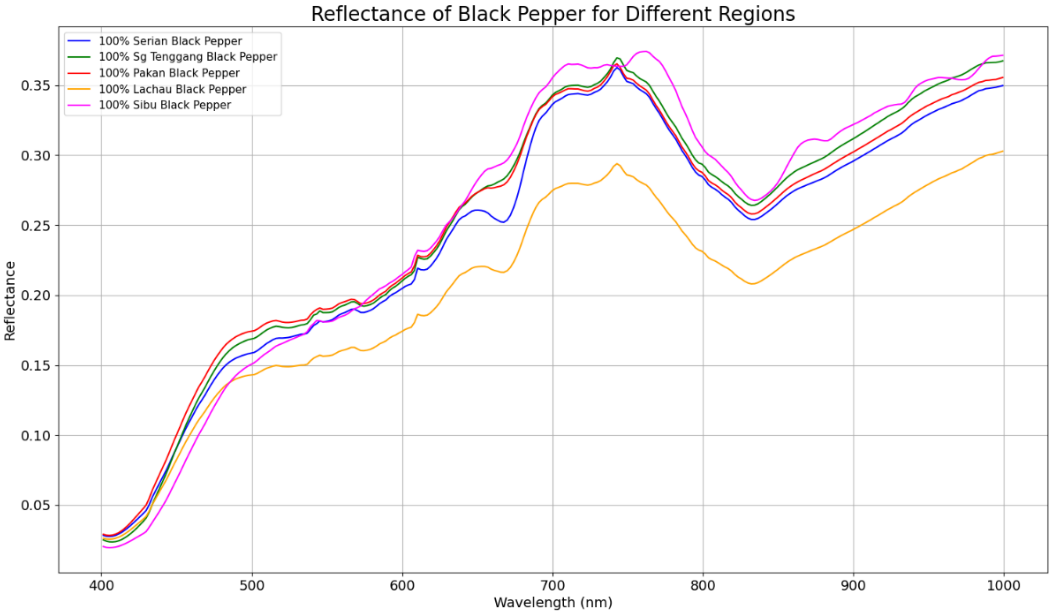


Figure 4.2: Reflectance of raw mean spectra of pure (100%) black pepper powder samples from different regions or geographical origins of Sarawak

less helpful to rely on the spectral features after 800 nm as its similarity could potentially cause conflicting predictions. The black pepper powder samples from other regions had similar patterns on the effects of adulteration, which can be seen in Figure B.1 and Figure B.2 from Appendix B. It was further observed that there is a significant peak around 750 – 780 nm, which could be attributed to O–H third overtones, indicating the presence of water, or C–H fourth overtones (Jha, 2016; Liu et al., 2014). Two alternating peaks were found between 630 – 700 nm, which was related to reflection of red bands due to dark brownish colour of black pepper powder samples.

On the other hand, from Figure 4.2, pure black pepper powder samples shared similar patterns with each other of different origins. Adulterated black pepper powder samples had similar patterns with the pure samples, with the exception of having lower reflectance. Lachau black pepper powder samples had the lowest reflectance compared to those of other origins, due to their darker appearance. Sibu had different spectral patterns compared to those of other origins, such as numerous peaks spotted on 700 – 800 nm and after 850 nm. There were no significant spectral differences for Serian, Sg Tenggang and Pakan black peppers, thus this required pre-processing or detailed analysis to clearly distinguish them.

4.1 Data Pre-processing and Exploration

Data pre-processing is usually performed on the spectral data to enhance the interpretability of the data. The data pre-processing methods used in this research were SG, SG with SNV, SG with 1st derivative and SG with 2nd derivative. SG and SG-SNV filters retained most of the spectral properties as the raw data since SG filter only smoothened the spectral data while SG-SNV filter normalised the smoothened data. The derivatives on SG filter magnify the differences or show the gradients present in the spectral data. Examples of these data pre-processing methods in action were shown in Figure 4.3 and Figure 4.4.

From Figure 4.3, as expected, SG only smoothened the spectra while SG-SNV normalised every spectrum. On the other hand, some prominent peaks such as 400 – 500 nm and 750 – 800 nm were spotted on SG-1st pre-processing method as well as a few spectral differences. For SG-2nd pre-processing method, there were hardly any notable spectral differences. Next, from Figure 4.4, it was observed that SG-1st and SG-2nd pre-processed mean spectra data had more spectral differences and prominent peaks than SG and SG-SNV pre-processing methods. Other than that, Sibu black pepper powder samples particularly had different spectral patterns than other regions.

PCA was then performed to transform and visualise the mean spectra data so that the data can be interpreted easily, hence allowing investigation and explanation which related to the research objectives. The number of principal components (PC) was set to reduce the dimensionality of the spectral data. PCA was applied on all raw and pre-processed mean spectra data using 10 PCs. After that, PCA score plots were made to visualise the scores, which are weights of each sample to projected PCs, according to various labels and targets, which are percentage authenticity or purity of black pepper and geographical origins. PCA score plots for first 3 PCs for all raw and pre-processed mean spectra data were illustrated in Figure 4.5 - Figure 4.9. Referring to these PCA score plots, the pattern of the scores after labelling could be observed and analysed on the possibility of clear discrimination of clusters or targets. From Figure 4.5, different states were observed for PCA score plots on comparison of two different PCs for raw mean spectra data for different classes of authenticity and geographical origins. On PC1 vs PC2, it was easier to spot the boundaries on different authenticity or purity of black pepper, but not for different geographical origins. Sibu black pepper powder samples appeared as a mixture among Serian, Sg Tenggang and Pakan samples. While for other comparisons which were PC1 vs PC3 and PC2 vs PC3, it was difficult to clearly discriminate the clusters and hence identifying the degree of adulteration and geographical origin, although to a lesser extent, it was still possible to identify the geographical origins based on PCA score plot for PC1 vs PC3. On the other hand, SG pre-processing data had similar effects with the raw mean spectra since it only smoothened the spectra, hence its PCA score plots in Figure 4.6 were unchanged. PCA score plots for SG-SNV pre-processing data in Figure 4.7 showed that for PC1 vs PC2, the distinction of clusters was visible to a smaller degree on the detection of degree of adulteration and classification of geographical origins. However, for SG-1st and SG-2nd pre-processed data, the PCA score plots in Figure 4.8 and Figure 4.9 illustrated that determination of degree of adulteration was slightly possible to smaller extent, but on the discrimination of geographical origins, Sibu labelled data were distant from other data, making the data pre-processing method a good candidate for further modelling purposes in the classification of geographical origins.

Table 4.1: Explained variability of first 3 PCs of raw and pre-processed mean spectra data

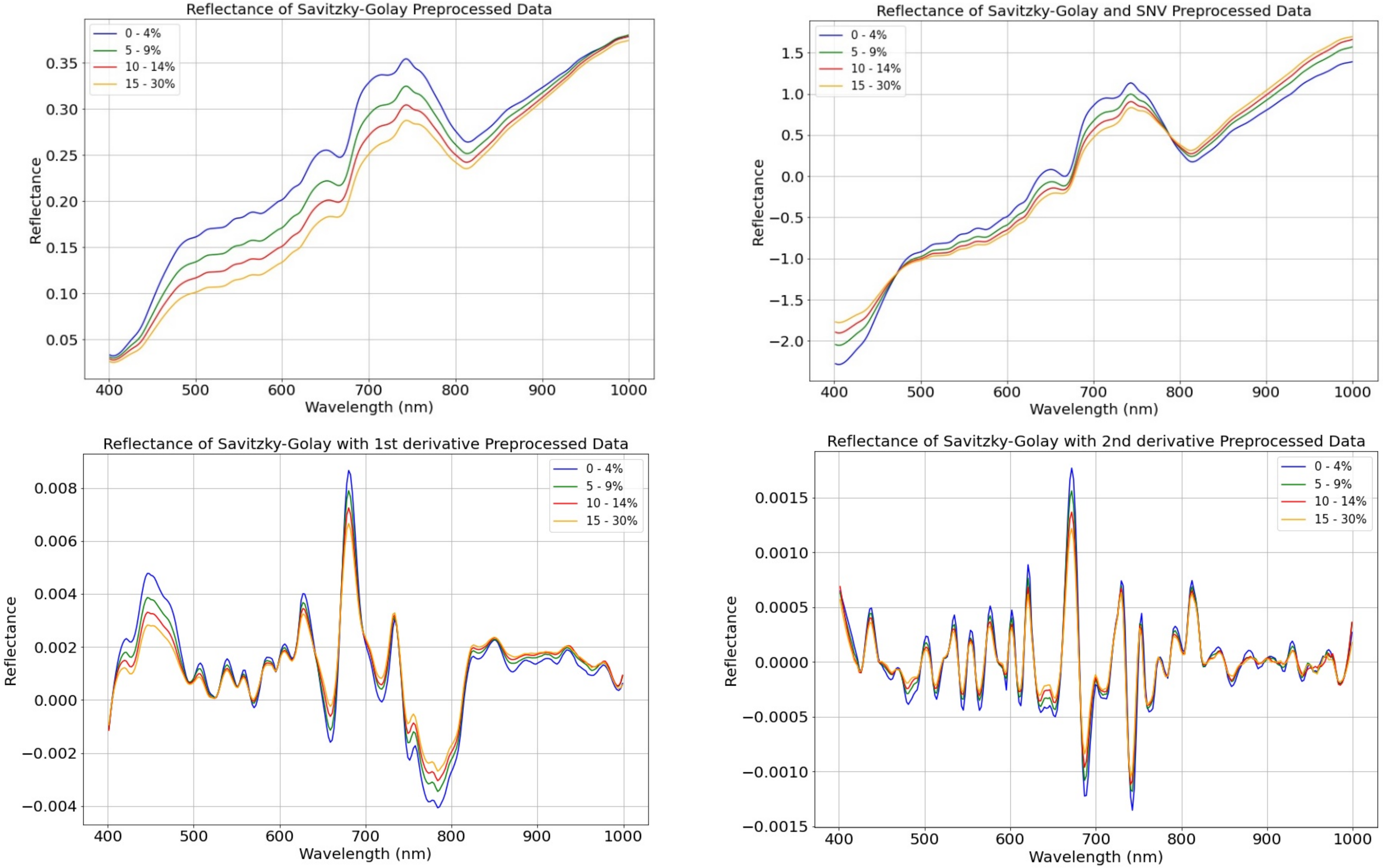
Pre-processing	**Explained Variability (%)**		
	PC1	**PC2**	**PC3**
Raw	78.34	17.52	3.06
SG	78.34	17.52	3.06
SG-SNV	83.59	9.02	4.36
SG-1st	46.15	24.37	12.36
SG-2nd	63.02	12.66	9.91

Explained variability was then determined where the degree of spread of data explained by PCs was calculated. The explained variability of first 3 PCs was illustrated in Table 4.1. Except SG-1st and SG-2nd, all mean spectra data had explained variability of more than 95% based on first 3 PCs, where the raw and SG pre-processed data were 98.92%, while SG-SNV was 96.98%. This is important because it is generally favourable to have a simpler model which contains lesser number of variables or PCs to explain most of the relationship contained in the data. The normalisation by SNV reduced the multiplicative scattering effects from the spectral data and subsequently provided a standardised baseline to allow clearer spread and better feature selection, as indicated with higher explained variable on first principal component (Barnes et al., 1989; Elmasry et al., 2012; Modupalli et al., 2021; Wu and Sun, 2013b). On the other hand, SG-1st and SG-2nd pre-processed mean spectra data had lower explained variability and hence required more principal components to fully explain the spread of the data.

Judging from the PCA score plots and explained variability, it is of interest to understand whether data exploration by PCA revealed any intriguing explanation on the HSI spectra data with regards to the authenticity and geographical origin of Sarawak black pepper powder samples. PCA loading plots were hence constructed to investigate the relationship between the spectral bands and PCs. Figure 4.10 – Figure 4.12 illustrated the PCA loading plots for all the raw and pre-processed mean spectra data. In Figure 4.10, there were slight resemblance of patterns for loading plot of PC1 on raw mean spectra data with the reflectance as displayed in Figure 4.3 and Figure 4.4, except on the wavelength range above 800 nm. Drastic change in loading was found starting from 600 nm for PC2, while some peaks were found between 600 nm and 800 nm for PC3. This could indicate spectral bands between 600 nm and 800 nm had significant impact for the results in score plots and explained variability.

For loading plots on SG pre-processed mean spectra data, it was hugely similar as those on raw mean spectra data. Next, in Figure 4.11, for SG-SNV pre-processed mean spectra data, the difference was clearly observed in the loading plot of PC1. Drastic change in loading values and many peaks were observed between 600 nm and 900 nm in both loading plots of PC2 and PC3. SG-SNV pre-processing made the spectral bands between 600 nm and 900 nm more clear and impactful, which further explained the normalisation effect of SNV. While for loading plots on SG-1st pre-processed mean spectra data, various peaks were observed in the loading plot of PC1, with the most prominent ones were in between 600 nm and 900 nm. Significant peak was observed around 800 nm for loading plots of PC2 and PC3, which explained separate clusters to be observed easily in PCA score plots as shown in Figure 4.8. Since the loading values were quite close to each other, this could be the reason why the explained variability was highly distributed among the PCs. In Figure 4.12, for loading plots on SG-2nd pre-processed mean spectra data, similar patterns could be observed as those on SG-1st pre-processed mean spectra data, hence explaining the distinct clustering observed in PCA score plots as shown in Figure 4.9.

Based on these observations, data pre-processing allowed the HSI mean spectral data to be interpreted easily and data exploration with PCA revealed which spectral bands in HSI mean spectra data were significant in qualitative manner. It was noted that PCA did not consider the relationship of independent variables with the dependent variable, thus the PCA score and loading plots might not be necessarily reflecting the relationship between wavelengths of mean spectra data and the target variables (Elmasry et al., 2012; September, 2011; Wu and Sun, 2013b). Model development and training was then proceeded after preliminary investigation on the HSI spectra data was finished since quantitative analysis was not shown in the PCA.



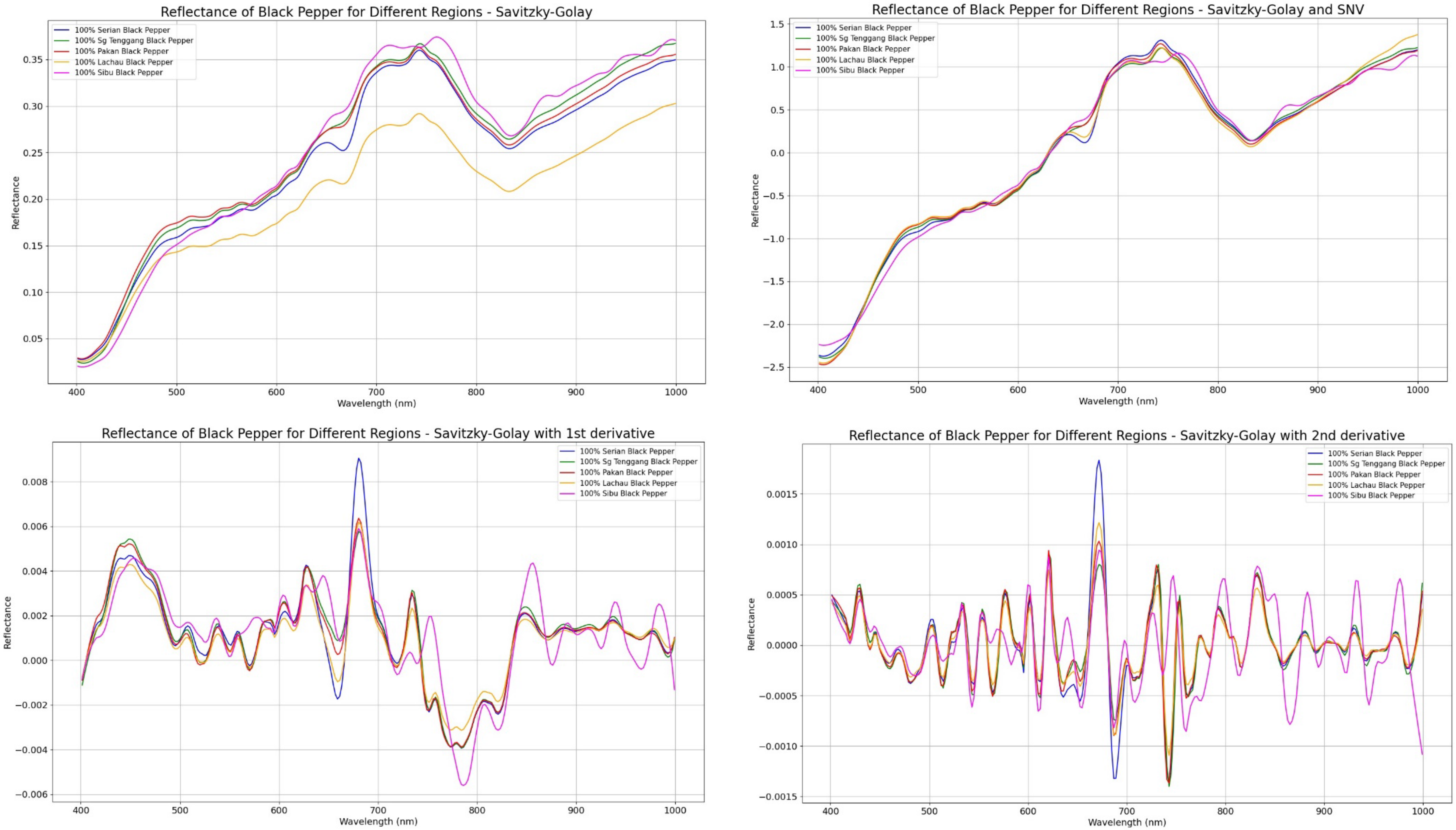
Reflectance of Savitzky-Golay Preprocessed Data

Reflectance of Savitzky-Golay and SNV Preprocessed Data

Reflectance of Savitzky-Golay with 1st derivative Preprocessed Data

Reflectance of Savitzky-Golay with 2nd derivative Preprocessed Data

Figure 4.3: Reflectance of black pepper powder samples from different range of degrees of adulteration using various data pre-processing methods



Reflectance of Black Pepper for Different Regions - Savitzky-Golay

Reflectance of Black Pepper for Different Regions - Savitzky-Golay and SNV

Reflectance of Black Pepper for Different Regions - Savitzky-Golay with 1st derivative

Reflectance of Black Pepper for Different Regions - Savitzky-Golay with 2nd derivative

Figure 4.4: Reflectance of pure (100%) black pepper powder samples from various origins using various data pre-processing methods

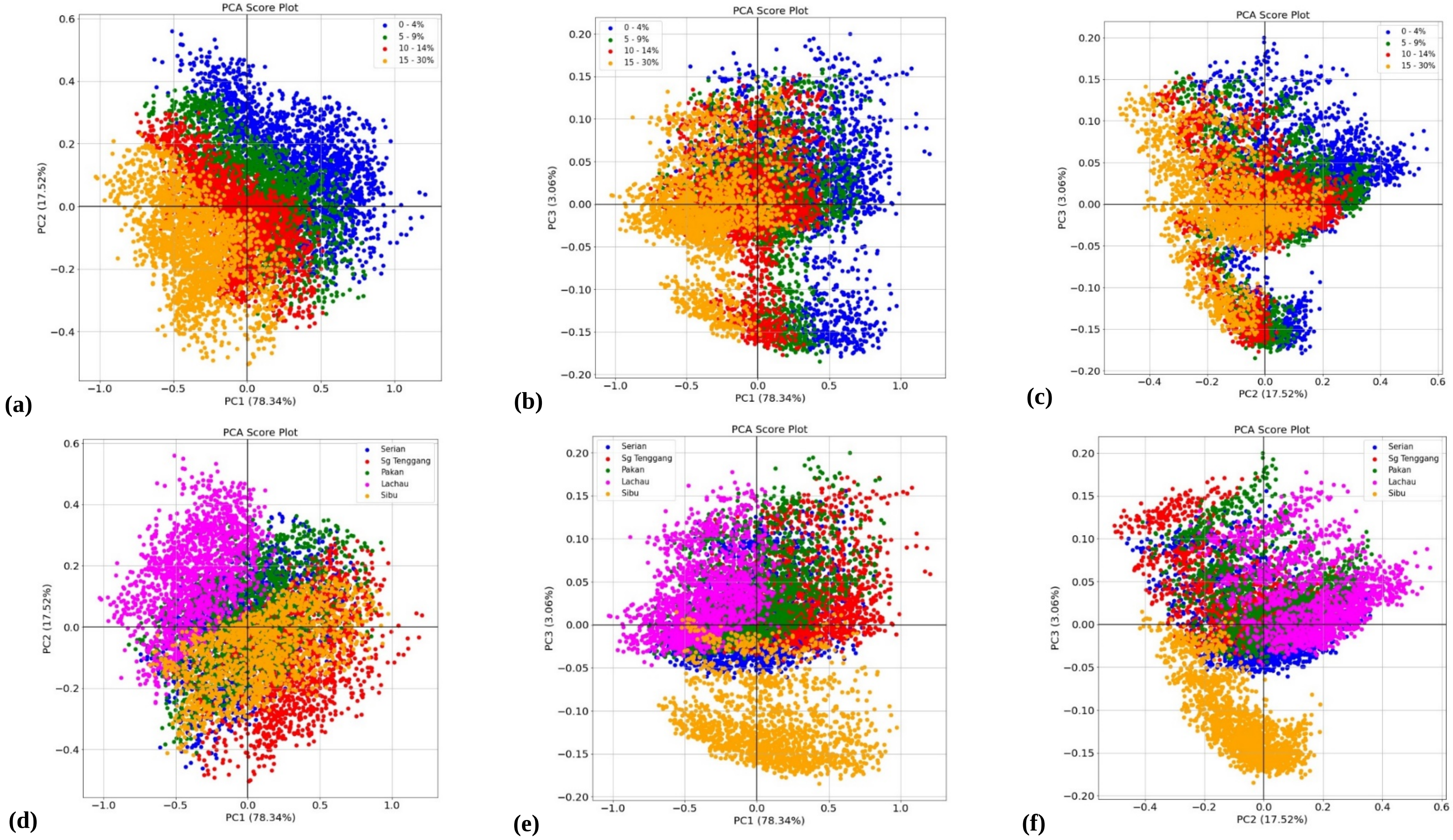


Figure 4.5: PCA score plots for first 3 principal components on raw mean spectra data: Based on percentage authenticity – (a) PC1 vs PC2, (b) PC1 vs PC3, (c) PC2 vs PC3; Based on geographical origins – (d) PC1 vs PC2, (e) PC1 vs PC3, (f) PC2 vs PC3

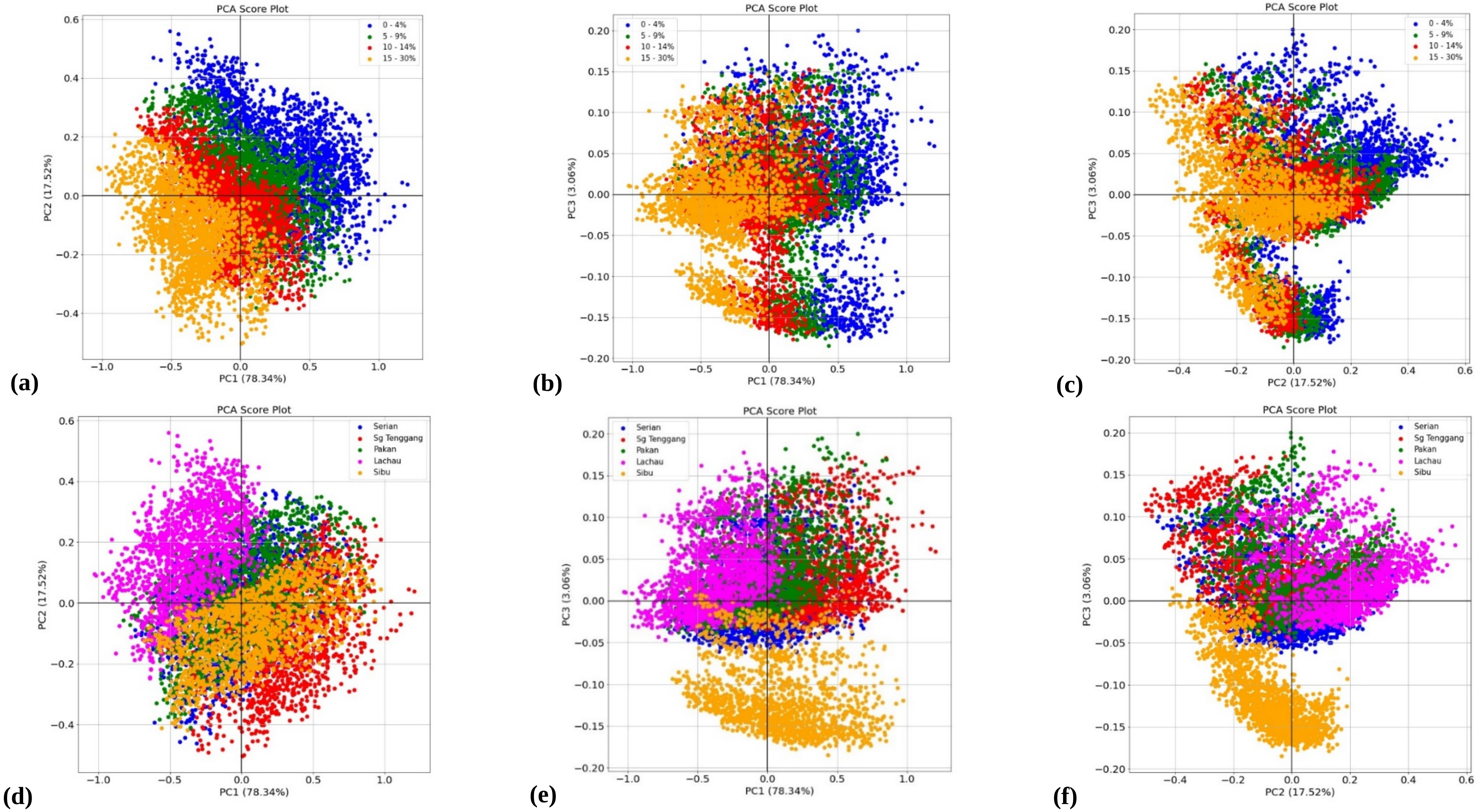


Figure 4.6: PCA score plots for first 3 principal components on SG pre-processed mean spectra data: Based on percentage authenticity – (a) PC1 vs PC2, (b) PC1 vs PC3, (c) PC2 vs PC3; Based on geographical origins – (d) PC1 vs PC2, (e) PC1 vs PC3, (f) PC2 vs PC3

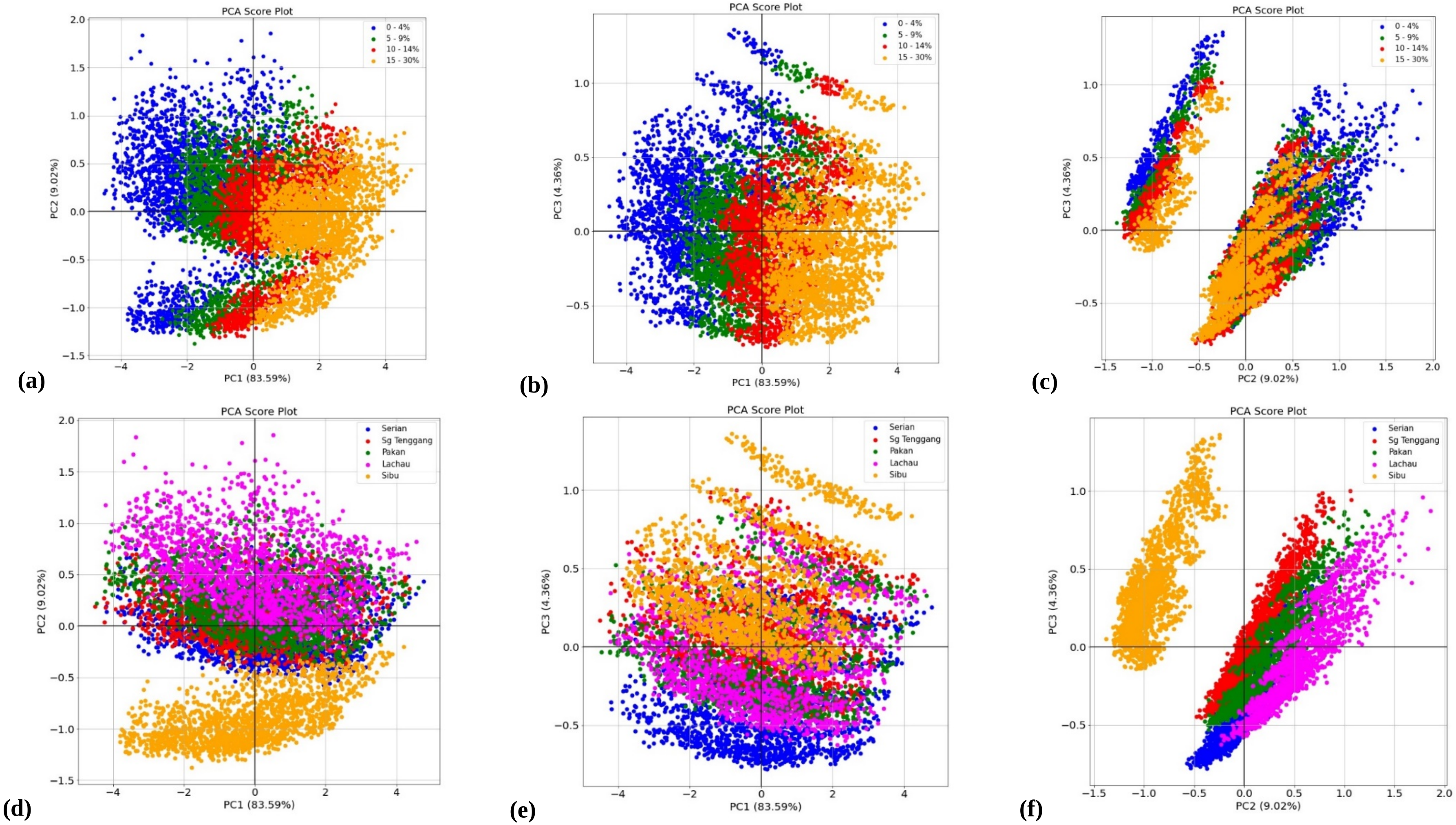


Figure 4.7: PCA score plots for first 3 principal components on SG-SNV pre-processed mean spectra data: Based on percentage authenticity – (a) PC1 vs PC2, (b) PC1 vs PC3, (c) PC2 vs PC3; Based on geographical origins – (d) PC1 vs PC2, (e) PC1 vs PC3, (f) PC2 vs PC3

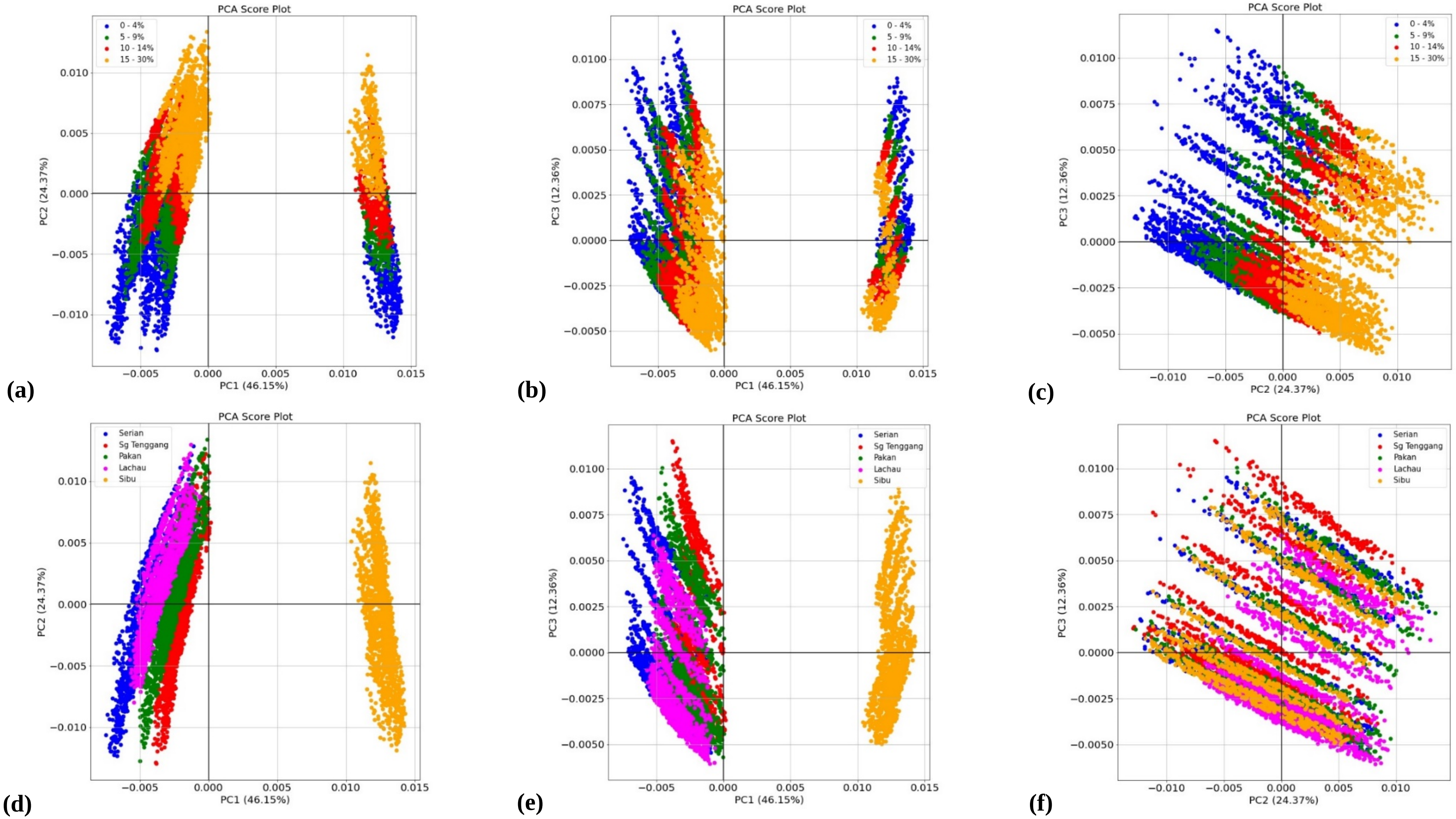


Figure 4.8: PCA score plots for first 3 principal components on SG-1st pre-processed mean spectra data: Based on percentage authenticity – (a) PC1 vs PC2, (b) PC1 vs PC3, (c) PC2 vs PC3; Based on geographical origins – (d) PC1 vs PC2, (e) PC1 vs PC3, (f) PC2 vs PC3

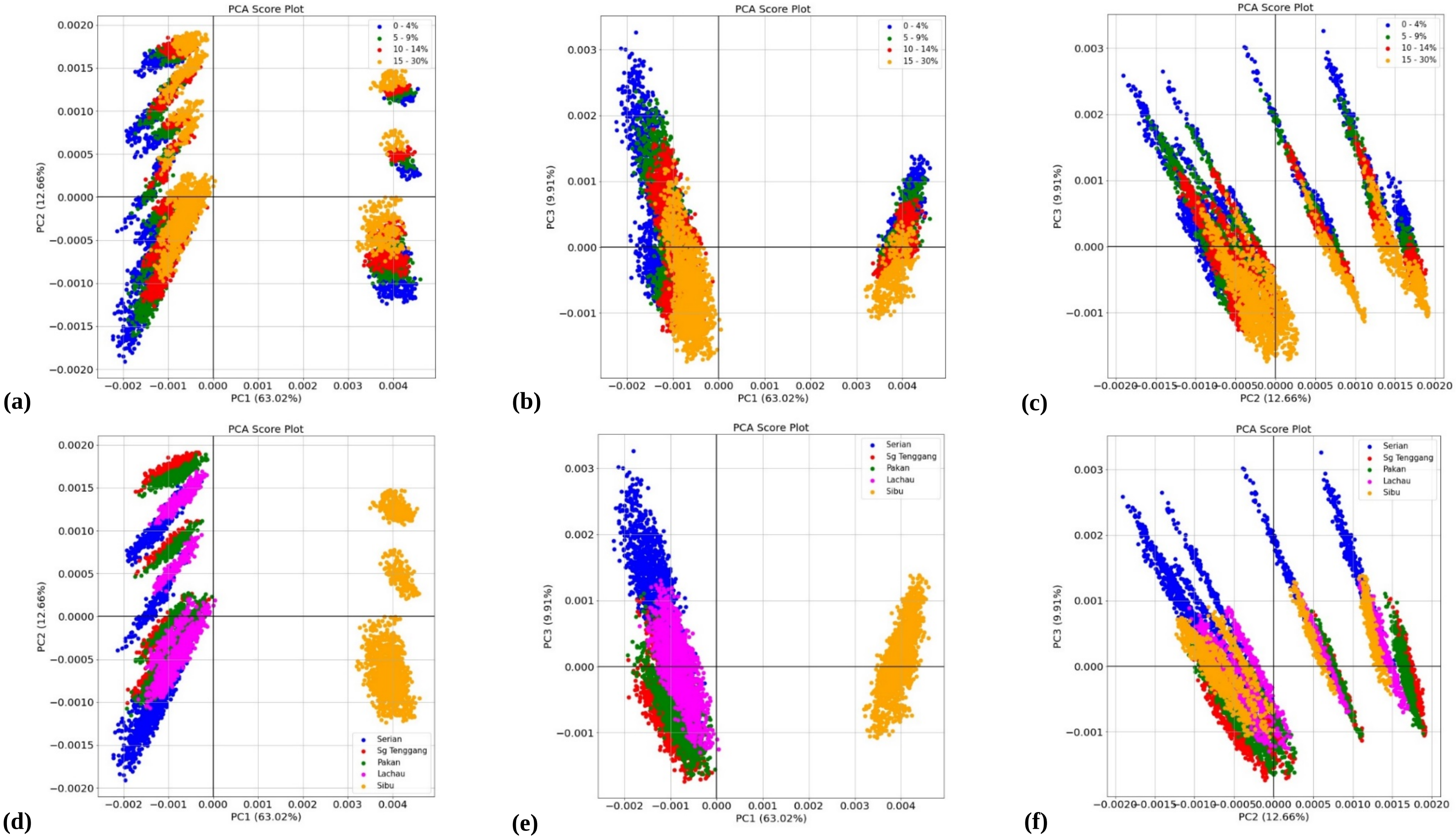


Figure 4.9: PCA score plots for first 3 principal components on SG-2nd pre-processed mean spectra data: Based on percentage authenticity – (a) PC1 vs PC2, (b) PC1 vs PC3, (c) PC2 vs PC3; Based on geographical origins – (d) PC1 vs PC2, (e) PC1 vs PC3, (f) PC2 vs PC3

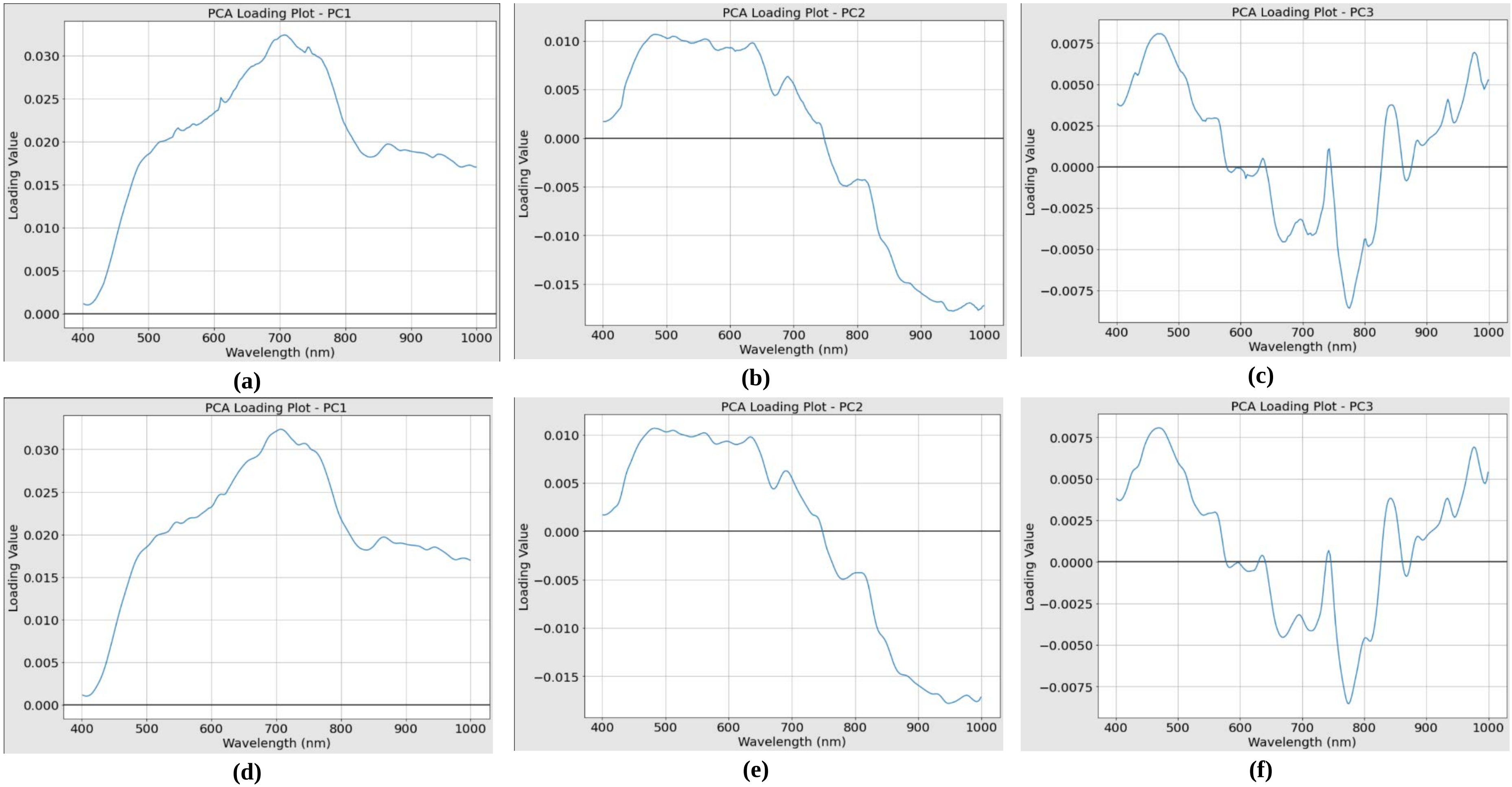
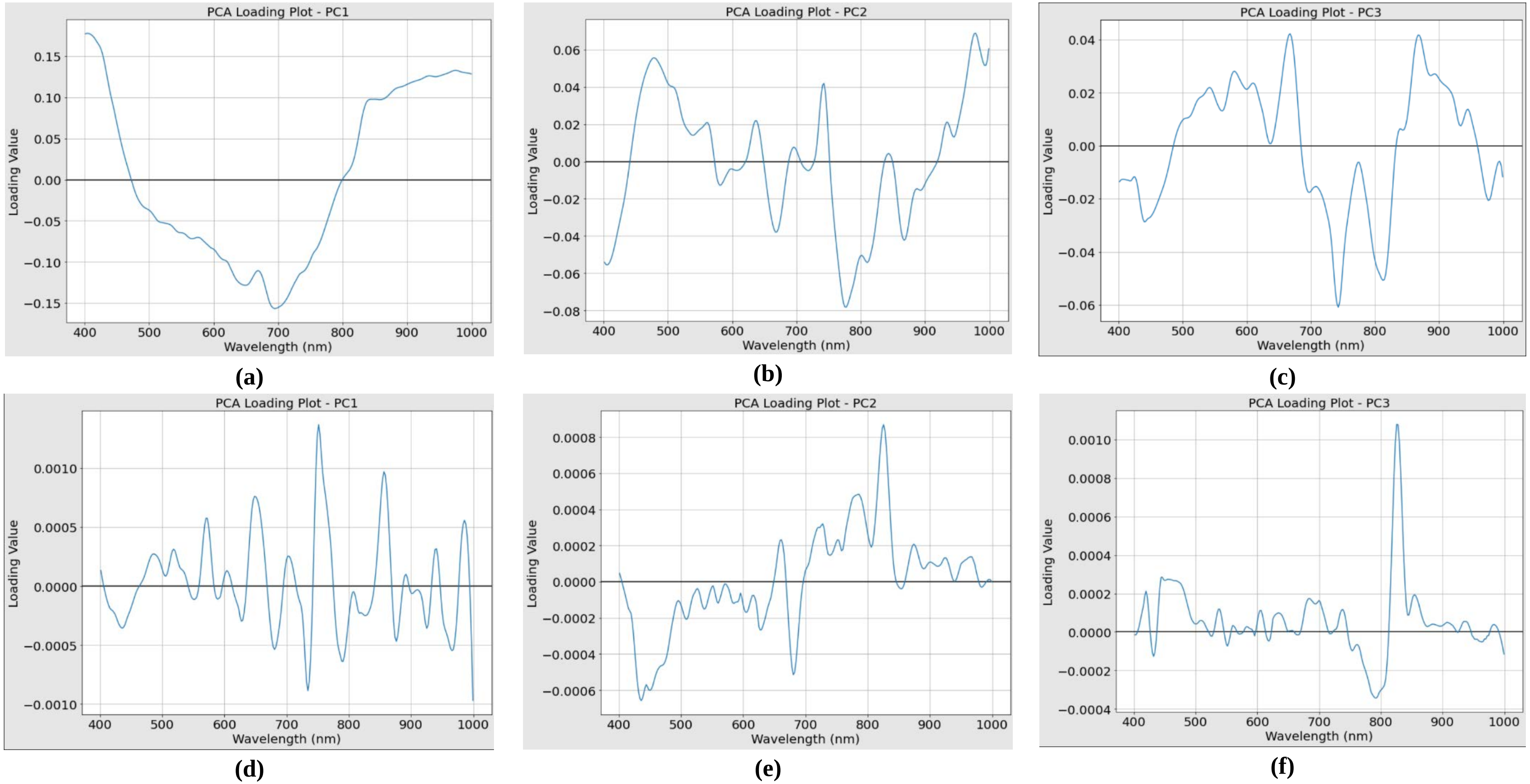


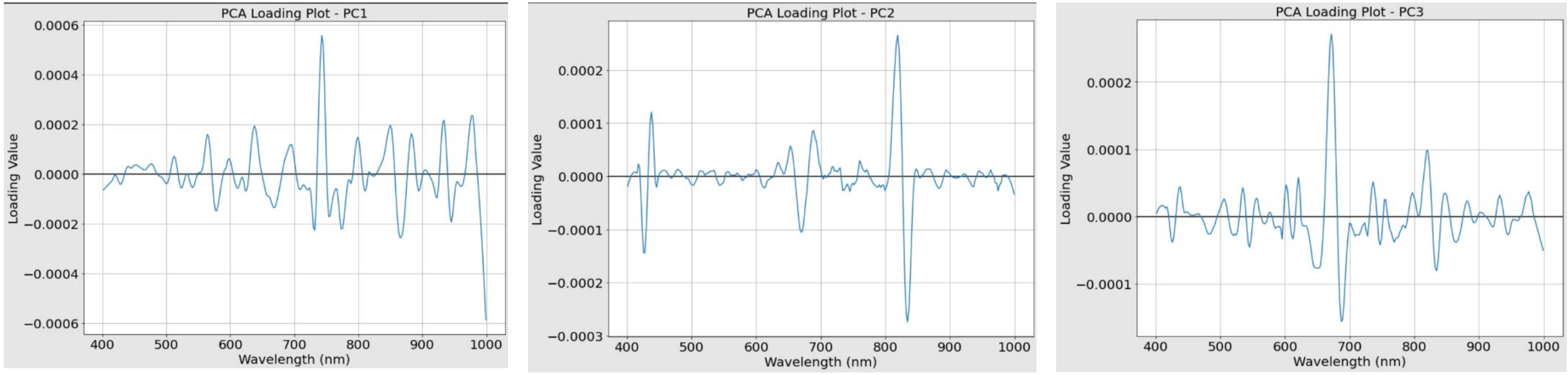
Figure 4.10: PCA loading plots for raw and SG pre-processed mean spectra data – (a) PC1, (b) PC2 and (c) PC3 for raw data; and (d) PC1, (e) PC2 and (f) PC3 for SG data



(a) (b) (c)

(d) (e) (f)

Figure 4.11: PCA loading plots for SG-SNV and SG-1st pre-processed mean spectra data – (a) PC1, (b) PC2 and (c) PC3 for SG-SNV data; and (d) PC1, (e) PC2 and (f) PC3 for SG-1st data



(a)

(b)

(c)

Figure 4.12: PCA loading plots for SG-2nd pre-processed mean spectra data – Based on (a) PC1, (b) PC2 and (c) PC3

4.2 Determination of Authenticity

Determining the authenticity, which is the purity of black pepper powder samples, was a regression task. The machine learning model was trained using supervised learning to learn the data and corresponding labels or targets. Machine learning techniques used in this research were PLS, SVR and DL. In DL, two model architectures were used: convolutional neural network (CNN) and modified stacked autoencoder (SAE).

4.2.1 Partial Least Square

PLS models were built and trained with all the raw and pre-processed mean spectra data. To obtain the best PLS model, the main parameter to be optimised is the number of latent variables. This optimisation was performed over a range of 50 latent variables to be based on the RMSE from the trained model tested on the validation data set. The final results could be seen in Table 4.2.

Table 4.2: Results of PLS models on regression of determination of authenticity of black pepper powder samples, bolded results are the best results among the testing data set

Pre-processing		Raw	SG	SG-SNV	SG-1st	SG-2nd
Best # Latent Variable		30	46	36	33	49
R^2	Train	0.9750	0.9556	0.9681	0.9513	0.9516
	Valid	0.9550	0.9555	0.9677	0.9517	0.9502
	Test	0.4327	0.4457	**0.4475**	0.4159	0.4298
RMSE	Train	0.0161	0.0164	0.0139	0.0171	0.0171
	Valid	0.0166	0.0166	0.0141	0.0173	0.0175
	Test	0.0343	0.0339	**0.0338**	0.0348	0.0344
MAPE (%)	Train	1.41	1.43	1.21	1.51	1.50
	Valid	1.48	1.46	1.24	1.53	1.56
	Test	2.91	2.89	**2.88**	3.03	2.98

From Table 4.2, the best results based on the testing data set were obtained on SG-SNV pre-processed mean spectra data. Its resulting R^2, RMSE and MAPE were 0.4475, 0.0338 and 2.88% respectively. Its optimisation on the best number of latent variables and subsequently parity plot were illustrated in Figure 4.13. Parity plots, plots of optimisation of latent variables and regression coefficients for PLS models with other data pre-processing methods could be referred to Figure B.3 and Figure B.4 in Appendix B.

It was observed that SG-1st and SG-2nd pre-processed data worsened the predictive performance of PLS model, contrary to most findings in the literature where the SG filter with derivatives should yield better estimates (McGoverin et al., 2012; Orrillo et al., 2019).

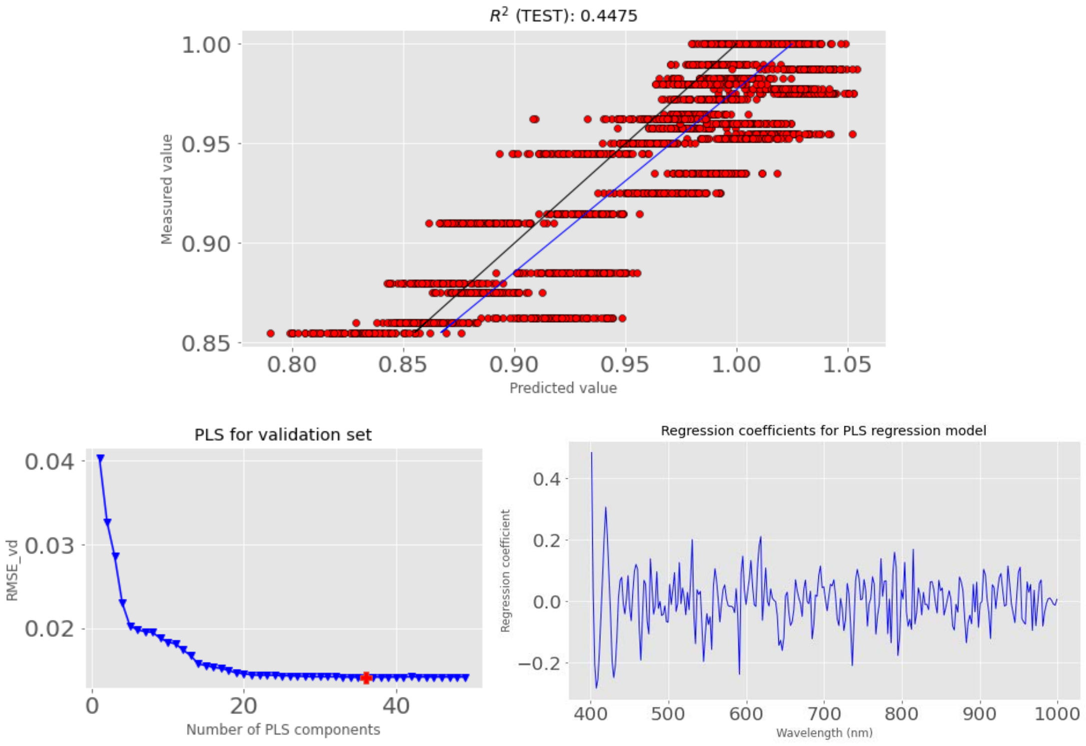


Figure 4.13: The best result of PLS model on SG-SNV pre-processed mean spectra data [Above – Parity plot, black line is 1:1 baseline, blue line is best fit line of predicted values; Below – (Left): Optimisation of number of latent variables based on validation RMSE, (Right): Regression coefficient of that resulting PLS model]

However, SG and SG-SNV pre-processing methods improved the predictive performance on testing data set, although to a lesser extent. It was observed that all PLS models such as in the parity plot of SG-SNV pre-processed PLS model yielded predicted values exceeding 1.0, causing the best fit line of PLS to project values beyond 1.0, thus further increasing discrepancies between predicted and true values. More importantly, all PLS models for all testing mean spectra data only managed to yield R^2 of less than 0.5, and it was only acceptable if R^2 is more than 0.9 (Yu et al., 2018). It was most likely due to non-linearity of the data, as the PLS model was a linear model and would yield high predictive performance on both training and testing data sets if the data was linear. Even after PLS transformation, the latent variables which captured the maximum variability could not extract representative features from the data easily.

Additionally, it was common that most machine learning models suffered from overfitting issues, where the PLS model had good training and validation predictive performance (R^2 more than 0.95), but worse testing predictive performance. This could be attributed to conflicting data where the models found the mean spectra of one testing sample belonging to certain label similar to the mean spectra of one trained sample but belonging to another label. The MAPE, which determined the mean absolute error among the predicted and measured values, was within acceptable margin of 10%, where more than 10% indicating the predictions strayed away from intended measured targets.

From Figure 4.13, plot of regression coefficients over wavelengths for the PLS model on SG-SNV mean spectra data was also included to investigate how spectral bands influence the final prediction results. From there, most prominent peaks were located below 750 nm, which is visible light region but fewer peaks could be located above 750 nm. For example, positive peaks on around 420, 520 and 615 nm and negative peaks on around 435, 460, 545, 595, 745 and 800 nm were possible parts of factors which influenced the PLS regression results. Nevertheless, different machine learning techniques were to be explored for better prediction of authenticity of Sarawak black pepper powder products.

4.2.2 Support Vector Regression

Different SVR models were created for each data pre-processing method. Grid search algorithm was used to find the optimal combination of C and γ for each SVR model of different data pre-processing method. The final results of all SVR models could be seen in Table 4.3. Parity plot of the best SVR model based on SG-SNV pre-processed data was plotted in Figure 4.14, with the rest of parity plots for other SVR models could be referred to Figure B.6 in Appendix B.

Table 4.3: Results from SVR model on regression of determination of authenticity of black pepper powder samples, bolded results are the best results among the testing data set

Pre-processing		Raw	SG	SG-SNV	SG-1st	SG-2nd
Best C		1	1	10	100	10
Best γ		10	10	0.001	10	10000
R^2	Train	0.9776	0.9773	0.9565	0.9654	0.9752
	Valid	0.9754	0.9751	0.9574	0.9666	0.9741
	Test	0.7855	0.7954	**0.8032**	0.7559	0.7197
RMSE	Train	0.0116	0.0117	0.0162	0.0144	0.0122
	Valid	0.0123	0.0124	0.0162	0.0143	0.0126
	Test	0.0211	0.0206	**0.0202**	0.0225	0.0241
MAPE (%)	Train	0.97	0.98	1.40	1.23	1.03
	Valid	1.03	1.04	1.40	1.24	1.06
	Test	1.72	**1.66**	1.75	1.98	2.09
# Support Vectors		1815	1833	2680	2334	1945

From Table 4.3, the best results were also from the SVR model for SG-SNV pre-processed mean spectra data where the R^2, RMSE and MAPE were 0.8032, 0.0202 and 1.75% respectively. Due to the standardisation of mean spectra data by SNV, the model predictive

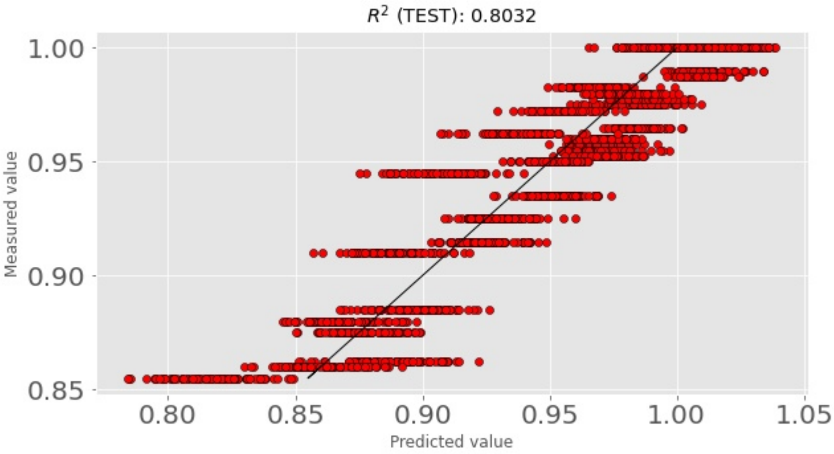


Figure 4.14: Parity plot of SVR model using SG-SNV pre-processed mean spectra data

performance could be improved. Additionally, the radial basis function kernel was a Gaussian function and normalisation usually followed Gaussian distribution (Goodfellow et al., 2016). All SVR models managed to perform better than PLS models, with most R^2 achieving more than 0.7, since the radial basis function kernel in SVR model could translate the mean spectra data to make it more separable, hence making SVR model more robust and able to account for non-linearity and unknown testing data set easier than PLS model (Goodfellow et al., 2016; Z. Zhou et al., 2019). The issue of data with conflicting labels was still present but SVR model seemed to deal with that appropriately. There was also a substantial reduction in MAPE compared to that of PLS model. Even so, from Figure 4.14, SVR model yielded predicted values exceeding 1.0, which was outside appropriate range in determining the authenticity of black pepper powder samples to begin with. Additionally, several predicted values were not close to measured values.

SVR models selected the input mean spectra data as support vectors for the decision function (i.e. radial basis function kernel), thus they were based on full spectra range instead of certain characteristic wavelengths or bands. The number of support vectors usually indicates the size of model to determine the best fit decision function. Referring to Table 4.3, somehow the SVR model using raw mean spectra data had the least number of support vectors. While SVR models yielded decent predictions in determining the authenticity of black pepper powder samples, it was of interest to investigate whether deep learning models performed better in this regression task or not.

4.2.3 Deep Learning

Deep learning (DL) model was a black box model entirely dependent on the data. CNN and SAE model architectures were considered in this research due to majority of researches in determining the quality of food and agricultural products applied these architectures (Saha and Manickavasagan, 2021). Both CNN and SAE models were not pre-trained (training the model with other mean spectra data before this training happened), tested and compared to investigate which model functions better for regression task.

4.2.3.1 Convolutional Neural Network

In CNN model, the convolution operations were responsible to extract locally correlated features present in the spectra. The design of the CNN model was based on the design from Han et al. (2021). There were total of 16 different model types to be assessed which involved altering of the kernel sizes in convolutional filters for each convolution layer, and could be referred to Table 4.4. The reason of using convolution kernel size of 1 or 3 was smaller kernel is able to capture finer features or details rather than obvious features (Han et al., 2021). The layer configuration or number of convolutional kernels in all convolutional layers was kept to be 16-32-64-128 for all model types, where the number being set to be base 2 was common in most works and that configuration generally worked well (Blazhko et al., 2021; Goodfellow et al., 2016; Han et al., 2021; Rong et al., 2019). As most spectra data were mostly similar to each other especially relationship among the bands or spectral features, it was beneficial to find for these local features so that the authenticity and geographical origins of black pepper powder samples easily could be predicted easily.

Table 4.4: Different DL CNN models with respective definitions of kernel sizes of each layer

Model Type	1D Kernel size (Layer 1-2-3-4)			
1	1	1	1	1
2	1	1	1	3
3	1	1	3	1
4	1	1	3	3
5	1	3	1	1
6	1	3	1	3
7	1	3	3	1
8	1	3	3	3
9	3	1	1	1
10	3	1	1	3
11	3	1	3	1
12	3	1	3	3
13	3	3	1	1
14	3	3	1	3
15	3	3	3	1
16	3	3	3	3

All CNN models had in average of around 300,000 trainable parameters. The final results on the testing data sets were displayed in Table 4.5. For the training and validation results, they could be referred to Table B.1 in Appendix B.

Table 4.5: Results from DL CNN models for regression, bolded results are the best results from testing data set for respective data pre-processing method

Indicator	Model Type	Pre-processing				
		Raw	SG	SG-SNV	SG-1st	SG-2nd
R^2 Testing	1	0.7272	0.7799	0.7449	0.6898	0.6525
	2	0.7466	0.7876	**0.8671**	0.8322	0.5471
	3	0.7347	0.8297	0.8663	0.8420	**0.7052**
	4	0.7439	0.7577	0.7679	0.8040	0.5899
	5	0.7740	0.7931	0.8372	0.7872	-1.6461
	6	0.7001	0.7579	0.7323	0.7643	0.6522
	7	0.7571	0.7568	0.8137	0.7760	0.6507
	8	0.7208	0.7898	0.8246	0.8271	0.4836
	9	0.8370	0.8509	0.8172	0.3479	-1.2065
	10	0.7816	0.7877	0.7530	-2.0652	-0.4153
	11	0.8284	**0.8663**	0.8412	0.1773	0.6151
	12	0.8153	0.8335	0.8018	0.7658	0.5344
	13	0.8307	0.8642	0.8319	**0.8430**	-2.1639
	14	0.8088	0.7078	0.8019	0.8025	0.5607
	15	0.7720	0.8471	0.8540	0.7593	0.4030
	16	**0.8507**	0.8176	0.8391	0.7883	0.6572
Indicator	**Model Type**	**Pre-processing**				
		Raw	**SG**	**SG-SNV**	**SG-1st**	**SG-2nd**
RMSE Testing	1	0.0238	0.0213	0.0230	0.0253	0.0268
	2	0.0231	0.0210	**0.0166**	0.0186	0.0306
	3	0.0234	0.0188	0.0166	0.0181	**0.0247**
	4	0.0230	0.0228	0.0219	0.0201	0.0291
	5	0.0211	0.0207	0.0184	0.0210	0.0740
	6	0.0249	0.0224	0.0236	0.0221	0.0268
	7	0.0224	0.0222	0.0196	0.0215	0.0269
	8	0.0233	0.0209	0.0191	0.0189	0.0327
	9	0.0184	0.0176	0.0195	0.0367	0.0676
	10	0.0213	0.0210	0.0226	0.0796	0.0541
	11	0.0188	**0.0166**	0.0181	0.0413	0.0282
	12	0.0196	0.0186	0.0203	0.0220	0.0310
	13	0.0187	0.0168	0.0187	**0.0180**	0.0809
	14	0.0199	0.0246	0.0202	0.0202	0.0302
	15	0.0217	0.0178	0.0174	0.0223	0.0352
	16	**0.0176**	0.0194	0.0183	0.0209	0.0266
Indicator	**Model Type**	**Pre-processing**				
		Raw	**SG**	**SG-SNV**	**SG-1st**	**SG-2nd**
	1	1.89	1.71	1.95	2.04	2.19

MAPE Testing (%)	**2**	1.85	1.60	**1.38**	1.55	2.42
	3	1.88	1.42	1.40	1.52	**2.03**
	4	1.84	1.83	1.80	1.62	2.29
	5	1.75	1.74	1.58	1.72	5.56
	6	2.03	1.87	1.90	1.79	2.21
	7	1.77	1.75	1.64	1.78	2.25
	8	1.92	1.76	1.55	**1.51**	2.68
	9	1.51	1.40	1.65	2.54	4.50
	10	1.79	1.76	1.90	5.80	4.30
	11	1.60	**1.39**	1.52	3.33	2.42
	12	1.72	1.55	1.79	1.86	2.46
	13	1.56	1.40	1.54	1.55	6.37
	14	1.61	1.97	1.72	1.73	2.38
	15	1.74	1.43	1.41	1.80	2.86
	16	**1.44**	1.51	1.56	1.77	2.17

Table 4.6: Example of DL CNN model architecture

Layer	**Output Shape**	**# of Parameters**	**Connected To**
Input Layer	(280, 1)	0	-
Conv1D_1 *[kernel size = 1]*	(280, 16)	32	Input Layer
BatchNormalization_1	(280, 16)	64	Conv1D_1
MaxPool1D_1	(140, 16)	0	BatchNormalization_1
Conv1D_2 *[kernel size = 1]*	(140, 32)	544	MaxPool1D_1
BatchNormalization_2	(140, 32)	128	Conv1D_2
MaxPool1D_2	(70, 32)	0	BatchNormalization_2
Conv1D_3 *[kernel size = 1]*	(70, 64)	2112	MaxPool1D_2
BatchNormalization_3	(70, 64)	256	Conv1D_3
MaxPool1D_3	(35, 64)	0	BatchNormalization_3
Conv1D_4 *[kernel size = 3]*	(35, 128)	24704	MaxPool1D_3
BatchNormalization_4	(35, 128)	512	Conv1D_4
MaxPool1D_4	(17, 128)	0	BatchNormalization_4
Flatten	(2176)	0	MaxPool1D_4
Dense_1	(128)	278656	Flatten
Dense_2	(128)	16512	Dense_1
Dense_3 *[authenticity]*	(1)	129	Dense_2
Dense_4 *[geog. origin]*	(5)	645	Dense_2
Total of trainable number of parameters = **323,814**			

From Table 4.5, the best result was the CNN model on SG-SNV pre-processed mean spectra data in Model Type 2 where the kernel sizes for each layer were [1-1-1-3]. An outline of the best CNN model was displayed in Table 4.6. Its resulting R^2, RMSE and MAPE were

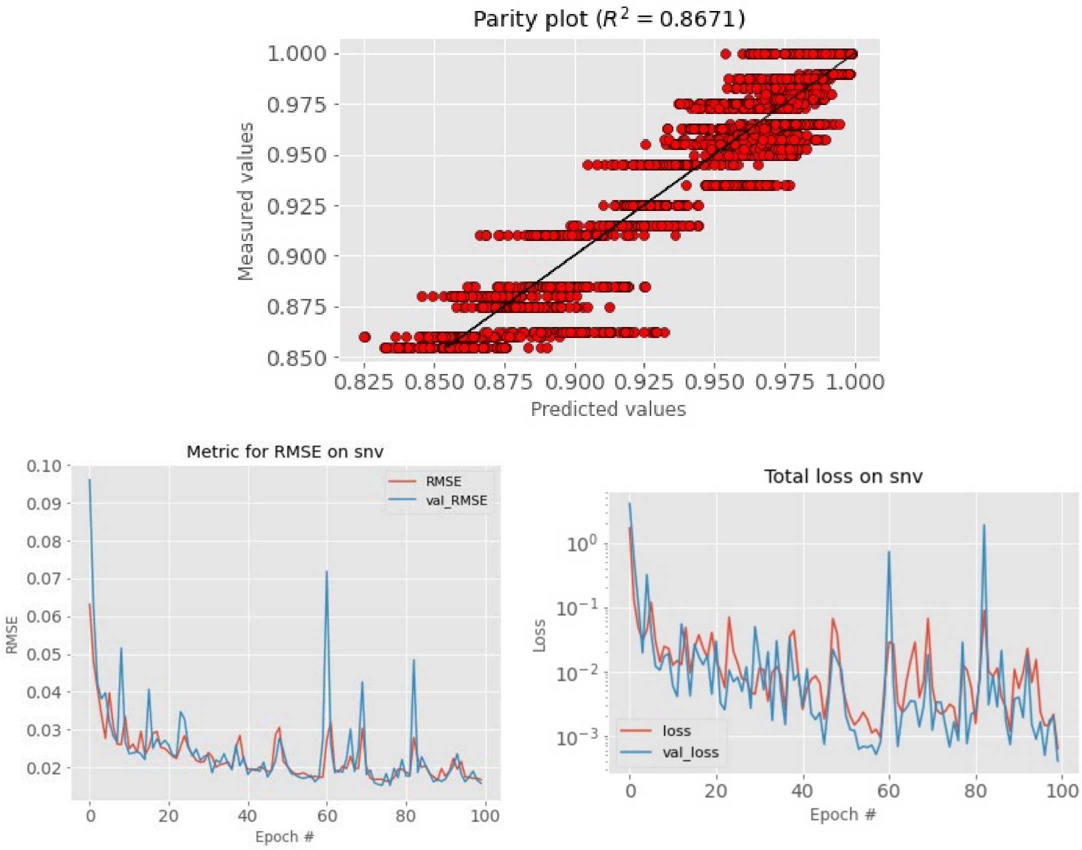


Figure 4.15: Best results of DL CNN model on SG-SNV pre-processed data (Model Type 2) [Above – Parity plot; Below – Loss and metric over epochs plots to monitor the generalizability of the model (Left) RMSE metric over epochs, (Right) Total loss over epochs

0.8671, 0.0166 and 1.38% respectively. As shown on its parity plot in Figure 4.15, thanks to sigmoid function which appropriately limited the output, this resulted in predictions with higher precision with the measured values. Plots of losses and RMSE metric over epochs were also shown in Figure 4.15 to assess the generalizability of the model, which was Model Type 2. From those plots, it was safe to indicate that the model did not overfit on validation data set, where the validation loss were usually higher than training loss on overfitting case (Goodfellow et al., 2016).

The Model Type 3 (layer configuration [1-1-3-1]) for SG-SNV pre-processed data had similar predictive performance as Type 2 with R^2, RMSE and MAPE were 0.8663, 0.0166 and 1.40% respectively. Another model which has closer predictive performance with SG-SNV pre-processed data was SG pre-processed model of Type 11 (layer configuration [3-1-3-1]). Although the CNN model had already decent predictive performance on raw mean spectra data where the best model type was Model Type 16 (layer configuration [3-3-3-3]), its R^2, RMSE and MAPE were 0.8507, 0.0176 and 1.44% respectively, the data pre-processing on mean spectra data had further improved the CNN model predictive performance, albeit to a smaller extent. Since it was highly desirable to get lightweight model, Model Type 2 for SG-SNV pre-processed data with the best predictive performance was hence selected.

On the contrary, SG with 2nd derivative pre-processing seemed to cause opposite effects on the predictive performance of DL CNN models, where the final training and validation predictive performance were worse, as per reference to Table B.1 in Appendix B. This could be attributed to the many redundant and conflicting information present in the derivatives of the data were trained, subsequently making wrong predictions. It was well noted that CNN models with batch normalisation hugely improved the generalizability of CNN models. Dropouts were considered but they were found to impede the predictive performance of the model since batch normalisation components were already present in regularising the outputs (Goodfellow et al., 2016; Yang et al., 2019).

4.2.3.2 Stacked Autoencoder

SAE models were also explored as the encoding part of SAE could extract features by mapping values through the reduced dimensional space. Most settings and configurations were similar to CNN modelling, but the convolution part was substituted with SAE encoding part. In average, there were 110,000 trainable parameters, making them faster to train than CNN models. Assessments for different model types where the layer configuration of encoding part was altered were performed as shown in Table 4.7. The final results of all SAE models for different data pre-processing methods on testing data sets could be seen in Table 4.8. For the training and validation results, they could be referred to Table B.2 in Appendix B.

Table 4.7: DL SAE models with different definitions of layer configuration

Model Type	Layer Configuration
1	(192, 16)
2	(192, 128, 16)
3	(192, 128, 64, 16)
4	(192, 128, 64, 32, 16)
5	(192, 128, 96, 64, 32, 16)

Table 4.8: Results from DL SAE model on regression, bolded results are the best results from testing data set for respective data pre-processing method

Pre-processing	Raw	SG	SG-SNV	SG-1st	SG-2nd
Model Type	R^2 Testing				
1	0.8647	**0.8807**	0.8649	0.8372	0.4421
2	**0.8801**	0.8424	0.8752	**0.8784**	**0.4634**
3	0.8456	0.7807	0.7720	0.7883	-0.1915

4	0.7369	0.8347	0.8665	0.8173	0.1285
5	0.8715	0.8504	**0.9010**	0.7509	0.3989
Model Type	**RMSE Testing**				
1	0.0167	**0.0157**	0.0167	0.0184	0.0340
2	**0.0158**	0.0181	0.0161	**0.0159**	**0.0333**
3	0.0179	0.0213	0.0219	0.0209	0.0497
4	0.0233	0.0185	0.0166	0.0194	0.0425
5	0.0163	0.0176	**0.0143**	0.0227	0.0353
Model Type	**MAPE Testing (%)**				
1	1.42	**1.34**	1.38	1.60	2.69
2	1.38	1.42	1.34	**1.30**	2.64
3	1.44	1.59	1.61	1.68	3.38
4	1.69	1.47	1.37	1.50	3.20
5	**1.36**	1.35	**1.17**	1.83	**2.60**

Table 4.9: Example of DL SAE model architecture

Layer	**Output Shape**	**# of Parameters**	**Connected To**
Input Layer	(280)	0	-
Dense_1	(192)	53952	Input Layer
BatchNormalization_1	(192)	768	Dense_1
Dense_2	(128)	24704	BatchNormalization_1
BatchNormalization_2	(128)	512	Dense_2
Dense_3	(96)	12384	BatchNormalization_2
BatchNormalization_3	(96)	384	Dense_3
Dense_4	(64)	6208	BatchNormalization_3
BatchNormalization_4	(64)	256	Dense_4
Dense_5	(32)	2080	BatchNormalization_4
BatchNormalization_5	(32)	128	Dense_5
Dense_6	(16)	528	BatchNormalization_5
BatchNormalization_6	(16)	64	Dense_6
Dense_7	(128)	2176	BatchNormalization_6
Dense_8	(128)	16512	Dense_7
Dense_9 *[authenticity]*	(1)	129	Dense_8
Dense_10 *[geog. origin]*	(5)	645	Dense_8
Total of trainable number of parameters = **120,374**			

As a result, the best DL SAE model was based on Model Type 5 (layer configuration [192-128-96-64-32-16]) and SG-SNV pre-processed mean spectra data, where the testing R^2, RMSE and MAPE were 0.9010, 0.0143 and 1.17% respectively. Its model architecture and parity plot could be referred to Table 4.9 and Figure 4.16 respectively. Batch normalisation components worked well with the SG-SNV pre-processed mean spectra data, resulting in high quality predictions. From Figure 4.16, plots of losses and RMSE metric over epochs were presented. There were no obvious signs of overfitting found in those plots, and eventually this resulted in the model with best generalisation.

It was found out that for all model types, all SAE models underwent similar trends in losses over epochs where it took numerous epochs for all these models to reduce losses to a reasonable degree (down to 10^{-3}), unlike CNN models where the loss reduction was gradual. It was inferred that the updating of weights in Dense functions of feature extractor section required several runs to stabilise, compared to convolution filters which were more stable to update. Interestingly, the training and validation performance for SAE models were consistently more stable than CNN. SAE models had comparable predictive performance in most regression tasks to predict numerical properties as demonstrated in most literature (Xin et al., 2020; Yang et al., 2019; Yu et al., 2018; C. Zhang et al., 2020).

Based on all the results from PLS, SVR, CNN and SAE models, it was evident that DL models performed better than most conventional machine learning techniques. Among all data pre-processing methods, SG-SNV pre-processed mean spectra data enabled all types of models to yield the best predictive performance in determining the authenticity of Sarawak black pepper powder samples, and DL-SAE model had the best predictive performance among all the models with R^2 of 0.9010. Additionally, Visible-NIR HSI had quite considerable success in determining the authenticity of powder samples, as most researches using Visible-

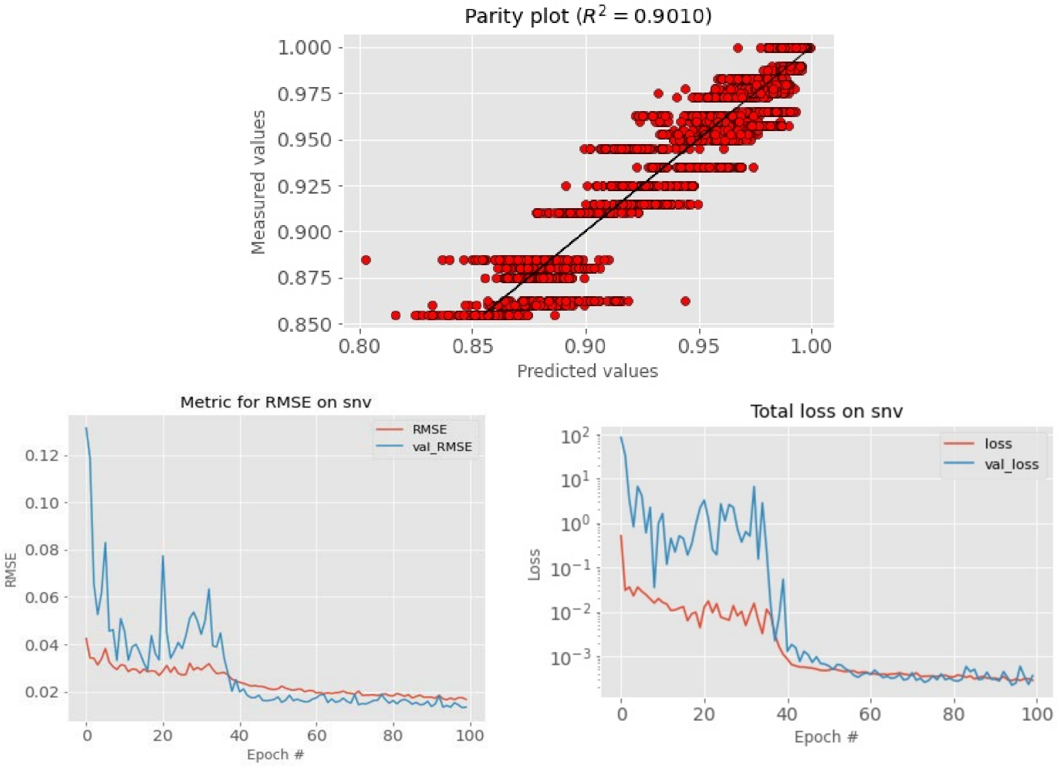


Figure 4.16: Best results of DL SAE model on SG-SNV pre-processed data (Model Type 5) [Above – Parity plot; Below – Loss and metric over epochs plots to monitor the generalizability of the model (Left) RMSE metric over epochs, (Right) Total loss over epochs

NIR HSI were towards classification and determination of quality of non-powder opaque objects due to distinct dissimilarity while for black pepper powder samples, NIR spectral range with further wavelength range was used instead (Hu et al., 2018; Lima et al., 2020; McGoverin et al., 2012; Orrillo et al., 2019).

4.3 Classification of Geographical Origin

Classification of geographical origin involves classification and thus the machine learning techniques were required to be adjusted accordingly. PLS with discriminant analysis (DA), SVM and DL were applied in classification task. For DL, the CNN and SAE models were trained as outlined in Section 4.2.3.1 to yield multiple outputs including classification of geographical origins.

4.3.1 Partial Least Square – Discriminant Analysis

PLS-DA was performed based on the threshold of predicted values by PLS regression to determine the class of the predicted sample. The parameter to be optimised was still the number of latent variables and the validation accuracy was compared. The final results were tabulated in Table 4.10. The best PLS-DA model with the minimum number of latent variables was based on the SG-SNV pre-processed mean spectra data with the highest accuracy of 68.07%, which was considered subpar. Based on its confusion matrix in Table 4.11, Sibu samples could be easily classified due to its uniqueness present in the mean spectra data, while samples of other regions could not be identified easily due to their spectral similarities. On the other hand, PLS-DA models were unable to learn the mean spectra data well with SG-1st and SG-2nd derivatives pre-processing methods, mainly because PLS-DA models were unfit for non-linear mean spectra data (Elmasry et al., 2012; Li et al., 2019; Petersson et al., 2016; Yang et al., 2019). While for models with other data pre-processing methods, the predictions were more towards misclassification of Pakan samples instead.

Table 4.10: Results from PLS-DA model on classification of geographical origins of black pepper powder samples, bolded results are the best results among the testing data set

Pre-processing	Best # Latent Variables	Accuracy (%)		
		Train	Valid	Test
Raw	37	98.96	98.72	63.94
SG	42	98.89	98.64	63.52
SG-SNV	33	97.39	97.82	**68.07**
SG-1st	49	98.59	98.64	67.82
SG-2nd	49	98.43	98.31	64.20

Table 4.11: Confusion matrix of PLS-DA model using SG-SNV pre-processed data

		Predicted Category				
		Serian	Sg Tenggang	Pakan	Lachau	Sibu
True Category	Serian	468	18	0	0	0
	Sg Tenggang	240	246	0	0	0
	Pakan	0	325	161	0	0
	Lachau	0	0	167	238	0
	Sibu	0	0	0	0	486

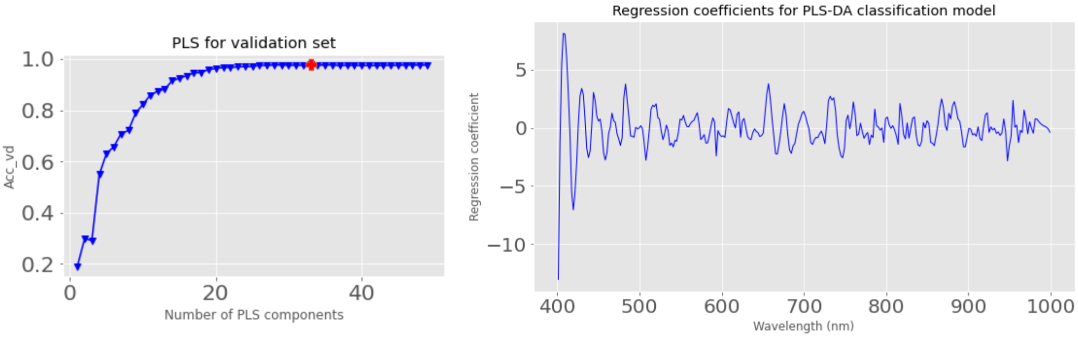


Figure 4.17: Plots of PLS-DA model on SG-SNV pre-processed mean spectra data, (Left): Optimisation of number of latent variables based on validation accuracy, (Right): Regression coefficient of that resulting PLS-DA model

Plots of optimisation of latent variables and regression coefficients of PLS-DA model on SG-SNV pre-processed data were displayed in Figure 4.17. Models of other pre-processing methods were shown in Figure B.5 of Appendix B. From the plot of regression coefficients in Figure 4.17, most coefficients were similar to each other and only a few notable peaks could be found on visible light region which is below 700 nm, such as around 420, 480 and 660 nm. Considering the low classification accuracy by the PLS-DA models, other machine learning techniques were then explored and assessed.

4.3.2 Support Vector Machine

SVM model was used specifically for classification tasks. Similar to how SVR modelling was performed, the final SVM models with the best C and γ on all mean spectra data were trained. The results were then displayed in Table 4.12. Table 4.13 displayed the confusion matrix of the best result which was from the SG-1st pre-processed SVM model. The best accuracy was now 100%, thanks to high capability of SVM to discriminate non-linear spectral data. The SG-1st pre-processed SVM model had the least number of support vectors, 15. It was easier for SVM to have higher predictive classification performance than SVR on determination of authenticity of black pepper powder samples because the mean spectra data were distinct enough in conjunction with discrete values of regions for training SVM model to classify the origins of black pepper powder samples. This resulted in easier construction of feature space by SVM model, hence lesser number of support vectors compared to SVR model which had to consider highly variable and continuous values in the determination of authenticity of black pepper.

Table 4.12: Results of SVM models on classification of geographical origins of black pepper powder samples, bolded results are the best results among the testing data set

Pre-processing	Best C	Best γ	# Support Vectors	Accuracy (%)		
				Train	Valid	Test
Raw	100	0.1	206	99.51	99.22	99.87
SG	100	0.1	206	99.49	99.22	99.87
SG-SNV	100	0.1	44	99.98	100	99.83
SG-1st	100	1000	**15**	100	100	**100**
SG-2nd	100	10000	23	100	100	99.62

Table 4.13: Confusion matrix of SVM model using SG-1st pre-processed data

		Predicted Category				
		Serian	Sg Tenggang	Pakan	Lachau	Sibu
True Category	Serian	486	0	0	0	0
	Sg Tenggang	0	486	0	0	0
	Pakan	0	0	486	0	0
	Lachau	0	0	0	405	0
	Sibu	0	0	0	0	486

4.3.3 Deep Learning

Deep learning models used for the classification task were CNN and SAE models. The settings, definition, configurations and assessment of different model types were from Sections 3.5.1 and 4.2.3, as all the models were trained to be multi-output, allowing simultaneous different outputs to be observed easily.

4.3.3.1 Convolutional Neural Network

Table 4.14 displayed the testing classification results of CNN models on all raw and pre-processed mean spectra data for all model types. Overall, except for SG-SNV and SG-2nd pre-processed mean spectra data, all CNN models on all data were able to yield 100% classification accuracy. Especially for the raw mean spectra data, CNN models were able to effectively extract spectral features from the data and classify the origins of black pepper powder samples, which was proven to function well without much pre-processing (Yang et al., 2019). Varying kernel sizes seemed to not affect the testing classification accuracy as it was inferred that most spectra data were distinct to be captured by convolutional filters and easily weighed on the predicting fully-connected layers to produce excellent classification results. Figure 4.18 showed the accuracy metric over epochs plot. From there, the model took only around 5 epochs to reach convergence, signifying the ease of training of DL models provided the definition of model architecture was performed properly.

Table 4.14: Results of DL CNN models for different kernel sizes of each layer on geographical origin classification

Pre-processing	Raw	SG	SG-SNV	SG-1st	SG-2nd
Model Type	**Accuracy Testing (%)**				
1	97.87	99.66	99.28	96.51	94.76
2	99.96	97.96	99.74	99.53	99.96
3	99.91	99.15	99.74	100.00	98.68
4	99.62	100.00	99.53	99.23	97.96
5	97.62	99.91	97.53	99.36	85.65
6	98.30	99.11	97.62	100.00	99.74
7	100.00	100.00	99.15	97.66	98.51
8	99.87	100.00	99.83	97.70	97.66
9	99.11	100.00	98.42	95.79	85.57
10	99.96	99.91	98.55	96.42	95.02
11	99.96	100.00	99.02	99.96	95.32
12	100.00	99.83	98.60	99.87	98.68
13	98.85	99.15	99.66	99.83	85.06
14	100.00	100.00	99.57	99.96	98.94
15	100.00	99.96	99.19	99.32	93.44
16	99.87	100.00	98.55	98.64	98.51

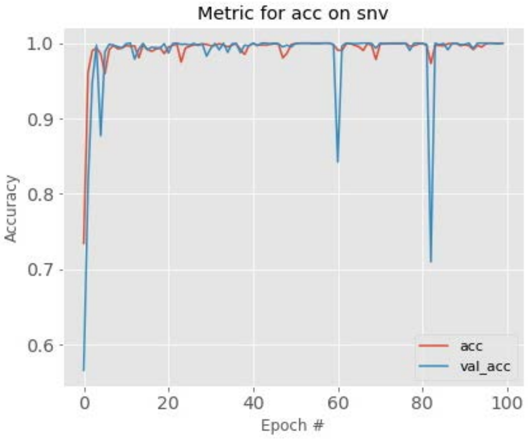


Figure 4.18: Accuracy metric over epochs for DL-CNN model (Model Type 2)

4.3.3.2 Stacked Autoencoder

Table 4.15 showed the results of DL SAE models on all raw and pre-processed mean spectra data for all model types. The results were similar to CNN models where all SAE models were able to yield 100% classification accuracy, even with varying number of nodes in the feature extractor layers. Figure 4.19 showed the accuracy metric over epochs plot. From there, the model took around 2 - 3 epochs to reach convergence, but the validation accuracy only stabilised after around 38 epochs due to updating of weights in the SAE model, which was explained in Section 4.2.3.2.

Table 4.15: Results of DL SAE models for different number of nodes of each layer on geographical origin classification

Pre-processing	Raw	SG	SG-SNV	SG-1st	SG-2nd
Model Type	Accuracy Testing (%)				
1	99.91	99.96	99.83	99.70	89.19
2	99.96	99.96	99.83	99.83	99.96
3	99.96	99.96	100.00	99.87	97.02
4	100.00	99.96	99.91	99.91	100.00
5	99.87	99.87	99.96	100.00	99.83

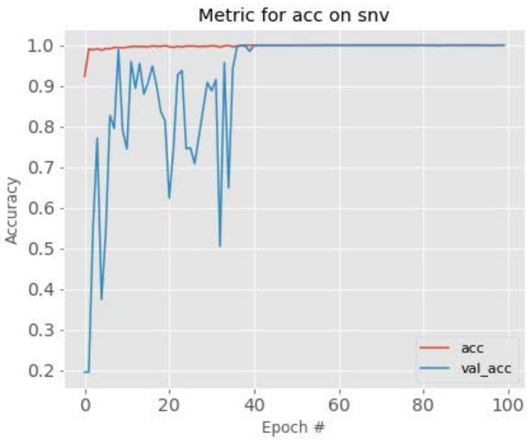


Figure 4.19: Accuracy metric over epochs for DL-SAE model (Model Type 5)

4.4 Summary

To determine the authenticity or degree of adulteration and geographical origin of Sarawak black pepper powder samples, HSI was used to gather the data of the samples by capturing the images of them, and multivariate data analysis was applied for data exploration, analysis and modelling purposes. During data exploration, mean spectra of these hyperspectral images were obtained, pre-processed (SG, SG-SNV, SG-1st, SG-2nd) and analysed, followed by PCA to screen for possible discrimination of clusters or targets. Sibu black pepper powder samples exhibited unique features compared to those of other regions. From PCA score plots, it was fairly possible to differentiate the difference in degree of adulteration and geographical origin of these samples. Next, models which included PLS, SVM, SVR, CNN and SAE were built and trained using raw and pre-processed mean spectra data to determine the authenticity and geographical origin of Sarawak black pepper powder samples. In assessing the authenticity of black pepper, DL SAE model on SG-SNV pre-processed mean spectra data had the best predictive performance on testing data set with R^2, RMSE and MAPE of 0.9010, 0.0143 and 1.17% respectively. While for the classification of geographical origin of Sarawak black pepper, SVM and DL models managed to achieve perfect classification accuracy on testing data set. As the controlled data set had considerable success in the determination of authenticity and classification of geographical origins of Sarawak black pepper powder samples, more external samples from the market would be acquired to serve as testing data sets and further assessed on the effectiveness of all the predictive models in the future study.

It was well noted that most research in the literature on detecting the authenticity and adulteration of black pepper was mainly focused on the further NIR wavelength range of more than 1000 nm. Considering the fact that Visible-NIR HSI mean spectra data which were focused on the visible light and close NIR regions were used, and most machine learning models were able to deliver high predictive performance, the main research objective in utilising the HSI with multivariate data analysis to effectively determine the authenticity and geographical origin of Sarawak black pepper powder samples was essentially achieved.

CHAPTER 5
PREDICTION OF CHEMICAL AND BIOLOGICAL ANALYTICAL PROPERTIES

Often, authenticity of black pepper powder products is tied with the internal quality of these products. Chemical and biological analyses were mainly performed to examine the chemical composition of contents of samples and micro-organism contents to determine the degree of contamination of the samples. The chemical composition of black pepper products was affected by factors such as growing climate, temperature, rainfall, soil conditions and quality, type of fertiliser used (Mercer et al., 2019; Sun et al., 2021). For example, carbohydrate, moisture, protein, volatile oil compounds contents. Analysis of the chemical compositions of essential or volatile oils present in black pepper products to assess the quality or geographical origin of black pepper had been done in the past (Li et al., 2020; Mercer et al., 2019).

HSI had been explored in the literature on determining the quality or chemical composition of various food and agricultural products (Khan et al., 2020; Liu et al., 2017). In order to study how the aforementioned factors affecting the final quality of the black pepper products, one could study the spectral differences present in the HSI data. In this research, all the Sarawak black pepper powder samples were sent for third party lab analyses. The lab analytical results were then used as targets to perform supervised training on various models to predict chemical and biological analytical properties. As outlined in Section 3.1, the range of purity of black pepper powder samples to be covered was from 90% to 100%, or 0% to 10% papaya seeds powder adulteration with the interval of 1%.

5.1 Preliminary Data Exploratory

The main contents of the lab analytical results were categorised into chemical and microbiological analyses. The chemical analysis included moisture content, ash content, acid insoluble ash content, non-volatile ether extract content, volatile oil content, lead and arsenic contents. The arsenic contents were found to be negligible in all black pepper powder samples, while the rest of the properties were found to be within the specified range of values set by IPC (IPC, 2015). Most of these methods were performed according to AOAC and ISO standards. While for the microbiological analysis, it included total plate, yeast, mould and *E. coli* counts as well as *salmonella spp*. Only plate and yeast count had numerical results while the others returned negligible outputs. Microbiological analyses were based on Australian Standard. All those properties could be referred to Table 5.1. From that table, ranges of values for those properties was included and they were compared with the standard ranges of values from MPB, which was one of the main parties for the laboratory quality assessment of black pepper products in Malaysia. Majority of those ranges of values were consistent with the standard ranges of values measured by MPB.

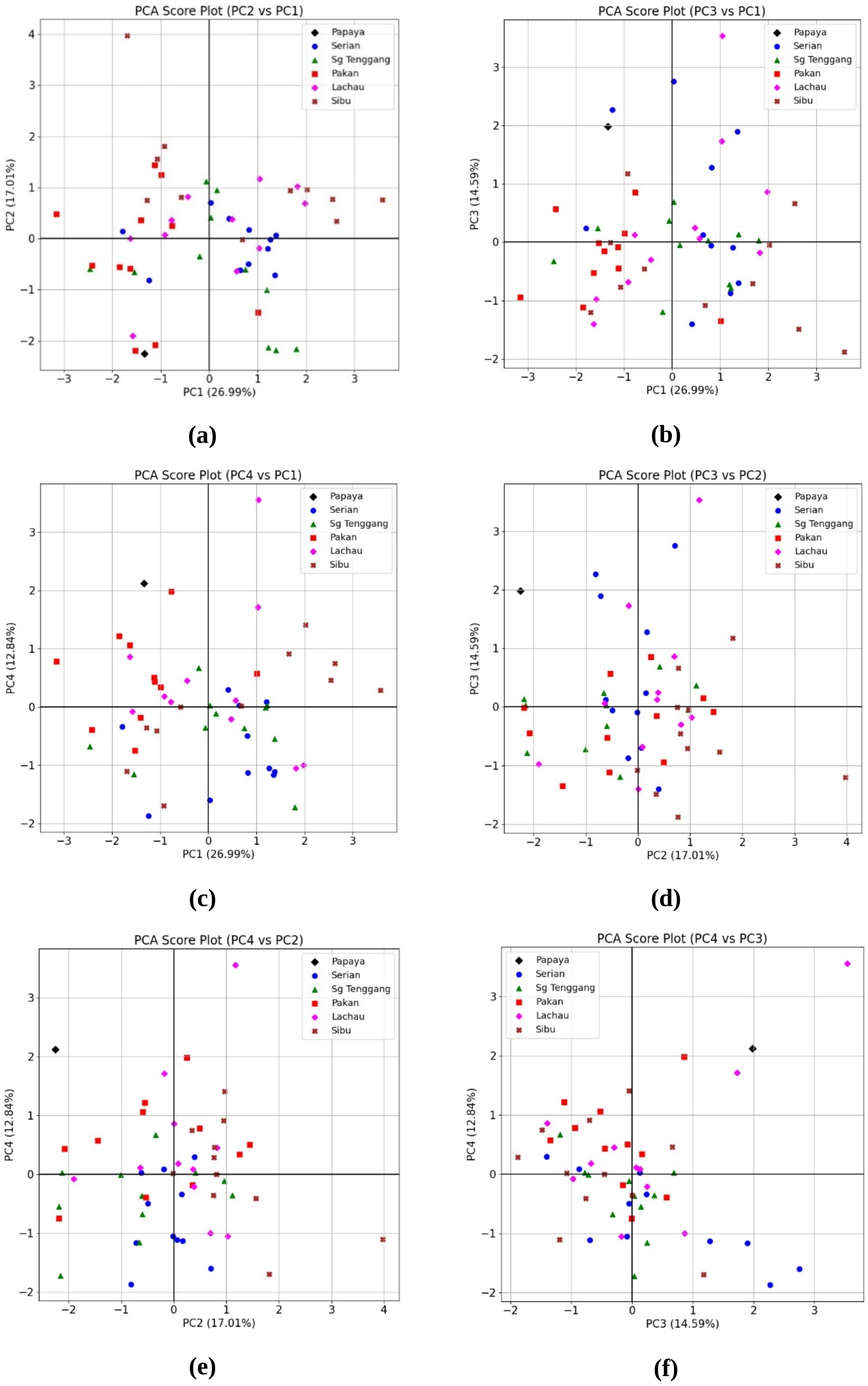
Table 5.1: Chemical and biological (microbiological) analytical properties to be used in subsequent modelling tasks and their respective range of values

Type of Properties	Parameter	Properties	Range from Lab	Standard Range from MPB
Chemical	1	Moisture (%)	6 – 14%	≤ 12%
	2	Ash Content (%)	0.3 – 5%	≤ 5%
	3	Acid Insoluble Ash (%)	0.1 – 1%	≤ 0.5%
	4	Non-volatile Ether Extract (%)	6 – 10%	≤ 6%
	5	Volatile Oil (%)	0.4 – 3%	≤ 2%
	6	Lead (mg/kg)	0.5 – 2.0 mg/kg	N/A
Biological	7	Total plate count (CFU/g)	10^3 – 10^5 CFU/g	≤ 10000 CFU/g
	8	Total yeast count (CFU/g)	10^3 – 10^5 CFU/g	≤ 1000 CFU/g

In particular, the volatile oil content was of main interest because of its contribution in the aroma and flavour of black pepper (Ravindran and Kallupurackal, 2001). The volatile oil contents of black pepper powder samples had been found to be random and have no definite relationship with the degree of adulteration. While for the geographical origins of Sarawak black pepper powder samples, the average volatile oil contents were as follows: Serian = 1.2%, Sungai Tenggang = 1.4%, Pakan = 1.0%, Lachau = 1.4%, Sibu = 1.6%. It was reported the volatile oil of black pepper was around 2.8%, which was higher than most average values of those black pepper powder samples under examination (Johny et al., 2020). The difference in the volatile oil contents could be attributed to different processing of the black pepper powder products and various environmental factors affecting the cultivation of these black pepper products (Johny et al., 2020; Ravindran and Kallupurackal, 2001).

PCA was performed on the chemical and microbiological analyses data. The data was pre-processed through scaling with SNV due to disparity of values present among all variables. The explained variability distribution for first 6 principal components (PC) was as follows:

- PC1 = 26.99%
- PC2 = 17.01%
- PC3 = 14.59%
- PC4 = 12.84%
- PC5 = 11.35%
- PC6 = 10.91%

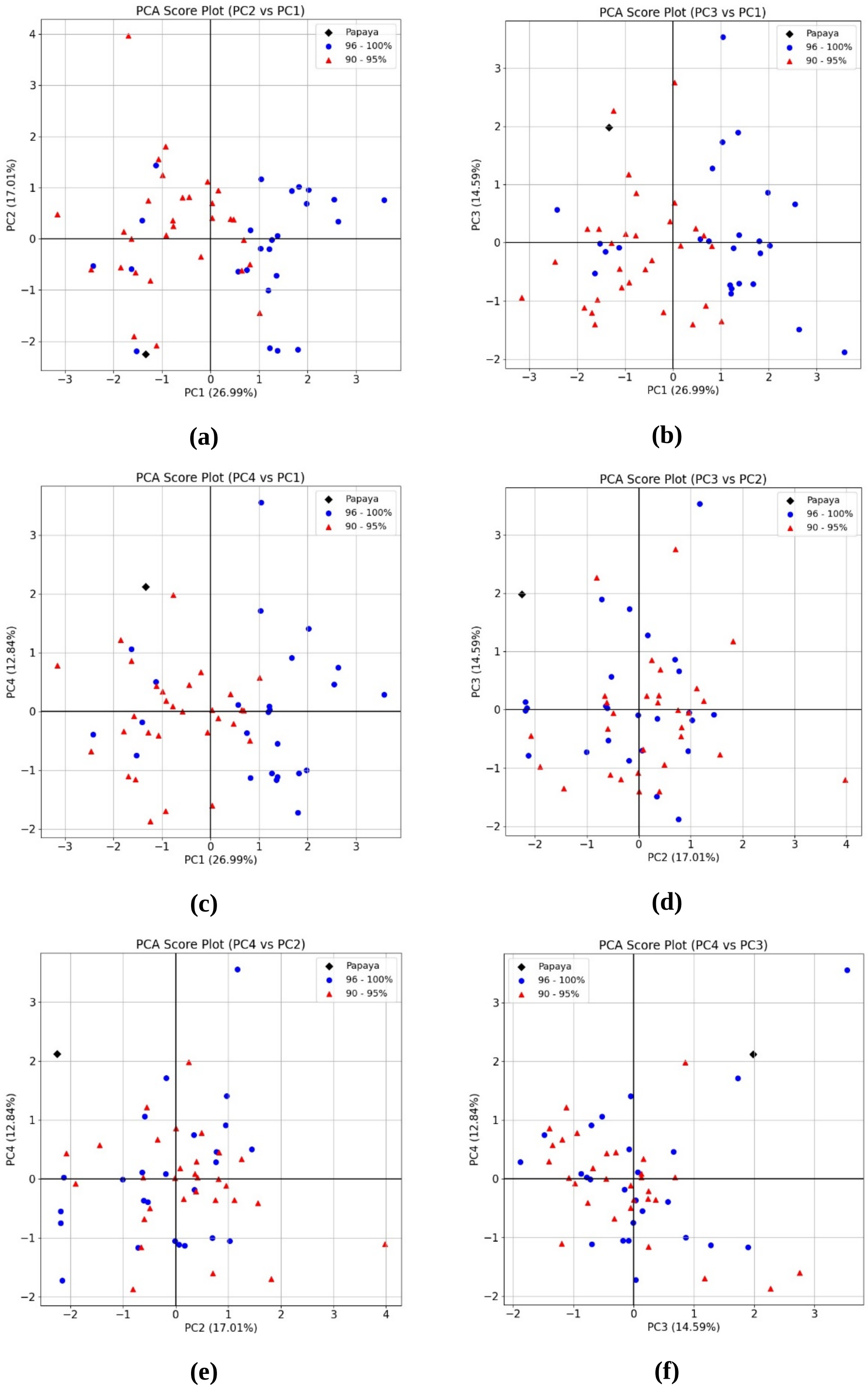


(a) **(b)**

(c) **(d)**

(e) **(f)**

Figure 5.1: PCA score plots for first 4 principal components on the chemical and microbiological analytical data, with the labels being the geographical origins of black pepper powder samples: (a) PC2 vs PC1, (b) PC3 vs PC1, (c) PC4 vs PC1, (d) PC3 vs PC2, (e) PC4 vs PC2, (f) PC4 vs PC3



(a) **(b)**

(c) **(d)**

(e) **(f)**

Figure 5.2: PCA score plots for first 4 principal components on the chemical and microbiological analytical data, with the labels being the range of purity or authenticity of black pepper powder samples: (a) PC2 vs PC1, (b) PC3 vs PC1, (c) PC4 vs PC1, (d) PC3 vs PC2, (e) PC4 vs PC2, (f) PC4 vs PC3

It took 6 PCs to reach 93.7% explained variability, which indicated that almost all the properties or variables were crucial in this data. Figure 5.1 and Figure 5.2 showed the PCA score plots for first 4 PCs with different labels of authenticity and geographical origin of black pepper powder samples. Papaya seeds powder was included in the score plots as a reference of adulterant of black pepper. From there, in terms of geographical origins, it was difficult to screen for any distinct cluster or clear boundary in differentiating geographical origins of black pepper. This was peculiar as it was demonstrated in the literature that different origins of black pepper had different chemical compositions and bioactivities (Li et al., 2020; Liang et al., 2021). However, in terms of authenticity of black pepper, PC2 vs PC1 seemed to have slightly clear distinction of clusters where higher purity of black pepper was more on the right side of plot while lower purity of black pepper was more towards left side of plot. Similar situation was observed in PC3 vs PC1 score plot, but not for other remaining comparisons of scores.

5.2 Model Development and Training

After the data exploration, Visible-NIR HSI mean spectra data were pre-processed and used along with the lab analytical results as targets to train the models to predict the chemical and biological (or microbiological) analytical properties. The details of all the properties could be referred to Table 5.1. As prediction of these properties was a regression task, PLS, SVR and DL (CNN and SAE) models were used and assessed on their predictive performance.

5.2.1 Partial Least Square

The PLS models were fed with both raw and pre-processed (SG, SG-SNV, SG-1st, SG-2nd) mean spectra data. Due to the difference in magnitudes of values in the lab results, models for chemical and biological analytical properties were trained separately. To obtain the best model, optimisation was made over a range of 50 latent variables and based on the RMSE from the trained model tested on the cross validation. The testing prediction results were tabulated in Table 5.2, along with the training and validation results as shown in Table C.1 of Appendix C. Parity plots for PLS models on chemical and biological properties were plotted. For example, parity plots on moisture content and total plate count could be seen in Figure 5.3. The rest of the parity plots could be referred to Figure C.1 in Appendix C. Additionally, plots of optimisation of latent variables, and regression coefficients for moisture content and total plate count were plotted in Figure 5.3 as well. The plots of regression coefficients for the remaining analytical properties could be referred to Figure C.2 in Appendix C.

Table 5.2: Results of PLS models on regression of prediction of chemical and microbiological properties with testing data set

Pre-processing		Raw	SG	SG-SNV	SG-1st	SG-2nd
Best # Latent Variable	**Chemical**	**40**	**40**	**40**	**49**	**48**
	Microbiological	**27**	**32**	**38**	**44**	**47**

Indicator	Pre-processing	Chemical						Microbiological	
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
R² Testing	**Raw**	0.6569	0.6149	0.5459	**0.2707**	**0.6387**	**0.4202**	0.3093	0.1954
	SG	0.6560	0.6150	0.5482	0.2602	0.6361	0.4093	0.3023	**0.1990**
	SG-SNV	0.6578	**0.6153**	**0.5508**	0.2479	0.6332	0.4016	0.3052	0.1644
	SG-1st	**0.6586**	0.6123	0.5394	0.2594	0.6244	0.4150	**0.3106**	0.1730
	SG-2nd	0.6480	0.6096	0.5263	0.2382	0.6148	0.4068	0.3041	0.1615
RMSE Testing	**Raw**	0.9594	0.7714	0.1286	0.4589	0.2615	0.2435	104930.95	22939.09
	SG	0.9607	0.7712	0.1283	0.4622	0.2624	0.2458	105461.75	22886.91
	SG-SNV	0.9582	0.7710	0.1279	0.4660	0.2635	0.2473	105244.34	23376.22
	SG-1st	0.9570	0.7740	0.1295	0.4624	0.2666	0.2446	104834.12	23255.12
	SG-2nd	0.9718	0.7766	0.1314	0.4690	0.2700	0.2463	105331.59	23417.33
MAPE Testing (%)	**Raw**	8.84	32.04	27.81	4.03	18.89	16.56	142.20	105.87
	SG	8.93	32.92	27.87	4.05	18.93	16.30	139.77	108.82
	SG-SNV	8.86	32.98	27.52	4.10	18.85	16.58	147.39	116.18
	SG-1st	8.90	32.65	27.85	4.03	19.25	16.38	149.26	110.71
	SG-2nd	9.08	32.30	28.34	4.14	19.66	16.55	142.01	109.68

In the case of prediction of chemical analytical properties, most PLS models for all data pre-processing methods required at least 40 latent variables to get the best testing predictive outputs. From Table 5.2, the best R^2 result came from PLS model on predicting moisture content, which was 0.6586, while the worse one was on non-volatile ether extract, which was 0.2707. Only prediction of moisture content and non-volatile ether extract had MAPE of less than 10%. All data pre-processing methods had similar testing predictive performance by the PLS models. In terms of regression coefficients as shown in Figure 5.3, it could be observed that peaks around 590 and 615 nm were noticeably more prominent.

While for the case of prediction of biological analytical properties, the PLS models had worse predictive performance, where all testing R^2 values were lower than 0.5. The randomness of values of biological properties was the main factor the PLS models were unable to yield good testing predictive performance, causing extremely high RMSE to be optimised. From the parity plots in Figure 5.3, it was observed that some predictions were too different than the intended measured values. The predictions were subpar due to non-linearity of the spectral data, randomness present in the lab analytical results and higher order microorganism populations. While for the regression coefficients, the coefficients were larger in magnitude due to high values of the biological analytical properties, and around 550 nm, a prominent peak was found while on other range, the peaks in the spectra were spread out quite evenly.

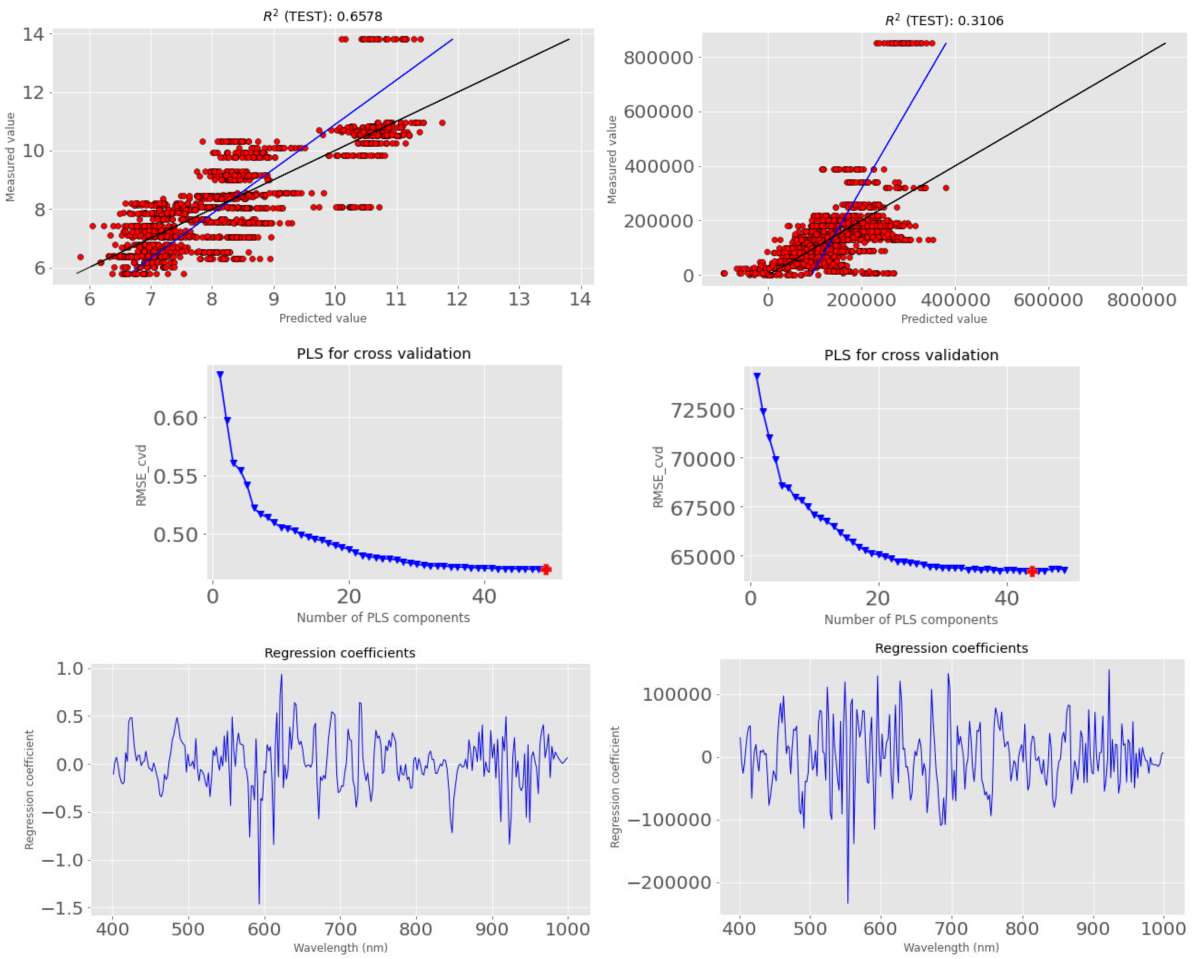


Figure 5.3: [Above] – Parity plots of PLS models, blue line represents the best fit line from the model; [Middle] – Finding the optimal number of PLS latent variables from cross validation RMSE; [Below] – Regression coefficients of PLS models; [Left side of all plots] – moisture content; [Right side of all plots] – total plate count

5.2.2 Support Vector Regression

Non-linear models were necessary to model for better prediction of chemical and biological analytical properties. SVR models were built and trained using all raw and pre-processed mean spectra data. Due to difference in magnitude of values in the lab results, separate SVR models were trained for prediction of chemical and biological analytical properties. The same parameters (C and γ) to be optimised were following the same range as outlined in Sections 3.5.2 and 4.2.2. The testing prediction results were then outlined in Table 5.3, with the training and validation results tabulated in Table C.2 of Appendix C.

Table 5.3: Results of SVR models on regression of prediction of chemical and microbiological properties with testing data set

Indicator	Pre-processing	Chemical						Microbiological	
		1	2	3	4	5	6	7	8
Best C, γ	**Raw**	100, 10	10, 100	100, 10	100, 10	100, 10	100, 10	100, 10	100, 100

	SG	100, 10	10, 100	100, 10	100, 10	100, 10	100, 10	100, 10	100, 100
	SG-SNV	100, 1	100, 1	100, 1	100, 1	100, 1	100, 1	100, 1	100, 1
	SG-1st	100, 10^5	100, 10^5	100, 10^4	100, 10^5	100, 10^5	10, 10^5	100, 10^4	100, 10^4
	SG-2nd	100, 10^5	100, 10^5	100, 10^5	100, 10^5	100, 10^5	100, 10^5	100, 10^5	100, 10^5
R^2 Testing	**Raw**	0.7032	0.6973	0.7414	0.2303	0.7253	0.4954	0.0277	0.0200
	SG	0.7020	0.6959	0.7390	0.2212	0.7177	0.4887	0.0277	0.0200
	SG-SNV	0.7105	0.7172	0.7793	0.3793	0.7450	0.5416	0.0094	0.0206
	SG-1st	**0.7934**	**0.7853**	**0.8184**	**0.5939**	**0.8782**	**0.6837**	**0.0302**	**0.0330**
	SG-2nd	0.7190	0.6656	0.7028	0.4423	0.7394	0.5739	0.0153	0.0317
RMSE Testing	**Raw**	0.8924	0.6839	0.0971	0.4715	0.2280	0.2271	124499.12	25315.91
	SG	0.8941	0.6855	0.0975	0.4742	0.2311	0.2287	124499.80	25315.69
	SG-SNV	0.8812	0.6610	0.0897	0.4234	0.2197	0.2165	125668.08	25308.12
	SG-1st	0.7446	0.5760	0.0813	0.3424	0.1519	0.1798	124341.81	25147.90
	SG-2nd	0.8682	0.7188	0.1041	0.4013	0.2221	0.2087	125293.46	25164.45
MAPE Testing (%)	**Raw**	7.23	28.41	17.08	3.35	13.98	10.73	225.85	144.62
	SG	7.26	28.47	17.11	3.38	14.18	10.81	225.84	144.62
	SG-SNV	6.90	29.33	15.68	3.03	12.53	10.68	243.47	145.43
	SG-1st	6.14	19.92	15.57	2.69	9.57	10.52	231.45	140.67
	SG-2nd	7.08	32.41	20.06	3.08	14.28	11.21	234.14	138.36

From Table 5.3, all SVR models with SG-1st data pre-processing method had the best testing predictive performance. During the prediction of chemical analytical properties, all SVR models had drastically better testing predictive performance than PLS models. The best testing R^2 result was achieved by the prediction of volatile oil content which was 0.8782, while the worst property was non-volatile ether extract, which had R^2 of only 0.5939. Examples of parity plots of SVR models with the best pre-processing method were plotted as shown in Figure 5.4. The moisture content that had the highest measured values were considered outliers, thus the SVR model did not consider these samples, which yielded highly scattered outputs. For prediction of microbiological properties, the model was unable to fit the data, hence the prediction of constant values. The rest of the parity plots for other properties could be referred to Figure C.3 in Appendix C.

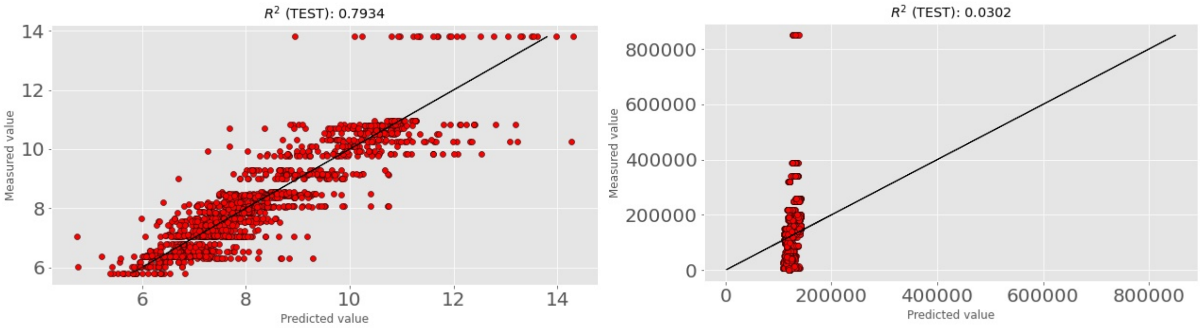


Figure 5.4: Parity plots of SVR models for (Left) moisture content and (Right) total plate count

There were many probable unique spectral features present in the visible light region of Visible-NIR HSI SG-1st pre-processed mean spectra, thus allowing SVR models to differentiate and identify them, although the relationship between these spectral features in visible light region and SVR was rather unclear, even in the literature (Sun et al., 2021). However, in the prediction of biological analytical properties, the testing predictive performance of SVR models were even worse than PLS models. Due to the randomness of the lab results and higher order microorganism populations, the SVR models was unable to decently fit the spectral data into the decision function after exhausting all the inputs to be support vectors. This prompted for more data on lab analyses for higher predictive performance by the models. Next, prediction of moisture content, non-volatile ether extract and volatile oil content yielded MAPE of less than 10%, explaining the good capability of SVR model in forecasting the chemical analytical properties.

5.2.3 Deep Learning

It was of interest to investigate whether deep learning is capable of predicting both chemical and biological analytical properties effectively or not. CNN and SAE model architectures were used, and most of the settings, hyperparameters, configurations and assessment of different model types were retained with the following exception: The predictor section was instead made up of two diverging separate fully-connected layers where each layer connected to individual chemical (6 nodes) and biological (2 nodes) properties output nodes. Both would have two layers of 128 neurons to be optimised.

5.2.3.1 Convolution Neural Network

The list of CNN model types to be assessed could be referred to Table 4.4. The testing prediction results were presented in Table 5.4. Most of the best models were obtained based on SG-1st and SG-2nd pre-processed mean spectra data. The training and validation results could be referred to Table C.3 in Appendix C. From Table 5.4, for prediction of chemical analytical properties, the best testing R^2 was achieved by moisture content of 0.7474 while the smallest best R^2 was the non-volatile ether extract with only 0.1767. While for the prediction of biological analytical properties, there was a major improvement on the predictive performance as indicated by testing R^2 on total plate and yeast count prediction, which were 0.5240 and 0.2280 respectively. It was most likely that the derivatives from SG-1st and SG-2nd pre-processing methods contained numerous representative features which could match with the nature of lab analytical results. In terms of MAPE, only moisture content and non-volatile ether extract properties managed to achieve error of less than 10%. Parity plots from CNN model with the best respective model types in Figure 5.5 were presented for the moisture content and total plate count, with the rest of properties could be referred to Figure C.4 in Appendix C. At the same time, loss over epochs plot was displayed and the exorbitantly high loss was due to high RMSE from the biological analytical properties prediction. Alterations on depth and complexity of model architecture or improvement on quality of data might be required in the future so that these results could be fit into the model.

Table 5.4: Results of DL CNN models on regression of prediction of chemical and microbiological properties with testing data set, bolded results represent the best results

Indicator	Pre-processing	Chemical						Microbiological	
		1	2	3	4	5	6	7	8
Best Model Type	Raw	14	7	7	7	14	16	16	14
	SG	4	4	1	8	8	3	8	7
	SG-SNV	8	4	8	3	4	8	8	16
	SG-1st	4	4	16	14	14	3	14	14
	SG-2nd	1	15	1	1	15	15	8	6
R^2 Testing	Raw	0.6725	0.6892	0.5002	-0.0086	0.6103	0.3537	0.4132	**0.2280**
	SG	0.7224	0.6970	0.5833	0.1519	0.6723	0.4544	0.4710	0.1695
	SG-SNV	0.6996	0.6964	0.6199	-0.0195	0.6745	0.4274	0.4479	0.1285
	SG-1st	**0.7474**	0.7255	0.5867	0.1327	0.6771	**0.4595**	0.4587	0.1969
	SG-2nd	0.7346	**0.7371**	**0.6529**	**0.1767**	**0.6881**	0.4485	**0.5240**	0.1504
RMSE Testing	Raw	0.9427	0.6909	0.1349	0.5273	0.2807	0.2503	98449.60	22225.82
	SG	0.8680	0.6821	0.1231	0.4835	0.2574	0.2299	93482.07	23052.54
	SG-SNV	0.9029	0.6829	0.1176	0.5301	0.2565	0.2356	95501.95	23615.00
	SG-1st	0.8278	0.6493	0.1226	0.4889	0.2556	0.2289	94561.84	22669.38
	SG-2nd	0.8486	0.6354	0.1124	0.4764	0.2511	0.2312	88673.83	23316.02
MAPE Testing (%)	Raw	8.74	27.12	26.80	4.94	18.64	17.61	135.15	115.78
	SG	8.13	24.76	22.29	3.95	15.86	15.52	131.18	130.22
	SG-SNV	8.01	25.59	22.08	4.34	15.60	16.33	109.44	133.02
	SG-1st	7.45	21.79	23.90	4.16	16.66	15.48	110.01	112.79
	SG-2nd	7.51	22.36	20.33	4.12	16.54	17.43	102.78	124.81

Table 5.5: Example of DL CNN model architecture (Model Type 15) for prediction of lab analytical properties

Layer	Output Shape	# of Parameters	Connected To
Input Layer	(280, 1)	0	-
Conv1D_1 *[kernel size = 3]*	(280, 16)	64	Input Layer
BatchNormalization_1	(280, 16)	64	Conv1D_1
MaxPool1D_1	(140, 16)	0	BatchNormalization_1
Conv1D_2 *[kernel size = 3]*	(140, 32)	1568	MaxPool1D_1
BatchNormalization_2	(140, 32)	128	Conv1D_2
MaxPool1D_2	(70, 32)	0	BatchNormalization_2
Conv1D_3 *[kernel size = 3]*	(70, 64)	6208	MaxPool1D_2
BatchNormalization_3	(70, 64)	256	Conv1D_3
MaxPool1D_3	(35, 64)	0	BatchNormalization_3
Conv1D_4 *[kernel size = 1]*	(35, 128)	8320	MaxPool1D_3
BatchNormalization_4	(35, 128)	512	Conv1D_4
MaxPool1D_4	(17, 128)	0	BatchNormalization_4

Flatten	(2176)	0	MaxPool1D_4
Dense_1	(128)	278656	Flatten
Dense_2	(128)	16512	Dense_1
Dense_3	(128)	278656	Flatten
Dense_4	(128)	16512	Dense_3
Dense_5 *[chemical output]*	(6)	774	Dense_2
Dense_6 *[biological output]*	(2)	258	Dense_4
Total of trainable number of parameters = **608,008**			

Model summary of one of good performing CNN models (Model Type 15 of layer configuration [3-3-3-1]) was displayed in Table 5.5. This, in turn, required high number of parameters (i.e. 600,000+) to be trained, due to two fully-connected sections to be optimised. Usually, layers of kernels with larger size were placed on the first few layers followed by layers of kernels with smaller size, so that the feature extraction efficiency could be maximised

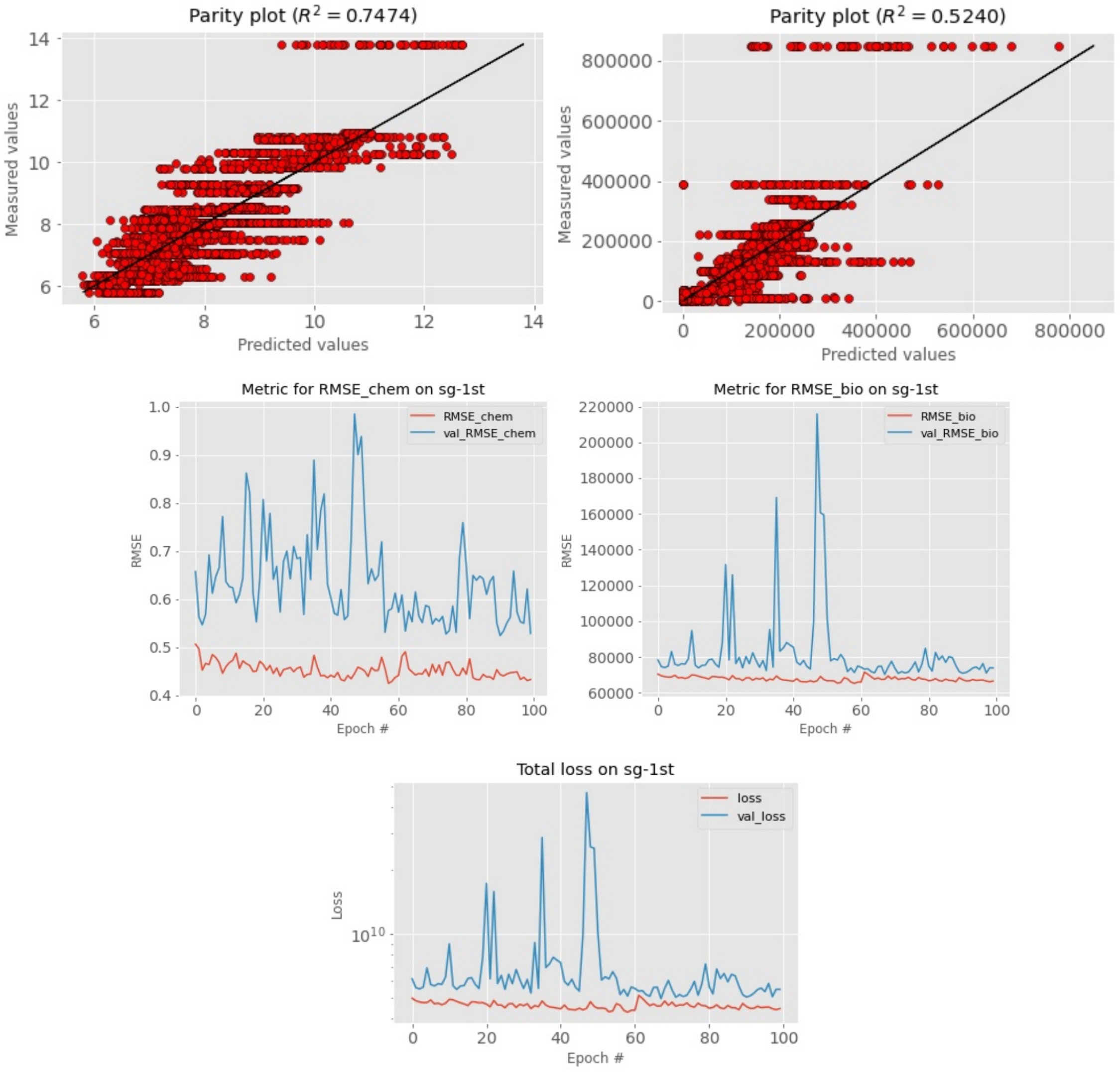


Figure 5.5: Above – Parity plots of DL CNN models of the best model types for (Left) moisture content [Best Model Type = 4] and (Right) total plate count [Best Model Type = 8]; Middle – RMSE metric over epochs plot of CNN model using SG-1st pre-processed mean spectra data; Below – Loss over epochs plot

(Goodfellow et al., 2016; Nagasubramanian et al., 2018; Saha and Manickavasagan, 2021; Yang et al., 2019). For example, from Table 5.4, Model Type 15 depicted that configuration and had the best predicted R^2 on ash and volatile oil contents, and Model Type 1 (layer configuration [1-1-1-1]) with the layers of smallest kernels had the best predicted R^2 on the acid insoluble ash and non-volatile ether extract. Model Types 3 ([1-1-3-1]) and 4 ([1-1-3-3]) had different configurations but still had better predictive performance. Despite that, all CNN models had decent predictive performance. They had worse testing predictive performance than SVR models in the prediction of chemical analytical properties but better in the prediction of biological analytical properties.

5.2.3.2 Stacked Autoencoder

The assessment on the different model types of the SAE models were referred to Table 4.7. The testing prediction results were tabulated in Table 5.6. The training and validation results could be referred to Table C.4 in Appendix C. Model summary of one of SAE models was displayed in Table 5.7. There were 130,000 trainable parameters in average, which was considerably faster to train than CNN models. Overall, the SAE model prediction results were the worst among all the models. From Table 5.6, for prediction of chemical analytical properties, the best testing R^2 was 0.6025 on the moisture content while the smallest best testing R^2 was 0.1077 on the non-volatile ether extract as well. On the other hand, only the total plate count was outputting reasonable predictions among the biological analytical properties. The parity plots of SAE models of best model types on the moisture content and total plate count were displayed in Figure 5.6, while the parity plots for the rest of other properties were in Figure C.5 of Appendix C. Plots of RMSE metric and losses over epochs on SAE model (SG-SNV pre-processed) were illustrated in Figure 5.6.

Table 5.6: Results of DL SAE models on regression of prediction of chemical and microbiological properties with testing data set

Indicator	Pre-processing	Chemical						Microbiological	
		1	2	3	4	5	6	7	8
Best Model Type	Raw	5	4	4	4	4	4	4	1
	SG	5	4	1	5	5	5	5	5
	SG-SNV	2	5	5	2	2	2	1	2
	SG-1^{st}	5	5	5	5	5	5	5	1
	SG-2^{nd}	4	2	5	3	5	4	1	5
R^2 Testing	Raw	0.5588	0.5216	0.4196	-0.3750	0.4509	0.2447	0.3911	-0.4721
	SG	-59.64	-4.1144	-7.1452	-1268.99	-4.8877	-4.1645	-10.031	-27.179
	SG-SNV	**0.6025**	0.5207	**0.4772**	**0.1077**	0.5098	**0.3346**	0.4069	-0.4448
	SG-1^{st}	0.5878	**0.5507**	0.4527	0.0065	**0.5538**	0.3229	**0.4247**	-0.0613
	SG-2^{nd}	-0.4869	-0.1676	0.0414	-0.0145	-0.2804	-0.0007	0.2568	**-0.0600**
RMSE Testing	Raw	1.088	0.8597	0.1454	0.6301	0.3224	0.2779	98523.46	31027.84
	SG	12.755	2.8111	0.5447	19.150	1.0556	0.7267	419346.21	135749.58

	SG-SNV	1.0327	0.8606	0.1380	0.5076	0.3046	0.2608	97235.04	30738.56
	SG-1st	1.0517	0.8332	0.1412	0.5356	0.2906	0.2631	95771.05	26345.42
	SG-2nd	1.9973	1.3432	0.1869	0.5412	0.4923	0.3199	108850.96	26328.81
MAPE Testing (%)	**Raw**	10.88	35.93	27.30	5.80	22.57	21.18	140.70	133.80
	SG	158.41	91.24	100.00	219.64	61.72	55.71	1433.61	539.06
	SG-SNV	9.61	35.08	30.53	4.28	21.53	17.10	174.48	132.69
	SG-1st	9.59	33.20	27.52	4.90	18.75	16.00	138.13	160.24
	SG-2nd	23.22	56.85	45.98	4.99	39.96	23.09	150.72	143.25

Table 5.7: Example of DL SAE model architecture (Model Type 5) for prediction of lab analytical properties

Layer	**Output Shape**	**# of Parameters**	**Connected To**
Input Layer	(280)	0	-
Dense_1	(192)	53952	Input Layer
BatchNormalization_1	(192)	768	Dense_1
Dense_2	(128)	24704	BatchNormalization_1
BatchNormalization_2	(128)	512	Dense_2
Dense_3	(96)	12384	BatchNormalization_2
BatchNormalization_3	(96)	384	Dense_3
Dense_4	(64)	6208	BatchNormalization_3
BatchNormalization_4	(64)	256	Dense_4
Dense_5	(32)	2080	BatchNormalization_4
BatchNormalization_5	(32)	128	Dense_5
Dense_6	(16)	528	BatchNormalization_5
BatchNormalization_6	(16)	64	Dense_6
Dense_7	(128)	2176	BatchNormalization_6
Dense_8	(128)	2176	Dense_7
Dense_9	(128)	16512	BatchNormalization_6
Dense_10	(128)	16512	Dense_9
Dense_11 *[chemical output]*	(6)	774	Dense_8
Dense_12 *[biological output]*	(2)	258	Dense_10
Total of trainable number of parameters = **139,320**			

Looking at the parity plots in Figure 5.6, there were numerous scattered predictions that seemed like outliers, especially for the prediction of total plate count. In the prediction of total yeast count, the model was unable to fit the spectral data to explain the relationship. It was inferred that randomness of values in the total yeast count results caused mapping functions in SAE models to fluctuate, which was shown in Figure 5.6, losses over epoch plot for prediction of biological analytical properties. Hence, this resulted in near constant values and negative or zero R^2. On the losses over epoch, the training loss was still reducing at the end of epochs while the validation loss remained unchanged. Yet, due to low or negative training predictive performance was found, increasing the number of epochs did not further help in boosting model predictive performance. Fundamentally, feature extractor section in

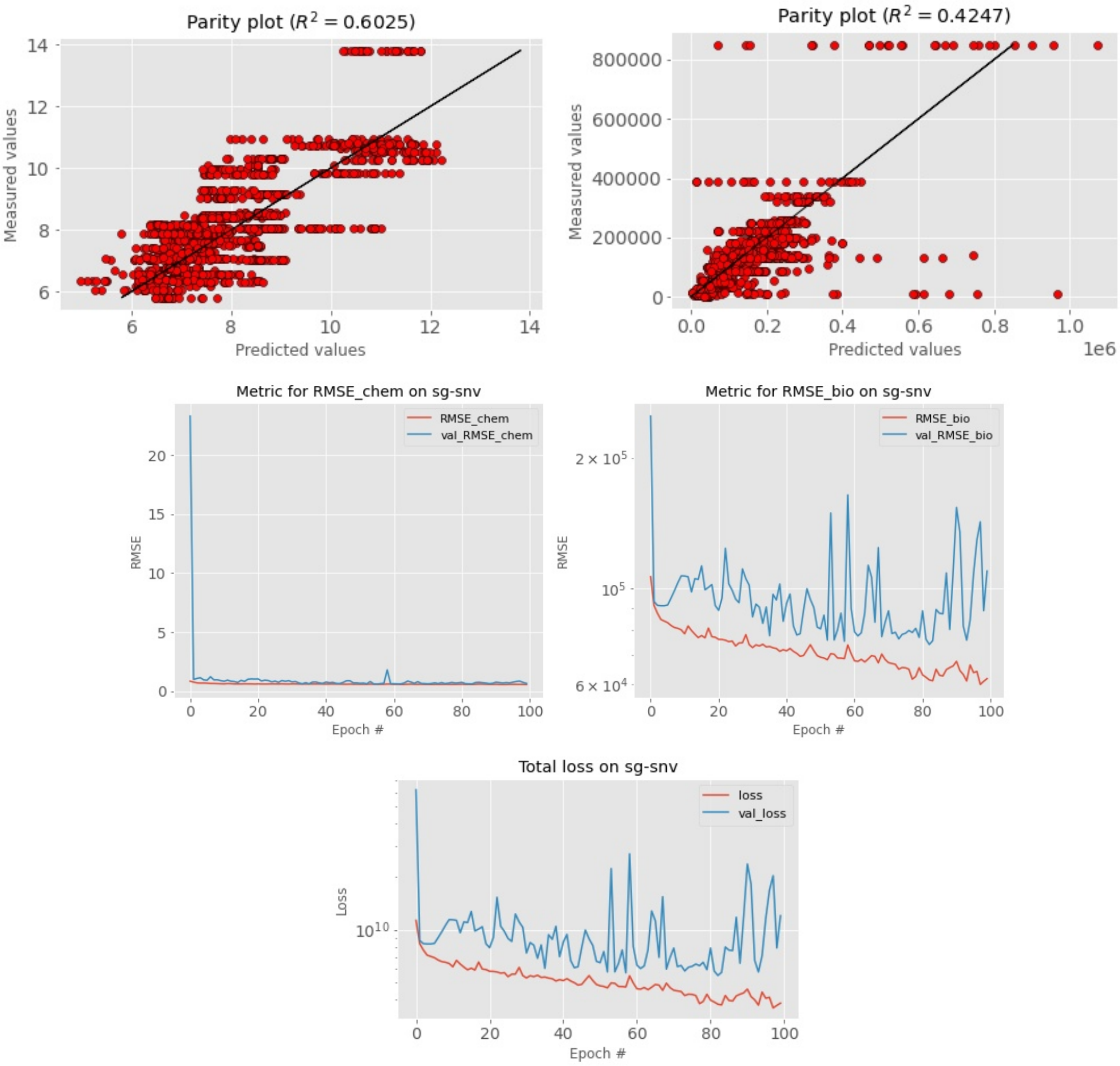


Figure 5.6: Above – Parity plots of DL SAE models of the best model types for (Left) moisture content [Best Model Type = 2] and (Right) total plate count [Best Model Type = 5]; Middle – RMSE metric over epochs plot of SAE model using SG-SNV pre-processed mean spectra data; Below – Loss over epochs plot

SAE models were made up of non-linear perceptrons. Updating convolution filters appeared to be more stable than updating the weights of perceptrons. Exploding gradients issue was more probable considering extremely large RMSE could cause updating of weights went haywire (Goodfellow et al., 2016; Paoletti et al., 2019; Shrestha and Mahmood, 2019).

5.3 Summary

As aforementioned, to assess the quality of black pepper powder products, detailed third-party lab analyses were inevitably required so that the quality of prediction of quality of black pepper powder products from HSI spectral analysis could be improved and more consistent. Yet, as these lab analyses were time-consuming, HSI along with initial lab analytical results were used to train predictive models to predict the quality or chemical composition of incoming black pepper powder samples. Biological, or specifically microbiological analytical results were also included in the prediction and then studied to investigate the extent and effectiveness of the predictive models.

Table 5.8: The best machine learning models with their respective data pre-processing methods on the prediction of chemical and microbiological analytical properties

Category	Properties	Best Model [Data Pre-processing Method]
Chemical	Moisture	SVR [SG-1st]
	Ash Content	SVR [SG-1st]
	Acid Insoluble Ash	SVR [SG-1st]
	Non-volatile Ether Extract	SVR [SG-1st]
	Volatile Oil	SVR [SG-1st]
	Lead	SVR [SG-1st]
Microbiological	Total plate count	CNN [SG-2nd]
	Total yeast count	CNN [Raw]

PLS, SVR, CNN and SAE models were trained and the results were tabulated in Table 5.8. From there, SVR model had the best predictions for all the chemical analytical properties, with the SG-1st pre-processing method on the mean spectra data. While for biological analytical properties, CNN model was the best. Moisture content and non-volatile ether extract were reliably predicted by the models. PLS model predictive performance was in between CNN and SVR models, while SAE model was the worst performing model. However, the randomness present in the lab analytical results which showed scattered and mixed clusters from PCA as well as complications from higher order microorganism populations, especially for biological analytical properties, had deteriorated the predictive performance of all models. Most importantly, the property of main constituent of black pepper which is piperine content was lacking from the lab analysis and might heavily improve the predictive performance of models as most research analysed piperine content to successfully verify the quality of black pepper powder products. Additionally, considering most chemical information resided outside Visible-NIR range (i.e. more than 1000 nm), Visible-NIR HSI data might be insufficient to reliably predict the chemical and biological analytical properties.

CHAPTER 6

CONCLUSION AND RECOMMENDATIONS

This research has demonstrated that the concept of using hyperspectral imaging (HSI) was proven as a rapid, non-destructive and affordable alternative to detect the authenticity or degree of adulteration in Sarawak black pepper powder samples and classify their geographical origins. By using Visible-NIR HSI which covered the visible light and near infrared region (400 – 1000 nm), hyperspectral images were taken followed by processing to mean spectra data sets. Data pre-processing which included SG, SG-SNV, SG-1st and SG-2nd were applied onto the data and compared. PCA revealed that from PC2 vs PC1 score plots, there was a slight clear and distinctive cluster on different degree of adulteration on all raw and pre-processed data. For different geographical origins, only SG-1st and SG-2nd pre-processed data revealed Sibu cluster separated from other clusters. In terms of explained variability for first 3 principal components, raw and SG pre-processing method explained the highest variability of 98.92% while SG-SNV explained 96.98%. On the contrary, SG-1st and SG-2nd pre-processing had less than 95% explained variability. This indicates that SG with derivative pre-processing required more components to fully explain the data, which could increase the complexity of the results.

Once the screening for different degrees of adulteration and geographical origin of black pepper powder samples was proven possible, various machine learning models (PLS, PLS-DA, SVR, SVM, DL CNN and DL SAE) were built and trained on all the raw and pre-processed mean spectra data. Among all the models, DL SAE model had the best testing predictive performance in the determination of authenticity of black pepper powder samples. The resulting best SAE model had 6 fully-connected layers on the feature extractor section, was trained on SG-SNV pre-processed data and had R^2, RMSE and MAPE of 0.9010, 0.0143 and 1.13% respectively. SVM, DL CNN and DL SAE models had generally high testing predictive performance in classification of geographical origin.

The utility of HSI data was tested further by assessing the prediction of chemical and biological analytical properties from various predictive models. While waiting for updated detailed but time-consuming lab analyses, the existing lab analytical results could be used to train the predictive models. Based on PCA, only PC2 vs PC1 score plots revealed minor discrimination on the clusters for different purity of black pepper powder samples. While for different geographical origins, PCA could not identify any direct or indirect separation of clusters. Explained variability could only go up to 93.7% using first 6 principal components. Next, PLS, SVR, DL CNN and DL SAE models were then trained with all the raw and pre-processed mean spectra data along with their respective results of lab analytical properties. SVR model had the best testing predictive performance among all on SG-1st pre-processed data in the prediction of chemical analytical properties, which had R^2 of up to 0.8782 for volatile oil content. It had the worst predictive performance in the prediction of biological analytical properties. On the other hand, DL CNN model performed the best in the prediction of biological analytical properties because of its capability to extract features which hugely related to the lab results, albeit with lower than normal R^2 of only 0.2280 on total yeast count.

Despite good predictive performance displayed by these models using Visible-NIR HSI mean spectra data, there was no clear and direct relationship to explain authenticity, geographical origin and quality of black pepper using Visible-NIR bands information. Visible-NIR HSI usually had higher success rate in most classification and determination of quality of non-powder opaque objects due to distinct dissimilarity. Visible-NIR HSI still lacked various chemical information such as overtones and bonding which resided beyond 1000 nm as most research reported more success in determination of authenticity and origin of black pepper (Hu et al., 2018; Lima et al., 2020; McGoverin et al., 2012; Orrillo et al., 2019; Wilde et al., 2019). Next, as it was costly to reproduce the lab analyses, these results were duplicated so that enough data could be fed to train the model for better generalizability. This caused conflicting predictions, where spectra data of similar values referred to drastically different targets or vice versa. Proper setup and sample preparation could help alleviate such this data quality issue while keeping the data abundant. Although deep learning models performed better than most conventional multivariate data analysis models, the deep learning model training took longer period to complete due to the nature of deep learning model training not easily understood and vast number of parameters to be optimised, thus finding the optimal design or architecture was a challenge (Paoletti et al., 2019; Shrestha and Mahmood, 2019; Yang et al., 2019). Simpler deep learning model architecture or network would be required to hasten the model training time and eventually be deployed in industrial real time quality monitoring and control applications not just to black pepper, but to other food and agricultural products.

Nevertheless, overall, it was an acceptable success that HSI could be potentially applied in determining the authenticity and geographical origin of black pepper powder samples. On the prediction of chemical and biological analytical properties of black pepper powder samples, it was only possible to some extent that HSI could be used to provide such prediction with considerable accuracy. There are more to be considered on the application of HSI in order to create the rapid, real-time, non-destructive and affordable online detector and estimator of quality, authenticity and geographical origin of Sarawak black pepper powder samples.

6.1 Recommendations

Future research work will be more focused on these following activities:

1. Exploration and acquisition of hyperspectral data of wavelength range beyond 400 – 1000 nm, which is 1000 – 1700 nm, as most literature on the detection of black pepper powder samples cover 1000 – 1700 nm which contains more chemical information and hence better determination of authenticity and geographical origin of black pepper powder samples
2. Piperine content, which is a major component of black pepper will be included in future lab analyses for increased accuracy in the prediction of chemical analytical properties
3. Hyperspectral data in three dimension will be explored for deep learning modelling as combination of spatial-spectral features may contain correlations that can improve the predictive performance of deep learning models
4. Moving stage may be modified to conveyor belt type to simulate real time continuous data gathering in an online quality assessment environment
5. As this research was assessed based on full spectra range, selection of optimal or feature wavelengths will be considered for comparative studies, because selection of optimal wavelengths reduces the model size and runtime, leading to more rapid detection of authenticity and geographical origin of black pepper powder samples
    - There were various dimensionality reduction algorithms studied, for example the commonly used PCA, wavelet transform, independent component analysis, genetic algorithm and successive projection algorithm (Anowar et al., 2021; Saha and Manickavasagan, 2021)
    - Selection of feature wavelengths is one such process of dimensionality reduction, where this process is generally incorporated in between the data pre-processing and model development steps to study the different effects or parameters of dimensionality reduction algorithms on the predictive performance of model.
    - For example, successive projection algorithm and principal component analysis are used to select important wavelengths from the data. Then, the data with selected wavelengths are sent to model development and training where the models are assessed accordingly (Orrillo et al., 2019; Pu et al., 2015; Xin et al., 2020).

REFERENCES

Abbas, O., Zadravec, M., Baeten, V., Mikuš, T., Lešić, T., Vulić, A., Prpić, J., Jemeršić, L., Pleadin, J., 2018. Analytical methods used for the authentication of food of animal origin. Food Chem. 246, 6–17. https://doi.org/10.1016/j.foodchem.2017.11.007

Al-Sarayreh, M., M. Reis, M., Qi Yan, W., Klette, R., 2018. Detection of Red-Meat Adulteration by Deep Spectral–Spatial Features in Hyperspectral Images. J. Imaging 4, 63. https://doi.org/10.3390/jimaging4050063

Al-Sarayreh, M., Reis, M.M., Yan, W.Q., Klette, R., 2020. Potential of deep learning and snapshot hyperspectral imaging for classification of species in meat. Food Control 117, 107332. https://doi.org/10.1016/j.foodcont.2020.107332

Amigo, J.M., Martí, I., Gowen, A., 2013. Chapter 9 - Hyperspectral Imaging and Chemometrics: A Perfect Combination for the Analysis of Food Structure, Composition and Quality, in: Marini, F. (Ed.), Data Handling in Science and Technology, Chemometrics in Food Chemistry. Elsevier, pp. 343–370. https://doi.org/10.1016/B978-0-444-59528-7.00009-0

Anowar, F., Sadaoui, S., Selim, B., 2021. Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE). Comput. Sci. Rev. 40, 100378. https://doi.org/10.1016/j.cosrev.2021.100378

ASTA, 2016. Identification and Prevention of Adulteration [WWW Document]. ASTA. URL https://www.astaspice.org/food-safety/ identification-prevention-adulteration-guidance-document/ (accessed 2.20.21).

Attrey, D.P., 2017. Detection of food adulterants/contaminants, in: Gupta, R.K., Dudeja, Singh Minhas (Eds.), Food Safety in the 21st Century. Academic Press, San Diego, pp. 129–143. https://doi.org/10.1016/B978-0-12-801773-9.00010-8

Bansal, S., Singh, A., Mangal, M., Mangal, A.K., Kumar, S., 2017. Food adulteration: Sources, health risks, and detection methods. Crit. Rev. Food Sci. Nutr. 57, 1174–1189. https://doi.org/10.1080/10408398.2014.967834

Barnes, R.J., Dhanoa, M.S., Lister, S.J., 1989. Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra. Appl. Spectrosc. 43, 772–777. https://doi.org/10.1366/0003702894202201

Bawden, T., 2015. Paprika dishes under investigation as nuts-for-spices scandal grows [WWW Document]. The Independent. URL http://www.independent.co.uk/news/uk/crime/paprika-dishes-under-investigation-as-nuts-for-spices-scandal-grows-10050199.html (accessed 7.11.19).

Bhargava, A., Bansal, A., 2018. Fruits and vegetables quality evaluation using computer vision: A review. J. King Saud Univ. - Comput. Inf. Sci. S131915781830209X. https://doi.org/10.1016/j.jksuci.2018.06.002

Bhattacharjee, P., Singhal, R.S., Gholap, A.S., 2003. Supercritical carbon dioxide extraction for identification of adulteration of black pepper with papaya seeds. J. Sci. Food Agric. 83, 783–786. https://doi.org/10.1002/jsfa.1406

Blazhko, U., Shapaval, V., Kovalev, V., Kohler, A., 2021. Comparison of augmentation and pre-processing for deep learning and chemometric classification of infrared spectra. Chemom. Intell. Lab. Syst. 215, 104367. https://doi.org/10.1016/j.chemolab.2021.104367

Chandra, P., Bajpai, V., Srivastva, M., Kumar, K.B.R., Kumar, B., 2014. Metabolic profiling of Piper species by direct analysis using real time mass spectrometry combined with principal component analysis. Anal. Methods 6, 4234. https://doi.org/10.1039/c4ay00246f

Chen, S., Xiong, J., Guo, W., Bu, R., Zheng, Z., Chen, Y., Yang, Z., Lin, R., 2019. Colored rice quality inspection system using machine vision. J. Cereal Sci. 88, 87–95. https://doi.org/10.1016/j.jcs.2019.05.010

Choi, J.-Y., Heo, S., Bae, S., Kim, J., Moon, K.-D., 2020. Discriminating the origin of basil seeds (Ocimum basilicum L.) using hyperspectral imaging analysis. LWT 118, 108715. https://doi.org/10.1016/j.lwt.2019.108715

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297. https://doi.org/10.1007/BF00994018

Coskun, O., 2016. Separation Tecniques: CHROMATOGRAPHY. North. Clin. Istanb. https://doi.org/10.14744/nci.2016.32757

Curl, C.L., Fenwick, G.R., 1983. On the determination of papaya seed adulteration of black pepper. Food Chem. 12, 241–247. https://doi.org/10.1016/0308-8146(83)90012-2

Danezis, G.P., Tsagkaris, A.S., Camin, F., Brusic, V., Georgiou, C.A., 2016. Food authentication: Techniques, trends & emerging approaches. TrAC Trends Anal. Chem., On-site and In-vivo Instrumentation and Applications 85, 123–132. https://doi.org/10.1016/j.trac.2016.02.026

Dhanoa, M.S., Lister, S.J., Sanderson, R., Barnes, R.J., 1994. The Link between Multiplicative Scatter Correction (MSC) and Standard Normal Variate (SNV) Transformations of NIR Spectra. J. Infrared Spectrosc. 2, 43–47. https://doi.org/10.1255/jnirs.30

Dhanya, K., Syamkumar, S., Sasikumar, B., 2009. Development and Application of SCAR Marker for the Detection of Papaya Seed Adulteration in Traded Black Pepper Powder. Food Biotechnol. 23, 97–106. https://doi.org/10.1080/08905430902873007

Di Rosa, A.R., Leone, F., Cheli, F., Chiofalo, V., 2017. Fusion of electronic nose, electronic tongue and computer vision for animal source food authentication and quality assessment – A review. J. Food Eng. 210, 62–75. https://doi.org/10.1016/j.jfoodeng.2017.04.024

Dissanayake, D.R.R.P., Herath, H.M.P.D., Dissanayake, M.D.M.I.M., Chamikara, M.D.M., Jayakody, M.M., Amaresekara, S.S.C., Kularathna, K.W.T.R., Karannagoda, N.N.H., Ishan, M., Sooriyapathirana, S.D.S.S., 2016. The Length Polymorphism of the Locus psbA-trnH is Idyllic to Detect the Adulterations of Black Pepper with Papaya Seeds and Chili. J. Agric. Sci. – Sri Lanka 11, 74–87. https://doi.org/10.4038/jas.v11i2.8120

Elmasry, G., Kamruzzaman, M., Sun, D.-W., Allen, P., 2012. Principles and Applications of Hyperspectral Imaging in Quality Evaluation of Agro-Food Products: A Review. Crit. Rev. Food Sci. Nutr. 52, 999–1023. https://doi.org/10.1080/10408398.2010.543495

Embuscado, M.E., 2019. Bioactives From Spices and Herbs, in: Melton, L., Shahidi, F., Varelis, P. (Eds.), Encyclopedia of Food Chemistry. Academic Press, Oxford, pp. 497–514. https://doi.org/10.1016/B978-0-08-100596-5.22355-X

Emmert-Streib, F., Yang, Z., Feng, H., Tripathi, S., Dehmer, M., 2020. An Introductory Review of Deep Learning for Prediction Models With Big Data. Front. Artif. Intell. 3, 1–23. https://doi.org/10.3389/frai.2020.00004

Entebang, H., Wong, S.-K., Zehnder Jarroop Augustine, M., 2020. DEVELOPMENT AND PERFORMANCE OF THE PEPPER INDUSTRY IN MALAYSIA: A CRITICAL REVIEW. Int. J. Bus. Soc. 21, 1402–1423.

Fan, J., Ma, C., Zhong, Y., 2019. A Selective Overview of Deep Learning. ArXiv190405526 Cs Math Stat.

FAO, 2019. Production Quantities of Pepper by Countries [WWW Document]. Food Agric. Organ. U. N. URL http://www.fao.org/faostat/en/#data/QC/visualize (accessed 6.20.19).

Galvin-King, P., Haughey, S.A., Elliott, C.T., 2021. Garlic adulteration detection using NIR and FTIR spectroscopy and chemometrics. J. Food Compos. Anal. 96, 103757. https://doi.org/10.1016/j.jfca.2020.103757

Galvin-King, P., Haughey, S.A., Elliott, C.T., 2018. Herb and spice fraud; the drivers, challenges and detection. Food Control 88, 85–97. https://doi.org/10.1016/j.foodcont.2017.12.031

Gao, Z., Shao, Y., Xuan, G., Wang, Y., Liu, Y., Han, X., 2020. Real-time hyperspectral imaging for the in-field estimation of strawberry ripeness with deep learning. Artif. Intell. Agric. 4, 31–38. https://doi.org/10.1016/j.aiia.2020.04.003

García-Mateos, G., Hernández-Hernández, J.L., Escarabajal-Henarejos, D., Jaén-Terrones, S., Molina-Martínez, J.M., 2015. Study and comparison of color models for automatic image analysis in irrigation management applications. Agric. Water Manag., New proposals in the automation and remote control of water management in agriculture: agromotic systems 151, 158–166. https://doi.org/10.1016/j.agwat.2014.08.010

Ge, Z., Song, Z., 2010. A comparative study of just-in-time-learning based methods for online soft sensor modeling. Chemom. Intell. Lab. Syst. 104, 306–317. https://doi.org/10.1016/j.chemolab.2010.09.008

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press.

Govindarajan, V.S., Stahl, W.H., 1977. Pepper — chemistry, technology, and quality evaluation. C R C Crit. Rev. Food Sci. Nutr. 9, 115–225. https://doi.org/10.1080/10408397709527233

Gowen, A.A., O'Donnell, C.P., Cullen, P.J., Downey, G., Frias, J.M., 2007. Hyperspectral imaging – an emerging process analytical tool for food quality and safety control. Trends Food Sci. Technol. 18, 590–598. https://doi.org/10.1016/j.tifs.2007.06.001

Gul, I., Nasrullah, N., Nissar, U., Saifi, M., Abdin, M.Z., 2018. Development of DNA and GC-MS Fingerprints for Authentication and Quality Control of Piper nigrum L. and Its Adulterant Carica papaya L. Food Anal. Methods 11, 1209–1222. https://doi.org/10.1007/s12161-017-1088-7

Guo, Z., Huang, W., Chen, L., Zhao, C., Peng, Y., 2013. Geographical classification of apple based on hyperspectral imaging, in: Sensing for Agriculture and Food Quality and Safety V. International Society for Optics and Photonics, p. 87210J. https://doi.org/10.1117/12.2015559

Han, Y., Liu, Z., Khoshelham, K., Bai, S.H., 2021. Quality estimation of nuts using deep learning classification of hyperspectral imagery. Comput. Electron. Agric. 180, 105868. https://doi.org/10.1016/j.compag.2020.105868

Hong, Z., Zhang, C., Kong, D., Qi, Z., He, Y., 2021. Identification of storage years of black tea using near-infrared hyperspectral imaging with deep learning methods. Infrared Phys. Technol. 114, 103666. https://doi.org/10.1016/j.infrared.2021.103666

Hu, L., Yin, C., Ma, S., Liu, Z., 2018. Assessing the authenticity of black pepper using diffuse reflectance mid-infrared Fourier transform spectroscopy coupled with chemometrics. Comput. Electron. Agric. 154, 491–500. https://doi.org/10.1016/j.compag.2018.09.029

IPC, 2021. Pepper Prices [WWW Document]. Int. PEPPER COMMUNITY. URL http://www.ipcnet.org/price/?p=d

IPC, 2018. Malaysia, INTERNATIONAL PEPPER COMMUNITY [WWW Document]. URL http://www.ipcnet.org/cp/?p=d&id=3&start=3 (accessed 6.19.19).

IPC, 2015. IPC Standard Specifications for Black/ White Pepper (Whole and Ground) and Whole Dehydrated Green Pepper [WWW Document]. URL http://www.fao.org/fao-who-codexalimentarius/sh-proxy/en/?lnk=1&url=https%253A%252F%252Fworkspace.fao.org%252Fsites%252Fcodex%252FMeetings%252FCX-736-03%252FCRD%252Fsc03_CRD07x.pdf

Jain, S.C., Menghani, E., Jain, R., 2007. Fluorescence and HPLC-Based Standardization of Piper nigrum Fruits. Int. J. Bot. 3, 208–213.

Jha, S.N., 2016. Chapter 6 - Spectroscopy and Chemometrics, in: Jha, S.N. (Ed.), Rapid Detection of Food Adulterants and Contaminants. Academic Press, San Diego, pp. 147–214. https://doi.org/10.1016/B978-0-12-420084-5.00006-8

Jiang, B., He, J., Yang, S., Fu, H., Li, T., Song, H., He, D., 2019. Fusion of machine vision technology and AlexNet-CNNs deep learning network for the detection of postharvest apple pesticide residues. Artif. Intell. Agric. 1, 1–8. https://doi.org/10.1016/j.aiia.2019.02.001

Jin, X., Jie, L., Wang, S., Qi, J.H., Li, W.S., 2018. Classifying Wheat Hyperspectral Pixels of Healthy Heads and Fusarium Head Blight Disease Using a Deep Neural Network in the Wild Field. Remote Sens. 10. https://doi.org/10.3390/rs10030395

Johny, F., Saupi, N., Ramaiya, S.D., 2020. Status of Pepper Farming and Flower Composition of Different Pepper Varieties in Sarawak. Pertanika J. Trop. Agric. Sci. 43. https://doi.org/10.47836/pjtas.43.4.04

Kamruzzaman, M., Makino, Y., Oshita, S., 2014. An appraisal of hyperspectral imaging for non-invasive authentication of geographical origin of beef and pork, in: International Conference of Agricultural Engineering. Zurich, pp. 6–10.

Kassie, F., Pool-Zobel, B., Parzefall, W., Knasmüller, S., 1999. Genotoxic effects of benzyl isothiocyanate, a natural chemopreventive agent. Mutagenesis 14, 595–604. https://doi.org/10.1093/mutage/14.6.595

Ke, J., Rao, L., Zhou, L., Chen, X., Zhang, Z., 2020. Non-destructive determination of volatile oil and moisture content and discrimination of geographical origins of Zanthoxylum bungeanum Maxim. by hyperspectral imaging. Infrared Phys. Technol. 105, 103185. https://doi.org/10.1016/j.infrared.2020.103185

Keong, M.S., 2017. Bright future for Sarawak pepper [WWW Document]. Star Online. URL https://www.thestar.com.my/news/nation/2017/10/25/malaysia-aims-to-be-worlds-top-supplier-of-the-premium-king-of-spices/ (accessed 6.19.19).

Kermanshai, R., McCarry, B.E., Rosenfeld, J., Summers, P.S., Weretilnyk, E.A., Sorger, G.J., 2001. Benzyl isothiocyanate is the chief or sole anthelmintic in papaya seed extracts. Phytochemistry 57, 427–435. https://doi.org/10.1016/S0031-9422(01)00077-2

Khan, A., Munir, M.T., Yu, W., Young, B.R., 2020. A Review Towards Hyperspectral Imaging for Real-Time Quality Control of Food Products with an Illustrative Case Study of Milk Powder Production. Food Bioprocess Technol. 13, 739–752. https://doi.org/10.1007/s11947-020-02433-w

Khan, A., Sohail, A., Zahoora, U., Qureshi, A.S., 2019. A Survey of the Recent Architectures of Deep Convolutional Neural Networks. ArXiv190106032 Cs.

Khan, S., Mirza, K.J., Anwar, F., Abdin, M.Z., 2010. Development of RAPD markers for authentication of Piper nigrum (L.). Environ. We Int. J. Sci. Technol. 5, 47–56.

Kherif, F., Latypova, A., 2020. Chapter 12 - Principal component analysis, in: Mechelli, A., Vieira, S. (Eds.), Machine Learning. Academic Press, pp. 209–225. https://doi.org/10.1016/B978-0-12-815739-8.00012-2

Kingma, D.P., Ba, J., 2017. Adam: A Method for Stochastic Optimization. ArXiv14126980 Cs.

Koirala, A., Walsh, K.B., Wang, Z., McCarthy, C., 2019. Deep learning – Method overview and review of use for fruit detection and yield estimation. Comput. Electron. Agric. 162, 219–234. https://doi.org/10.1016/j.compag.2019.04.017

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet Classification with Deep Convolutional Neural Networks, in: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, NIPS'12. Curran Associates Inc., USA, pp. 1097–1105.

Kucharska-Ambrożej, K., Karpinska, J., 2020. The application of spectroscopic techniques in combination with chemometrics for detection adulteration of some herbs and spices. Microchem. J. 153, 104278. https://doi.org/10.1016/j.microc.2019.104278

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444. https://doi.org/10.1038/nature14539

Li, S., Song, W., Fang, L., Chen, Y., Ghamisi, P., Benediktsson, J.A., 2019. Deep Learning for Hyperspectral Image Classification: An Overview. IEEE Trans. Geosci. Remote Sens. 1–20. https://doi.org/10.1109/TGRS.2019.2907932

Li, Y., Zhang, C., Pan, S., Chen, L., Liu, M., Yang, K., Zeng, X., Tian, J., 2020. Analysis of chemical components and biological activities of essential oils from black and white pepper (Piper nigrum L.) in five provinces of southern China. LWT 117, 108644. https://doi.org/10.1016/j.lwt.2019.108644

Li, Y., Zhang, H., Shen, Q., 2017. Spectral–Spatial Classification of Hyperspectral Imagery with 3D Convolutional Neural Network. Remote Sens. 9, 67. https://doi.org/10.3390/rs9010067

Liang, J., Sun, J., Chen, P., Frazier, J., Benefield, V., Zhang, M., 2021. Chemical analysis and classification of black pepper (Piper nigrum L.) based on their country of origin using mass spectrometric methods and chemometrics. Food Res. Int. 140, 109877. https://doi.org/10.1016/j.foodres.2020.109877

Lima, A.B.S. de, Batista, A.S., Jesus, J.C. de, Silva, J. de J., Araújo, A.C.M. de, Santos, L.S., 2020. Fast quantitative detection of black pepper and cumin adulterations by near-infrared spectroscopy and multivariate modeling. Food Control 107, 106802. https://doi.org/10.1016/j.foodcont.2019.106802

Liu, D., Sun, D.-W., Zeng, X.-A., 2014. Recent Advances in Wavelength Selection Techniques for Hyperspectral Image Processing in the Food Industry. Food Bioprocess Technol. 7, 307–323. https://doi.org/10.1007/s11947-013-1193-6

Liu, Y., Pu, H., Sun, D.-W., 2017. Hyperspectral imaging technique for evaluating food quality and safety during various processes: A review of recent applications. Trends Food Sci. Technol. 69, 25–35. https://doi.org/10.1016/j.tifs.2017.08.013

Lohumi, S., Lee, S., Lee, H., Cho, B.-K., 2015. A review of vibrational spectroscopic techniques for the detection of food authenticity and adulteration. Trends Food Sci. Technol. 46, 85–98. https://doi.org/10.1016/j.tifs.2015.08.003

Ma, T., Tsuchikawa, S., Inagaki, T., 2020. Rapid and non-destructive seed viability prediction using near-infrared hyperspectral imaging coupled with a deep learning approach. Comput. Electron. Agric. 177, 105683. https://doi.org/10.1016/j.compag.2020.105683

Madan, M.M., Singhal, R.S., Kulkarni, P.R., 1996. An approach into the detection of authenticity of black pepper (Piper nigrum L.) oleoresin. J. Spices Aromat. Crops 5, 64–67.

Mahesh, S., Jayas, D.S., Paliwal, J., White, N.D.G., 2015. Hyperspectral imaging to classify and monitor quality of agricultural materials. J. Stored Prod. Res. 61, 17–26. https://doi.org/10.1016/j.jspr.2015.01.006

Manley, M., 2014. Near-infrared spectroscopy and hyperspectral imaging: non-destructive analysis of biological materials. Chem. Soc. Rev. 43, 8200–8214. https://doi.org/10.1039/C4CS00062E

McGoverin, C.M., September, D.J.F., Geladi, P., Manley, M., 2012. Near Infrared and Mid-Infrared Spectroscopy for the Quantification of Adulterants in Ground Black Pepper. J. Infrared Spectrosc. 20, 521–528. https://doi.org/10.1255/jnirs.1008

Mercer, Z.J.A., Chua, H.S., Mahon, P., Hwang, S.S., Ng, S.M., 2019. Authentication of geographical growth origin of black pepper (piper nigrum l.) based on volatile organic compounds profile: A case study for Malaysia and India black peppers, in: 2019 IEEE International Symposium on Olfaction and Electronic Nose (ISOEN). Presented at the 2019 IEEE International Symposium on Olfaction and Electronic Nose (ISOEN), pp. 1–3. https://doi.org/10.1109/ISOEN.2019.8823265

Modi, A.S., 2018. Review Article on Deep Learning Approaches. Presented at the 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 1635–1639. https://doi.org/10.1109/ICCONS.2018.8663057

Modupalli, N., Naik, M., Sunil, C.K., Natarajan, V., 2021. Emerging non-destructive methods for quality and safety monitoring of spices. Trends Food Sci. Technol. 108, 133–147. https://doi.org/10.1016/j.tifs.2020.12.021

Mohd Uzir Mahidin, 2020a. Selected Agricultural Indicators, Malaysia [WWW Document]. Dep. Stat. Malays. Off. Portal. URL https://www.dosm.gov.my/v1/index.php?r=column/cthemeByCat&cat=72&bul_id=RXVKUVJ5TitHM0cwYWxlOHcxU3dKdz09&menu_id=Z0VTZGU1UHBUT1VJMFlpaXRRR0xpdz09 (accessed 6.20.19).

Mohd Uzir Mahidin, 2020b. Malaysia Economic Performance Fourth Quarter 2019 [WWW Document]. Dep. Stat. Malays. Off. Portal. URL https://www.dosm.gov.my/v1/index.php?r=column/cthemeByCat&cat=100&bul_id =WWk2MDA3R1k1SlVsTjlzU3FZcjVlUT09&menu_id=TE5CRUZCblh4ZTZMO DZIbmk2aWRRQT09 (accessed 2.13.20).

MPB, 2021. Official Portal Of Malaysian Pepper Board - MAIN [WWW Document]. URL https://www.mpb.gov.my/mpb/index.php/en/

Nagasubramanian, K., Jones, S., Singh, A.K., Singh, A., Ganapathysubramanian, B., Sarkar, S., 2018. Explaining hyperspectral imaging based plant disease identification: 3D CNN and saliency maps. ArXiv180408831 Cs.

Ohta, Y.-I., Kanade, T., Sakai, T., 1980. Color information for region segmentation. Comput. Graph. Image Process. 13, 222–241. https://doi.org/10.1016/0146-664X(80)90047-7

Oliveira, M.M., Cruz-Tirado, J.P., Barbin, D.F., 2019. Nontargeted Analytical Methods as a Powerful Tool for the Authentication of Spices and Herbs: A Review. Compr. Rev. Food Sci. Food Saf. 18, 670–689. https://doi.org/10.1111/1541-4337.12436

Orrillo, I., Cruz-Tirado, J.P., Cardenas, A., Oruna, M., Carnero, A., Barbin, D.F., Siche, R., 2019. Hyperspectral imaging as a powerful tool for identification of papaya seeds in black pepper. Food Control 101, 45–52. https://doi.org/10.1016/j.foodcont.2019.02.036

P. R. Goswami, K. R. Jain, 2013. Non-destructive quality evaluation in spice industry with specific reference to black pepper (Piper Nigrum L.). Presented at the 2013 Nirma University International Conference on Engineering (NUiCONE), pp. 1–5. https://doi.org/10.1109/NUiCONE.2013.6780106

Paoletti, M.E., Haut, J.M., Plaza, J., Plaza, A., 2019. Deep learning classifiers for hyperspectral imaging: A review. ISPRS J. Photogramm. Remote Sens. 158, 279–317. https://doi.org/10.1016/j.isprsjprs.2019.09.006

Paradkar, M.M., Singhal, R.S., Kulkarni, P.R., 2001. A new TLC method to detect the presence of ground papaya seed in ground black pepper. J. Sci. Food Agric. 81, 1322–1325. https://doi.org/10.1002/jsfa.946

Park, B., Lu, R. (Eds.), 2015. Hyperspectral Imaging Technology in Food and Agriculture, Food Engineering Series. Springer New York, New York, NY. https://doi.org/10.1007/978-1-4939-2836-1

Parvathy, V.A., Swetha, V.P., Sheeja, T.E., Leela, N.K., Chempakam, B., Sasikumar, B., 2014. DNA Barcoding to Detect Chilli Adulteration in Traded Black Pepper Powder. Food Biotechnol. 28, 25–40. https://doi.org/10.1080/08905436.2013.870078

Patel, K.K., Kar, A., Jha, S.N., Khan, M.A., 2012. Machine vision system: a tool for quality inspection of food and agricultural products. J. Food Sci. Technol. 49, 123–141. https://doi.org/10.1007/s13197-011-0321-4

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Petersson, H., Gustafsson, D., Bergstrom, D., 2016. Hyperspectral image analysis using deep learning — A review. Presented at the 2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA), pp. 1–6. https://doi.org/10.1109/IPTA.2016.7820963

Pu, H., Kamruzzaman, M., Sun, D.-W., 2015. Selection of feature wavelengths for developing multispectral imaging systems for quality, safety and authenticity of muscle foods-a review. Trends Food Sci. Technol. 45, 86–104. https://doi.org/10.1016/j.tifs.2015.05.006

Qin, J., Vasefi, F., Hellberg, R.S., Akhbardeh, A., Isaacs, R.B., Yilmaz, A.G., Hwang, C., Baek, I., Schmidt, W.F., Kim, M.S., 2020. Detection of fish fillet substitution and mislabeling using multimode hyperspectral imaging techniques. Food Control 114, 107234. https://doi.org/10.1016/j.foodcont.2020.107234

Qureshi, R., Uzair, M., Khurshid, K., Yan, H., 2019. Hyperspectral document image processing: Applications, challenges and future prospects. Pattern Recognit. 90, 12–22. https://doi.org/10.1016/j.patcog.2019.01.026

Ravindran, P.N., Kallupurackal, J.A., 2001. 7. Black Pepper, in: Handbook of Herbs and Spices. Woodhead Publishing.

Reinholds, I., Bartkevics, V., Silvis, I.C.J., van Ruth, S.M., Esslinger, S., 2015. Analytical techniques combined with chemometrics for authentication and determination of contaminants in condiments: A review. J. Food Compos. Anal. 44, 56–72. https://doi.org/10.1016/j.jfca.2015.05.004

Rivera-Pérez, A., Romero-González, R., Garrido Frenich, A., 2021. Feasibility of Applying Untargeted Metabolomics with GC-Orbitrap-HRMS and Chemometrics for Authentication of Black Pepper (Piper nigrum L.) and Identification of Geographical and Processing Markers. J. Agric. Food Chem. 69, 5547–5558. https://doi.org/10.1021/acs.jafc.1c01515

Rodriguez-Saona, L.E., Giusti, M.M., Shotts, M., 2016. Advances in Infrared Spectroscopy for Food Authenticity Testing, in: Downey, G. (Ed.), Advances in Food Authenticity Testing, Woodhead Publishing Series in Food Science, Technology and Nutrition. Woodhead Publishing, pp. 71–116. https://doi.org/10.1016/B978-0-08-100220-9.00004-7

Rong, D., Xie, L., Ying, Y., 2019. Computer vision detection of foreign objects in walnuts using deep learning. Comput. Electron. Agric. 162, 1001–1010. https://doi.org/10.1016/j.compag.2019.05.019

Saha, D., Manickavasagan, A., 2021. Machine learning techniques for analysis of hyperspectral images to determine quality of food products: A review. Curr. Res. Food Sci. 4, 28–44. https://doi.org/10.1016/j.crfs.2021.01.002

Savitzky, Abraham., Golay, M.J.E., 1964. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Anal. Chem. 36, 1627–1639. https://doi.org/10.1021/ac60214a047

Scholkopf, B., Kah-Kay Sung, Burges, C.J.C., Girosi, F., Niyogi, P., Poggio, T., Vapnik, V., 1997. Comparing support vector machines with Gaussian kernels to radial basis function classifiers. IEEE Trans. Signal Process. 45, 2758–2765. https://doi.org/10.1109/78.650102

September, D.J.F., 2011. Detection and quantification of spice adulteration by near infrared hyperspectral imaging (Thesis). Stellenbosch : University of Stellenbosch.

Shih, P., Liu, C., 2005. Comparative Assessment of Content-Based Face Image Retrieval in Different Color Spaces, in: Kanade, T., Jain, A., Ratha, N.K. (Eds.), Audio- and Video-Based Biometric Person Authentication, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 1039–1048.

Shityakov, S., Bigdelian, E., Hussein, A.A., Hussain, M.B., Tripathi, Y.C., Khan, M.U., Shariati, M.A., 2019. Phytochemical and pharmacological attributes of piperine: A bioactive ingredient of black pepper. Eur. J. Med. Chem. 176, 149–161. https://doi.org/10.1016/j.ejmech.2019.04.002

Shorten, P.R., Leath, S.R., Schmidt, J., Ghamkhar, K., 2019. Predicting the quality of ryegrass using hyperspectral imaging. Plant Methods 15, 63. https://doi.org/10.1186/s13007-019-0448-2

Shrestha, A., Mahmood, A., 2019. Review of Deep Learning Algorithms and Architectures. IEEE Access 7, 53040–53065. https://doi.org/10.1109/ACCESS.2019.2912200

Signoroni, A., Savardi, M., Baronio, A., Benini, S., 2019. Deep Learning Meets Hyperspectral Image Analysis: A Multidisciplinary Review. J. Imaging 5, 52. https://doi.org/10.3390/jimaging5050052

SMA, 2019. Agriculture [WWW Document]. Sarawak Multimed. Auth. URL https://www.sma.gov.my/pages.php?mod=webpage&sub=page&id=18 (accessed 7.15.19).

Sørensen, K.M., Khakimov, B., Engelsen, S.B., 2016. The use of rapid spectroscopic screening methods to detect adulteration of food raw materials and ingredients. Curr. Opin. Food Sci., Innovation in food science • Foodomics technologies 10, 45–51. https://doi.org/10.1016/j.cofs.2016.08.001

Sousa, A.I., Ferreira, I.M.P.L.V.O., Faria, M.A., 2019. Sensitive detection of Piper nigrum L. adulterants by a novel screening approach based on qPCR. Food Chem. 283, 596–603. https://doi.org/10.1016/j.foodchem.2019.01.062

Steinbrener, J., Posch, K., Leitner, R., 2019. Hyperspectral fruit and vegetable classification using convolutional neural networks. Comput. Electron. Agric. 162, 364–372. https://doi.org/10.1016/j.compag.2019.04.019

Sun, Y., Li, Y., Pan, L., Abbas, A., Jiang, Y., Wang, X., 2021. Authentication of the geographic origin of Yangshan region peaches based on hyperspectral imaging. Postharvest Biol. Technol. 171, 111320. https://doi.org/10.1016/j.postharvbio.2020.111320

Suthaharan, S., 2016. Support Vector Machine, in: Suthaharan, S. (Ed.), Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning, Integrated Series in Information Systems. Springer US, Boston, MA, pp. 207–235. https://doi.org/10.1007/978-1-4899-7641-3_9

Taheri-Garavand, A., Fatahi, S., Omid, M., Makino, Y., 2019. Meat quality evaluation based on computer vision technique: A review. Meat Sci. 156, 183–195. https://doi.org/10.1016/j.meatsci.2019.06.002

Terrillon, J.-C., M. N. Shirazi, H. Fukamachi, S. Akamatsu, 2000. Comparative performance of different skin chrominance models and chrominance spaces for the automatic detection of human faces in color images, in: Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580). pp. 54–61. https://doi.org/10.1109/AFGR.2000.840612

Tremlová, B., 2001. Evidence of spice black pepper adulteration. Czech J. Food Sci. 19, 235–239. https://doi.org/10.17221/6613-CJFS

Vadivel, V., Ravichandran, N., Rajalakshmi, P., Brindha, P., Gopal, A., Kumaravelu, C., 2018. Microscopic, phytochemical, HPTLC, GC–MS and NIRS methods to differentiate herbal adulterants: Pepper and papaya seeds. J. Herb. Med. 11, 36–45. https://doi.org/10.1016/j.hermed.2018.01.004

Veraverbeke, S., Dennison, P., Gitas, I., Hulley, G., Kalashnikova, O., Katagis, T., Kuai, L., Meng, R., Roberts, D., Stavros, N., 2018. Hyperspectral remote sensing of fire: State-of-the-art and future perspectives. Remote Sens. Environ. 216, 105–121. https://doi.org/10.1016/j.rse.2018.06.020

Vithu, P., Moses, J.A., 2016. Machine vision system for food grain quality evaluation: A review. Trends Food Sci. Technol. 56, 13–20. https://doi.org/10.1016/j.tifs.2016.07.011

Wang, T., Han, Q., Qin, H., Li, Y., 2019. Geographical Origin Identification of Glycyrrhiza Uralensis Fisch Seeds by Hyperspectral Imaging Technology. Presented at the 2019 International Conference on High Performance Big Data and Intelligent Systems (HPBD IS), pp. 200–203. https://doi.org/10.1109/HPBDIS.2019.8735461

Weng, S., Tang, P., Yuan, H., Guo, B., Yu, S., Huang, L., Xu, C., 2020. Hyperspectral imaging for accurate determination of rice variety using a deep learning network with multi-feature fusion. Spectrochim. Acta. A. Mol. Biomol. Spectrosc. 234, 118237. https://doi.org/10.1016/j.saa.2020.118237

Wilde, A.S., Haughey, S.A., Galvin-King, P., Elliott, C.T., 2019. The feasibility of applying NIR and FT-IR fingerprinting to detect adulteration in black pepper. Food Control 100, 1–7. https://doi.org/10.1016/j.foodcont.2018.12.039

Wu, D., Sun, D.-W., 2013a. Colour measurements by computer vision for food quality control – A review. Trends Food Sci. Technol. 29, 5–20. https://doi.org/10.1016/j.tifs.2012.08.004

Wu, D., Sun, D.-W., 2013b. Advanced applications of hyperspectral imaging technology for food quality and safety analysis and assessment: A review — Part I: Fundamentals. Innov. Food Sci. Emerg. Technol. 19, 1–14. https://doi.org/10.1016/j.ifset.2013.04.014

Wu, D., Sun, D.-W., 2013c. Advanced applications of hyperspectral imaging technology for food quality and safety analysis and assessment: A review — Part II: Applications. Innov. Food Sci. Emerg. Technol. 19, 15–28. https://doi.org/10.1016/j.ifset.2013.04.016

Xin, Z., Jun, S., Yan, T., Quansheng, C., Xiaohong, W., Yingying, H., 2020. A deep learning based regression method on hyperspectral data for rapid prediction of cadmium residue in lettuce leaves. Chemom. Intell. Lab. Syst. 200, 103996. https://doi.org/10.1016/j.chemolab.2020.103996

Xu, J.-L., Riccioli, C., Sun, D.-W., 2017. Comparison of hyperspectral imaging and computer vision for automatic differentiation of organically and conventionally farmed salmon. J. Food Eng. 196, 170–182. https://doi.org/10.1016/j.jfoodeng.2016.10.021

Xu, M., Wang, J., Gu, S., 2019. Rapid identification of tea quality by E-nose and computer vision combining with a synergetic data fusion strategy. J. Food Eng. 241, 10–17. https://doi.org/10.1016/j.jfoodeng.2018.07.020

Yang, J., Xu, J., Zhang, X., Wu, C., Lin, T., Ying, Y., 2019. Deep learning for vibrational spectral analysis: Recent progress and a practical guide. Anal. Chim. Acta. https://doi.org/10.1016/j.aca.2019.06.012

Yu, X., Lu, H., Liu, Q., 2018. Deep-learning-based regression model and hyperspectral imaging for rapid detection of nitrogen concentration in oilseed rape (Brassica napus L.) leaf. Chemom. Intell. Lab. Syst. 172, 188–193. https://doi.org/10.1016/j.chemolab.2017.12.010

Yu, X., Wang, J., Wen, S., Yang, J., Zhang, F., 2019. A deep learning based feature extraction method on hyperspectral images for nondestructive prediction of TVB-N content in Pacific white shrimp (Litopenaeus vannamei). Biosyst. Eng. 178, 244–255. https://doi.org/10.1016/j.biosystemseng.2018.11.018

Zhang, C., Wu, W., Zhou, L., Cheng, H., Ye, X., He, Y., 2020. Developing deep learning based regression approaches for determination of chemical compositions in dry black goji berries (Lycium ruthenicum Murr.) using near-infrared hyperspectral imaging. Food Chem. 319, 126536. https://doi.org/10.1016/j.foodchem.2020.126536

Zhang, L., Sun, H., Rao, Z., Ji, H., 2020. Hyperspectral imaging technology combined with deep forest model to identify frost-damaged rice seeds. Spectrochim. Acta. A. Mol. Biomol. Spectrosc. 229, 117973. https://doi.org/10.1016/j.saa.2019.117973

Zhang, M., de B. Harrington, P., Chen, P., 2015. Classification of Cultivation Locations of Black Pepper (Piper nigrum L.) using Gas Chromatography and Chemometrics. Curr. Chromatogr. 2, 145–151.

Zhang, M., Shi, Y., Sun, W., Wu, L., Xiong, C., Zhu, Z., Zhao, H., Zhang, B., Wang, C., Liu, X., 2019. An efficient DNA barcoding based method for the authentication and adulteration detection of the powdered natural spices. Food Control 106745. https://doi.org/10.1016/j.foodcont.2019.106745

Zhou, L., Zhang, C., Liu, F., Qiu, Z., He, Y., 2019. Application of Deep Learning in Food: A Review. Compr. Rev. Food Sci. Food Saf. 18, 1793–1811. https://doi.org/10.1111/1541-4337.12492

Zhou, Z., Morel, J., Parsons, D., Kucheryavskiy, S.V., Gustavsson, A.-M., 2019. Estimation of yield and quality of legume and grass mixtures using partial least squares and support vector machine analysis of spectral data. Comput. Electron. Agric. 162, 246–253. https://doi.org/10.1016/j.compag.2019.03.038

APPENDIX A

NEAR INFRARED BANDS AND SPECTRA STRUCTURE

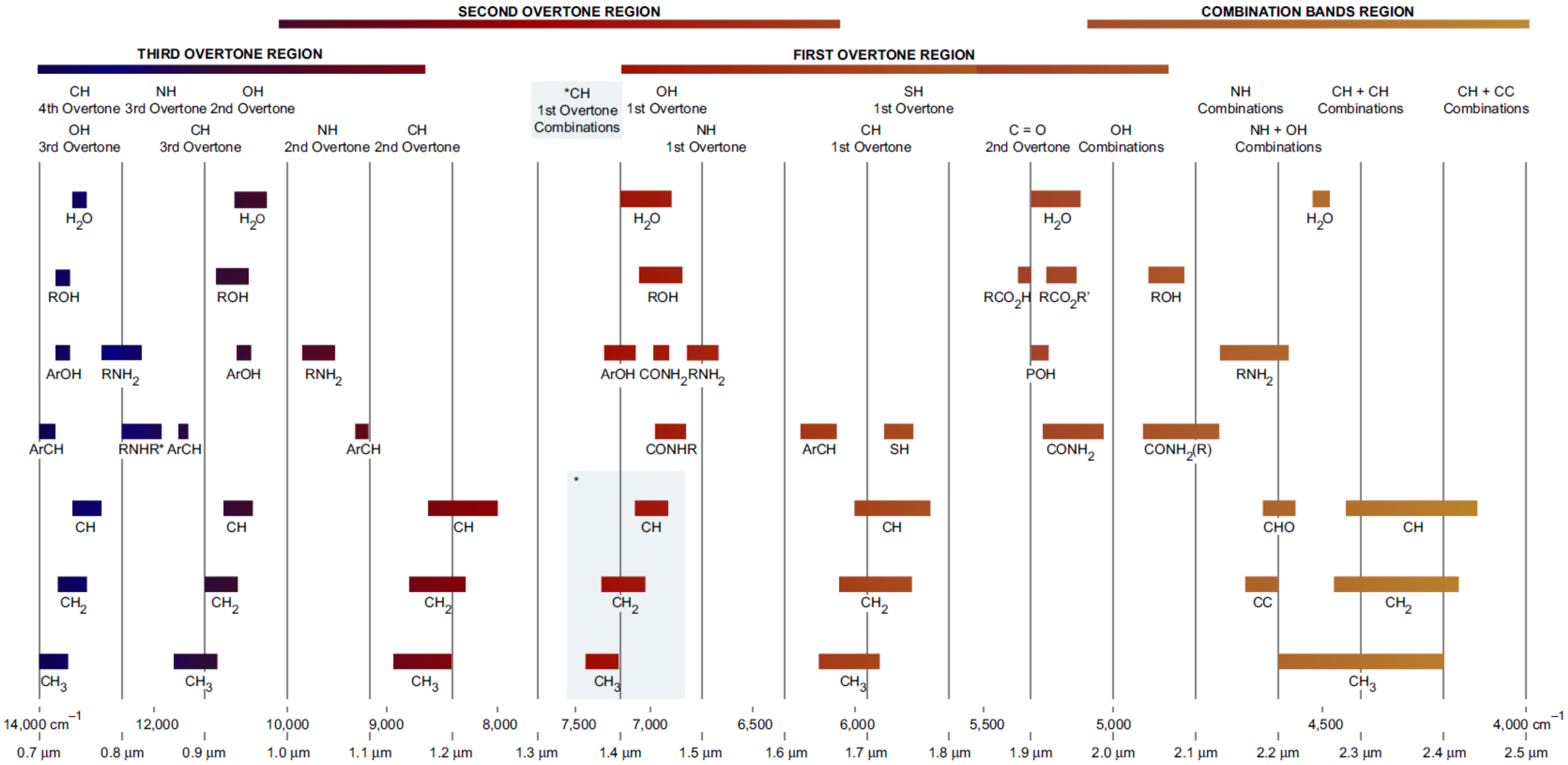


Figure A.1: NIR bands with associated spectra structure (Jha, 2016)

APPENDIX B

SUPPLEMENTAL FIGURES AND TABLES FROM CHAPTER 4

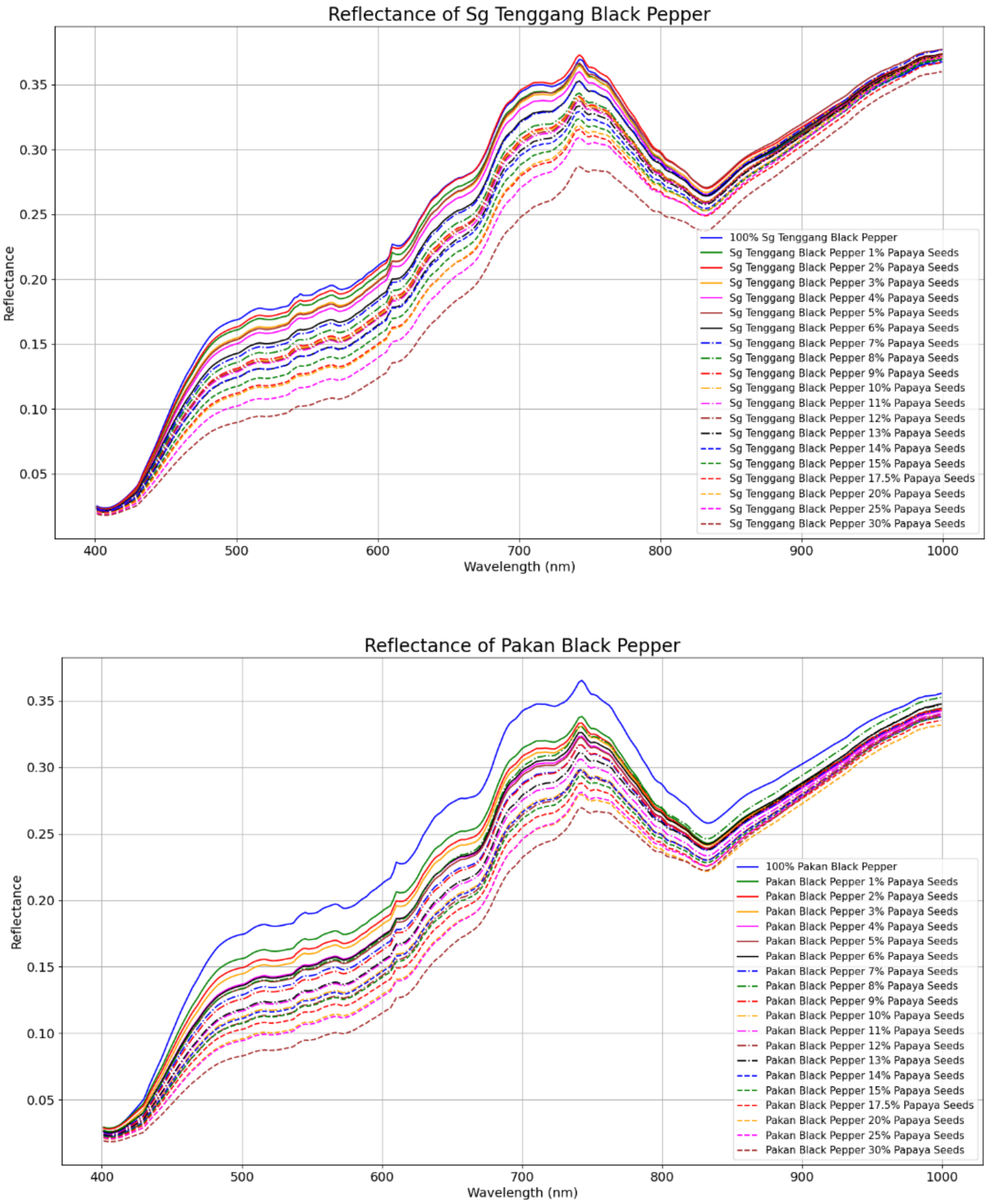


Figure B.1: Reflectance of raw mean spectra data of pure and adulterated black pepper powder samples [Above: From Sg Tenggang; Below: From Pakan]

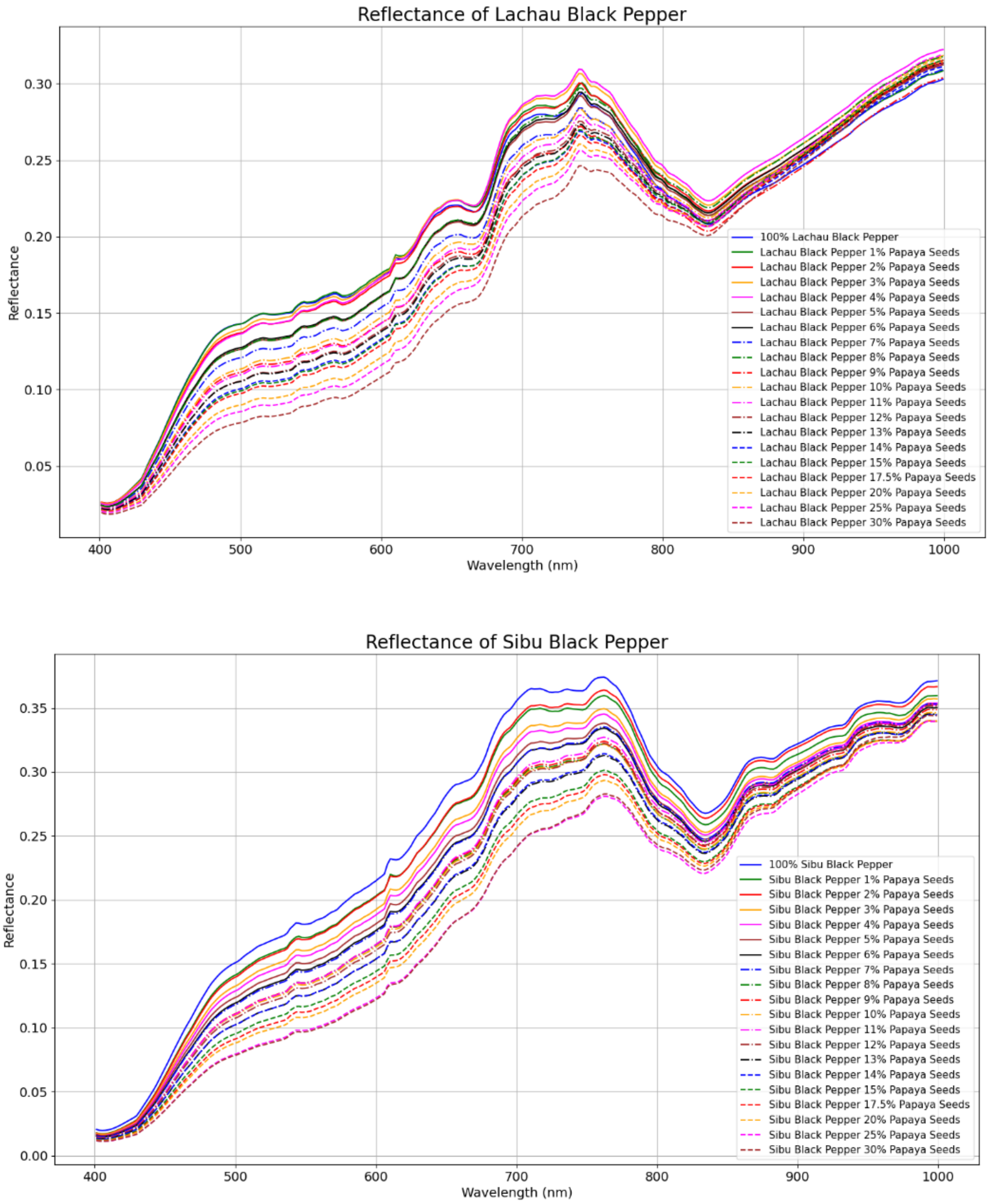


Figure B.2: (Continued from Figure B.) Reflectance of raw mean spectra data of pure and adulterated black pepper powder samples [Above: From Lachau; Below: From Sibu]

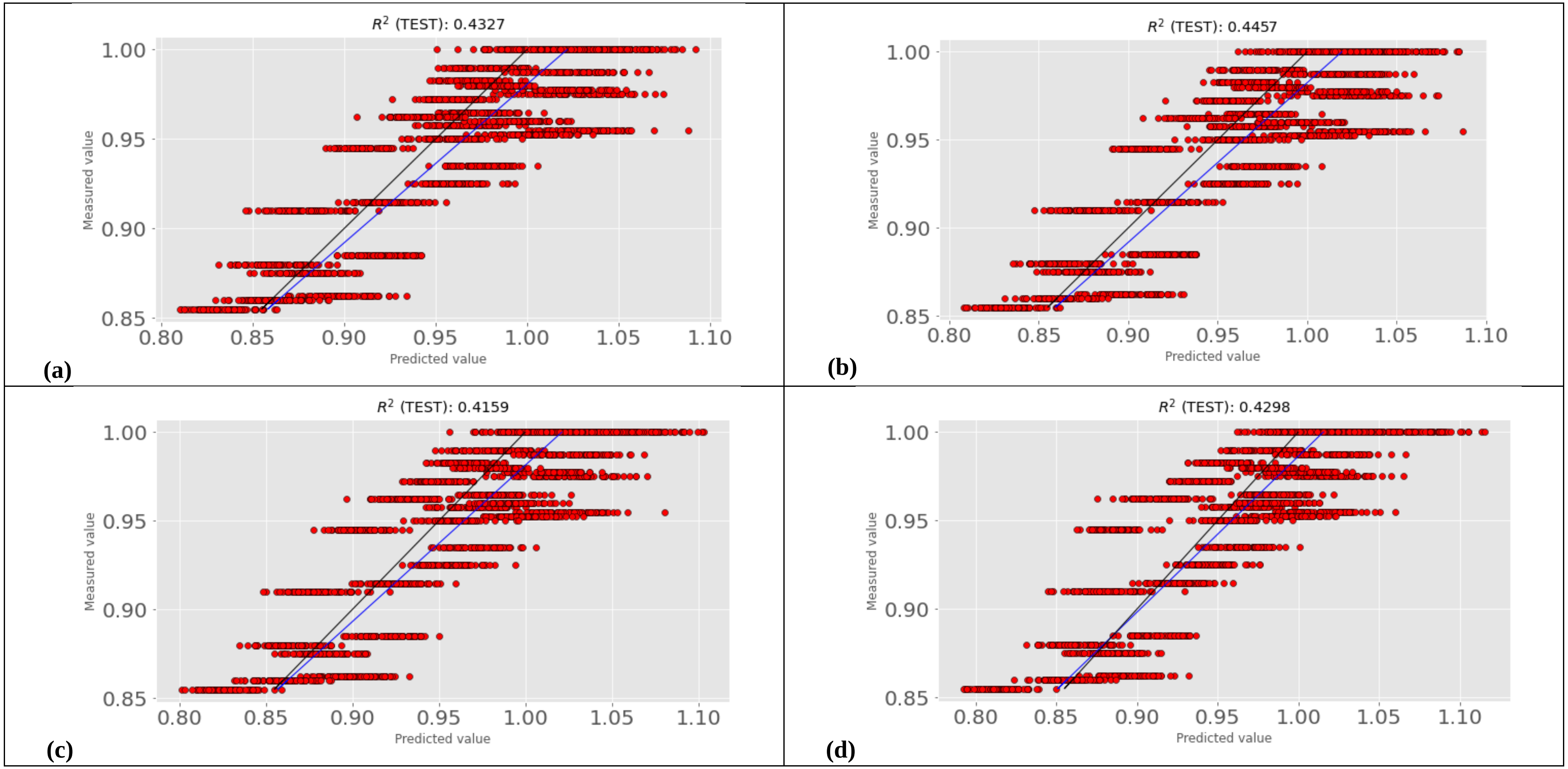


Figure B.3: Parity plots of PLS models for determination of authenticity of black pepper powder samples with different data pre-processing methods [(a): Raw/without pre-processing; (b): SG; (c): SG-1st; (d): SG-2nd]

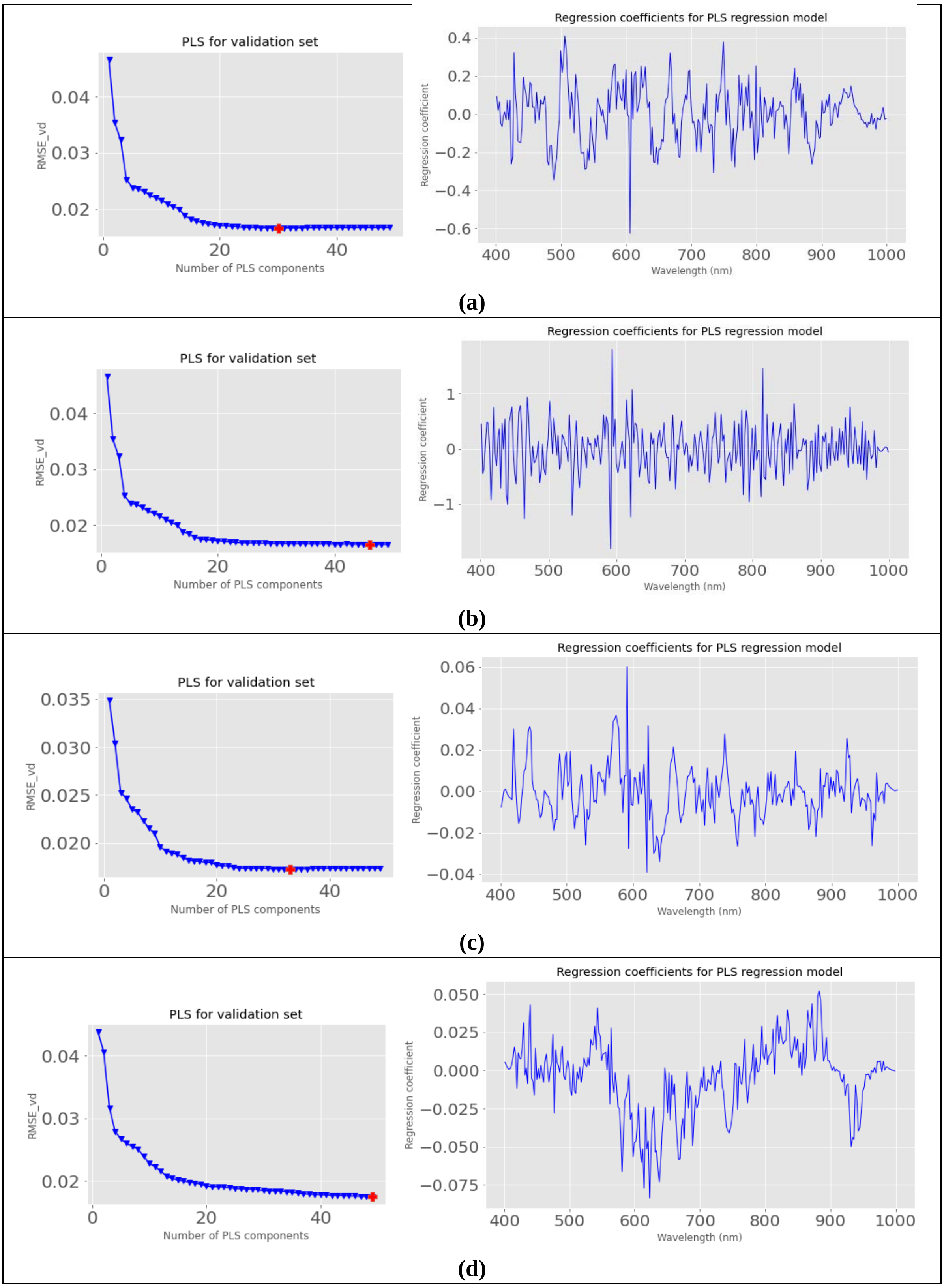
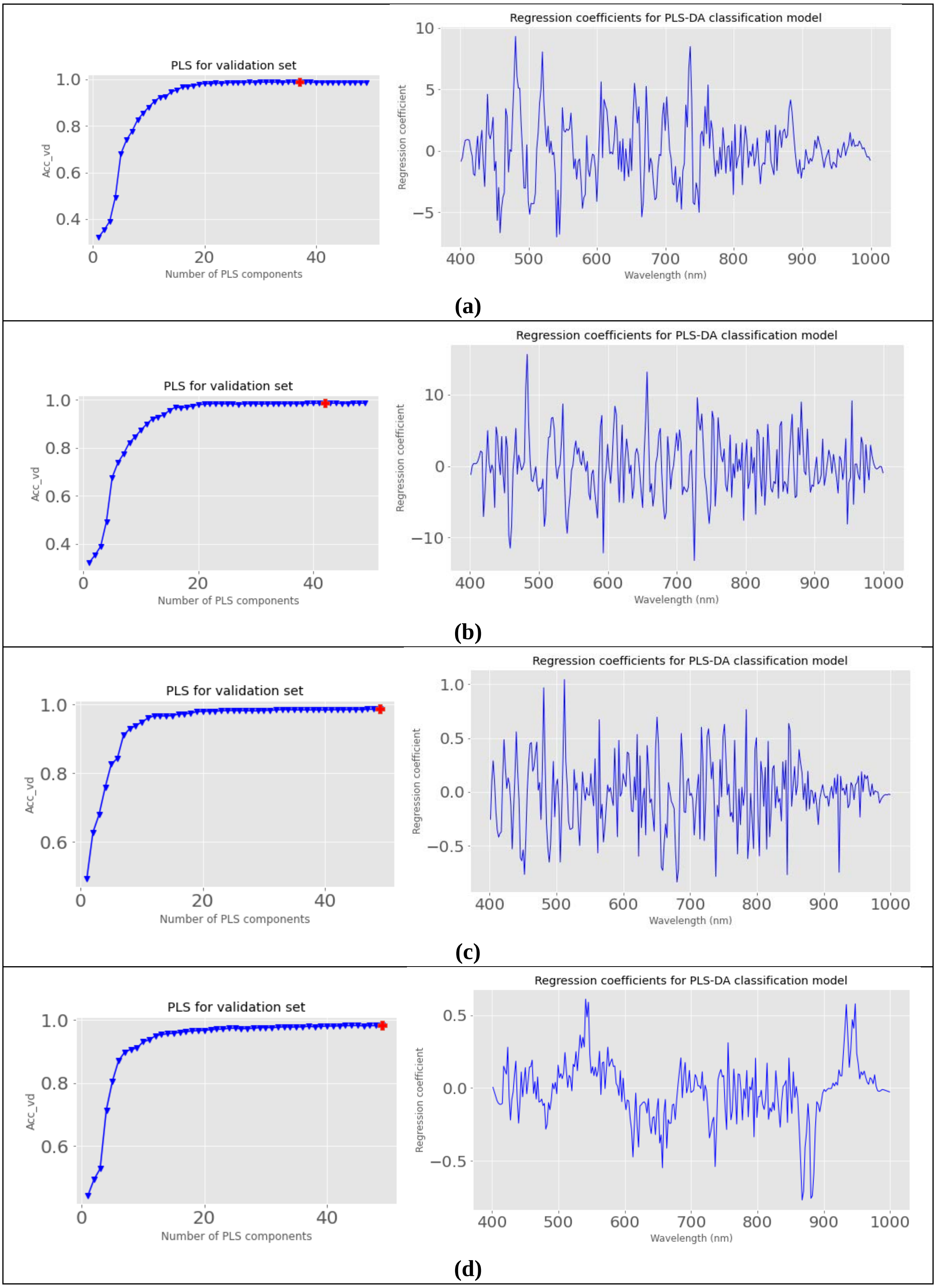


Figure B.4: The best result of PLS model on various pre-processed mean spectra data: (a) Raw/Without pre-processing; (b) SG; (c) SG-1st; (d) SG-2nd [Left – Optimisation of number of latent variables based on validation RMSE, Right – Regression coefficient of that resulting PLS model]



(a)

(b)

(c)

(d)

Figure B.5: The best result of PLS-DA model on various pre-processed mean spectra data: (a) Raw/Without pre-processing; (b) SG; (c) SG-1st; (d) SG-2nd [Left – Optimisation of number of latent variables based on validation RMSE, Right – Regression coefficient of that resulting PLS model]

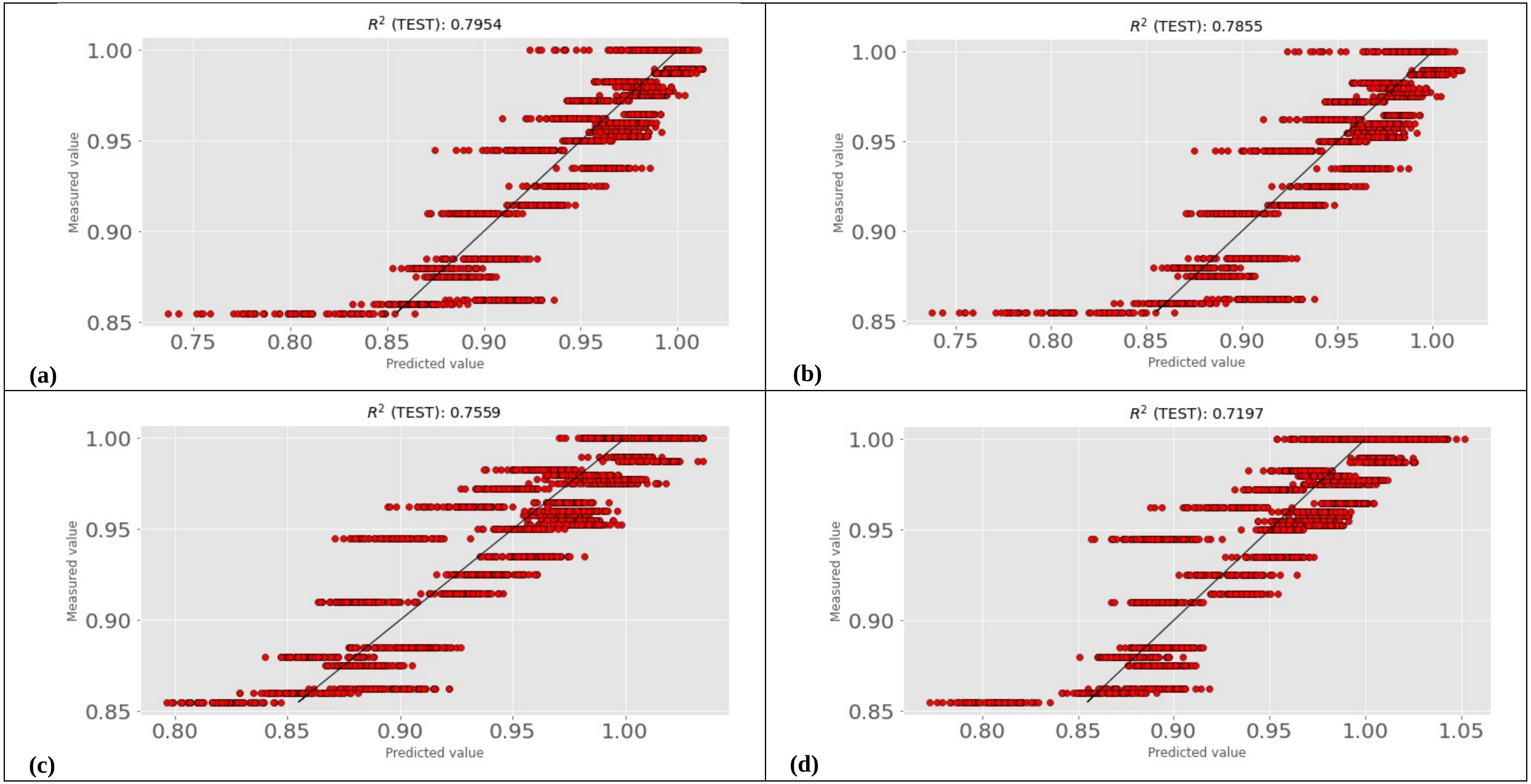


Figure B.6: Parity plots of SVR models for determination of authenticity of black pepper powder samples with different data pre-processing methods [(a): Raw/without pre-processing; (b): SG; (c): SG-1st; (d): SG-2nd]

Table B.1: Training and validation results of DL CNN models for determination of authenticity of black pepper powder samples

Indicator	Model Type	Pre-processing					Indicator	Model Type	Pre-processing				
		Raw	SG	SG-SNV	SG-1st	SG-2nd			Raw	SG	SG-SNV	SG-1st	SG-2nd
R^2 Training	1	0.9397	0.9541	0.9278	0.9557	0.9452	R^2 Validation	1	0.9405	0.9533	0.9268	0.9532	0.9436
	2	0.9493	0.9611	0.9499	0.9643	0.9219		2	0.9475	0.9599	0.9498	0.9636	0.9217
	3	0.9365	0.9730	0.9582	0.9664	0.9714		3	0.9376	0.9714	0.9590	0.9678	0.9701
	4	0.9473	0.9718	0.9491	0.9747	0.9766		4	0.9469	0.9710	0.9491	0.9735	0.9746
	5	0.9727	0.9565	0.9605	0.9718	-1.1675		5	0.9716	0.9576	0.9617	0.9706	-1.0943
	6	0.9775	0.9744	0.9741	0.9741	0.9592		6	0.9759	0.9736	0.9738	0.9736	0.9589
	7	0.9707	0.9763	0.9565	0.9607	0.9605		7	0.9696	0.9743	0.9566	0.9611	0.9594
	8	0.9733	0.9761	0.9649	0.9440	0.9546		8	0.9727	0.9745	0.9654	0.9431	0.9535
	9	0.9726	0.9699	0.9663	0.3713	0.5409		9	0.9720	0.9688	0.9676	0.4044	0.5241
	10	0.9741	0.9755	0.9738	-0.8937	-0.1000		10	0.9727	0.9744	0.9739	-0.8082	-0.1003
	11	0.9744	0.9750	0.9693	0.4931	0.9020		11	0.9720	0.9747	0.9690	0.4844	0.8896
	12	0.9717	0.9723	0.9724	0.9657	0.9019		12	0.9705	0.9722	0.9730	0.9656	0.9084
	13	0.9720	0.9712	0.9590	0.9589	-1.3221		13	0.9708	0.9696	0.9614	0.9584	-1.3232
	14	0.9711	0.9787	0.9537	0.9704	0.9674		14	0.9709	0.9782	0.9557	0.9681	0.9647
	15	0.9645	0.9731	0.9671	0.9610	0.9063		15	0.9657	0.9727	0.9673	0.9609	0.9005
	16	0.9721	0.9736	0.9673	0.9719	0.9576		16	0.9706	0.9728	0.9681	0.9711	0.9567
Indicator	Model Type	Pre-processing					Indicator	Model Type	Pre-processing				
		Raw	SG	SG-SNV	SG-1st	SG-2nd			Raw	SG	SG-SNV	SG-1st	SG-2nd
RMSE Training	1	0.0191	0.0167	0.0209	0.0164	0.0182	RMSE Validation	1	0.0190	0.0168	0.0211	0.0169	0.0185
	2	0.0175	0.0154	0.0174	0.0147	0.0218		2	0.0179	0.0156	0.0175	0.0149	0.0218
	3	0.0196	0.0128	0.0159	0.0143	0.0132		3	0.0195	0.0132	0.0158	0.0140	0.0135

	4	0.0179	0.0131	0.0176	0.0124	0.0119		**4**	0.0180	0.0133	0.0176	0.0127	0.0124
	5	0.0129	0.0162	0.0155	0.0131	0.1146		**5**	0.0131	0.0160	0.0153	0.0134	0.1128
	6	0.0117	0.0124	0.0125	0.0125	0.0157		**6**	0.0121	0.0127	0.0126	0.0127	0.0158
	7	0.0133	0.0120	0.0162	0.0154	0.0155		**7**	0.0136	0.0125	0.0163	0.0154	0.0157
	8	0.0127	0.0120	0.0146	0.0184	0.0166		**8**	0.0129	0.0124	0.0145	0.0186	0.0168
	9	0.0129	0.0135	0.0143	0.0617	0.0528		**9**	0.0130	0.0138	0.0140	0.0602	0.0538
	10	0.0125	0.0122	0.0126	0.1071	0.0817		**10**	0.0129	0.0125	0.0126	0.1048	0.0818
	11	0.0125	0.0123	0.0136	0.0554	0.0244		**11**	0.0131	0.0124	0.0137	0.0560	0.0259
	12	0.0131	0.0130	0.0129	0.0144	0.0244		**12**	0.0134	0.0130	0.0128	0.0145	0.0236
	13	0.0130	0.0132	0.0158	0.0158	0.1186		**13**	0.0133	0.0136	0.0153	0.0159	0.1188
	14	0.0132	0.0114	0.0168	0.0134	0.0141		**14**	0.0133	0.0115	0.0164	0.0139	0.0147
	15	0.0147	0.0128	0.0141	0.0154	0.0238		**15**	0.0144	0.0129	0.0141	0.0154	0.0246
	16	0.0130	0.0127	0.0141	0.0131	0.0160		**16**	0.0134	0.0129	0.0139	0.0133	0.0162
Indicator	**Model Type**	**Pre-processing**					**Indicator**	**Model Type**	**Pre-processing**				
		Raw	**SG**	**SG-SNV**	**SG-1st**	**SG-2nd**			**Raw**	**SG**	**SG-SNV**	**SG-1st**	**SG-2nd**
MAPE Training (%)	**1**	1.73	1.47	1.92	1.45	1.64	**MAPE Validation (%)**	**1**	1.73	1.49	1.93	1.50	1.66
	2	1.53	1.33	1.53	1.30	2.00		**2**	1.58	1.36	1.53	1.30	1.99
	3	1.75	1.10	1.39	1.24	1.13		**3**	1.75	1.15	1.39	1.23	1.16
	4	1.58	1.12	1.53	1.06	1.03		**4**	1.62	1.15	1.55	1.08	1.07
	5	1.12	1.44	1.37	1.12	9.37		**5**	1.14	1.43	1.35	1.15	9.17
	6	1.00	1.07	1.08	1.05	1.35		**6**	1.04	1.09	1.10	1.08	1.36
	7	1.18	1.03	1.43	1.36	1.35		**7**	1.20	1.08	1.46	1.37	1.37
	8	1.09	1.04	1.26	1.60	1.47		**8**	1.13	1.08	1.27	1.63	1.49
	9	1.10	1.18	1.23	5.61	4.59		**9**	1.12	1.21	1.21	5.46	4.66
	10	1.07	1.04	1.09	10.06	7.87		**10**	1.11	1.08	1.10	9.78	7.88
	11	1.08	1.05	1.18	4.47	2.06		**11**	1.14	1.06	1.19	4.50	2.17

	12	1.13	1.11	1.11	1.26	2.18		**12**	1.17	1.12	1.12	1.26	2.13
	13	1.12	1.12	1.38	1.38	10.79		**13**	1.15	1.17	1.36	1.39	10.81
	14	1.15	0.96	1.51	1.18	1.21		**14**	1.16	0.98	1.49	1.23	1.26
	15	1.28	1.09	1.23	1.34	2.09		**15**	1.27	1.11	1.23	1.35	2.18
	16	1.11	1.07	1.21	1.11	1.43		**16**	1.17	1.11	1.20	1.14	1.44
Indicator	**Model Type**	**Pre-processing**					**Indicator**	**Model Type**	**Pre-processing**				
		Raw	**SG**	**SG-SNV**	**SG-1st**	**SG-2nd**			**Raw**	**SG**	**SG-SNV**	**SG-1st**	**SG-2nd**
Accuracy Training (%)	**1**	99.96	99.98	99.96	99.98	99.89	**Accuracy Validation (%)**	**1**	99.96	100	99.96	100	99.71
	2	99.98	100	100	100	99.77		**2**	100	100	100	100	99.75
	3	99.98	100	99.98	100	100		**3**	100	100	100	100	100
	4	99.98	100	100	100	100		**4**	100	100	100	100	100
	5	100	100	100	99.98	90.81		**5**	100	100	100	100	91.48
	6	100	100	100	100	99.98		**6**	100	100	100	100	100
	7	100	100	100	100	100		**7**	100	100	100	100	100
	8	100	100	100	99.96	99.98		**8**	100	100	100	100	100
	9	100	100	100	99.84	92.57		**9**	100	100	100	99.92	92.18
	10	100	100	100	99.03	96.47		**10**	100	100	100	98.97	96.67
	11	100	100	100	99.54	99.95		**11**	100	100	100	99.59	100
	12	100	100	100	100	99.40		**12**	100	100	100	100	99.59
	13	100	100	100	100	94.13		**13**	100	100	100	100	94.32
	14	100	100	100	100	100		**14**	100	100	100	100	100
	15	100	100	100	100	99.93		**15**	100	100	100	100	100
	16	100	100	100	100	100		**16**	100	100	100	100	100

Table B.2: Training and validation results of DL SAE models for determination of authenticity of black pepper powder samples

Indicator	Model Type	Pre-processing					Indicator	Model Type	Pre-processing				
		Raw	SG	SG-SNV	SG-1st	SG-2nd			Raw	SG	SG-SNV	SG-1st	SG-2nd
R^2 Training	1	0.9594	0.9683	0.9615	0.9709	0.9446	R^2 Validation	1	0.9583	0.9661	0.9607	0.9706	0.9444
	2	0.9622	0.9663	0.9666	0.9680	0.9563		2	0.9606	0.9636	0.9643	0.9676	0.9552
	3	0.9660	0.9704	0.9579	0.9708	0.9553		3	0.9645	0.9678	0.9569	0.9695	0.9558
	4	0.9607	0.9661	0.9666	0.9747	0.9208		4	0.9587	0.9641	0.9637	0.9729	0.9233
	5	0.9624	0.9647	0.9680	0.9748	0.9524		5	0.9599	0.9646	0.9671	0.9723	0.9526
Indicator	Model Type	Pre-processing					Indicator	Model Type	Pre-processing				
		Raw	SG	SG-SNV	SG-1st	SG-2nd			Raw	SG	SG-SNV	SG-1st	SG-2nd
RMSE Training	1	0.0157	0.0139	0.0153	0.0133	0.0183	RMSE Validation	1	0.0159	0.0144	0.0155	0.0134	0.0184
	2	0.0151	0.0143	0.0142	0.0139	0.0163		2	0.0155	0.0149	0.0147	0.0140	0.0165
	3	0.0144	0.0134	0.0160	0.0133	0.0165		3	0.0147	0.0140	0.0162	0.0136	0.0164
	4	0.0154	0.0143	0.0142	0.0124	0.0219		4	0.0159	0.0148	0.0148	0.0128	0.0216
	5	0.0151	0.0146	0.0139	0.0124	0.0170		5	0.0156	0.0147	0.0141	0.0130	0.0170
Indicator	Model Type	Pre-processing					Indicator	Model Type	Pre-processing				
		Raw	SG	SG-SNV	SG-1st	SG-2nd			Raw	SG	SG-SNV	SG-1st	SG-2nd
MAPE Training (%)	1	1.38	1.20	1.33	1.14	1.63	MAPE Validation (%)	1	1.41	1.23	1.35	1.15	1.65
	2	1.34	1.24	1.24	1.19	1.41		2	1.38	1.29	1.27	1.22	1.44
	3	1.26	1.15	1.41	1.14	1.45		3	1.30	1.22	1.44	1.18	1.47
	4	1.33	1.23	1.24	1.06	1.96		4	1.39	1.29	1.28	1.10	1.96
	5	1.31	1.27	1.21	1.05	1.56		5	1.36	1.30	1.21	1.12	1.55
Indicator		Pre-processing					Indicator		Pre-processing				

	Model Type	Raw	SG	SG-SNV	SG-1st	SG-2nd		Model Type	Raw	SG	SG-SNV	SG-1st	SG-2nd
Accuracy Training (%)	1	100	99.98	99.98	100	100	Accuracy Validation (%)	1	100	100	99.96	100	100
	2	100	100	99.98	100	100		2	100	100	100	100	100
	3	100	100	100	100	100		3	100	100	100	100	100
	4	100	100	100	100	100		4	100	100	100	100	100
	5	99.98	100	100	100	99.98		5	100	100	100	100	100

APPENDIX C

SUPPLEMENTAL FIGURES AND TABLES FROM CHAPTER 5

Table C.1: Training and validation results of PLS models for prediction of chemical and biological analytical properties

Indicator	Pre-processing	Chemical						Microbiological	
		1	2	3	4	5	6	7	8
R^2 Training	Raw	0.6745	0.6851	0.6268	0.2814	0.6839	0.4676	0.4010	0.2909
	SG	0.6750	0.6560	0.6110	0.2872	0.6751	0.4463	0.3546	0.2583
	SG-SNV	0.6751	0.6622	0.6076	0.2799	0.6716	0.4300	0.3697	0.2295
	SG-1st	0.6808	0.6637	0.6079	0.2853	0.6643	0.4423	0.3753	0.2753
	SG-2nd	0.6722	0.6819	0.6212	0.2904	0.6853	0.4549	0.4019	0.2850
RMSE Training	Raw	0.9398	0.6954	0.1165	0.4451	0.2528	0.2271	99471.54	21302.01
	SG	0.9392	0.7269	0.1190	0.4433	0.2563	0.2317	103256.63	21785.85
	SG-SNV	0.9390	0.7202	0.1195	0.4455	0.2577	0.2350	102036.89	22203.78
	SG-1st	0.9307	0.7187	0.1195	0.4439	0.2606	0.2325	101585.66	21534.83
	SG-2nd	0.9432	0.6989	0.1174	0.4422	0.2522	0.2298	99394.42	21390.37
MAPE Training (%)	Raw	8.82	26.91	24.12	3.91	17.92	16.17	168.59	100.42
	SG	8.80	29.01	24.75	3.91	18.18	16.17	154.13	102.36
	SG-SNV	8.74	28.79	24.72	3.94	18.07	16.60	166.42	108.69
	SG-1st	8.76	28.30	24.60	3.92	18.37	16.42	163.51	101.18
	SG-2nd	8.85	27.27	24.20	3.90	17.68	16.38	174.43	101.11
R^2 Validation	Raw	0.6577	0.6357	0.5655	0.2422	0.6468	0.4128	0.3173	0.2059
	SG	0.6587	0.6360	0.5686	0.2481	0.6423	0.4198	0.3214	0.2148
	SG-SNV	0.6607	0.6342	0.5683	0.2433	0.6372	0.4053	0.3265	0.1744
	SG-1st	0.6579	0.6345	0.5605	0.2386	0.6343	0.4093	0.3214	0.2096
	SG-2nd	0.6507	0.6220	0.5513	0.2279	0.6316	0.4029	0.3103	0.1634
RMSE Validation	Raw	0.9638	0.7480	0.1258	0.4570	0.2673	0.2386	106197.69	22542.23
	SG	0.9624	0.7477	0.1253	0.4553	0.2690	0.2371	105877.46	22415.56
	SG-SNV	0.9595	0.7495	0.1253	0.4567	0.2709	0.2401	105480.14	22985.34
	SG-1st	0.9635	0.7492	0.1265	0.4581	0.2719	0.2393	105874.99	22489.53
	SG-2nd	0.9736	0.7620	0.1278	0.4613	0.2729	0.2406	106740.49	23137.59
MAPE Validation (%)	Raw	9.04	29.15	26.05	4.02	18.92	16.98	181.45	105.84
	SG	9.02	29.88	26.08	4.02	19.03	16.57	160.13	105.19
	SG-SNV	8.93	29.98	25.97	4.04	18.96	16.98	172.35	112.30
	SG-1st	9.06	29.59	26.05	4.05	19.14	16.91	171.85	105.28
	SG-2nd	9.12	29.91	26.33	4.07	19.12	17.16	190.72	108.68

Table C.2: Training and validation results of SVR models for prediction of chemical and biological analytical properties

Indicator	Pre-processing	Chemical						Microbiological	
		1	2	3	4	5	6	7	8
R^2 Training	Raw	0.7371	0.8217	0.8214	0.3473	0.7802	0.5726	0.0221	0.0225
	SG	0.7338	0.8193	0.8109	0.3313	0.7703	0.5579	0.0221	0.0225
	SG-SNV	0.8064	0.7941	0.9031	0.5848	0.8735	0.7094	0.0093	0.0149
	SG-1st	0.9960	0.9954	0.9198	0.9994	0.9995	0.9534	0.0268	0.0218
	SG-2nd	0.7969	0.7601	0.8926	0.6398	0.8544	0.7654	0.0142	0.0188
RMSE Training	Raw	0.8446	0.5233	0.0806	0.4241	0.2108	0.2035	127094.61	25009.87
	SG	0.8499	0.5268	0.0830	0.4293	0.2155	0.207	127095.43	25009.62
	SG-SNV	0.7248	0.5624	0.0594	0.3383	0.1599	0.1678	127924.14	25107.36
	SG-1st	0.1038	0.0838	0.0540	0.0132	0.0098	0.0672	126791.72	25019.24

	SG-2nd	0.7424	0.607	0.0625	0.3151	0.1716	0.1508	127606.39	25057.09
MAPE Training (%)	**Raw**	6.65	20.59	12.35	2.89	11.45	9.01	273.40	141.71
	SG	6.74	20.73	13.01	2.95	11.86	9.36	273.38	141.71
	SG-SNV	5.01	23.51	8.05	2.06	7.26	6.32	294.45	142.56
	SG-1st	0.25	1.89	8.22	0.12	0.82	2.16	280.36	138.25
	SG-2nd	5.18	25.54	8.93	1.97	7.99	5.77	283.19	136.12
R^2 Validation	**Raw**	0.6943	0.6969	0.7626	0.2215	0.7182	0.5003	0.0203	0.0154
	SG	0.6926	0.6959	0.7601	0.2110	0.7123	0.4936	0.0203	0.0155
	SG-SNV	0.7245	0.7292	0.8003	0.3474	0.7573	0.5567	0.0075	0.0100
	SG-1st	0.8139	0.8179	0.8169	0.5762	0.8628	0.6776	0.0247	0.0169
	SG-2nd	0.7224	0.6846	0.6948	0.4385	0.7064	0.5815	0.0118	0.0162
RMSE Validation	**Raw**	0.9109	0.6823	0.0929	0.4632	0.2387	0.2201	127211.67	25099.93
	SG	0.9133	0.6835	0.0934	0.4663	0.2412	0.2215	127212.18	25099.69
	SG-SNV	0.8647	0.6449	0.0852	0.4241	0.2215	0.2073	128044.71	25169.60
	SG-1st	0.7106	0.5288	0.0816	0.3418	0.1666	0.1768	126930.82	25081.02
	SG-2nd	0.8679	0.6959	0.1054	0.3934	0.2437	0.2014	127765.18	25090.61
MAPE Validation (%)	**Raw**	7.65	27.08	15.70	3.39	13.95	10.96	274.17	142.04
	SG	7.70	27.12	15.82	3.41	14.12	11.09	274.16	142.04
	SG-SNV	6.98	26.99	14.43	3.08	12.23	10.64	294.95	142.88
	SG-1st	5.90	18.05	15.41	2.63	10.04	10.77	282.22	138.80
	SG-2nd	7.15	29.78	20.45	3.06	14.90	11.73	284.01	136.66
# Support Vectors	**Raw**	3060	3022	2559	2946	2861	2709	3113	3113
	SG	3045	3013	2522	2935	2880	2721	3113	3113
	SG-SNV	3035	3020	2526	2910	2856	2755	3113	3113
	SG-1st	3049	3031	2586	2887	2845	2838	3113	3113
	SG-2nd	3057	3062	2750	2974	2977	2905	3112	3113

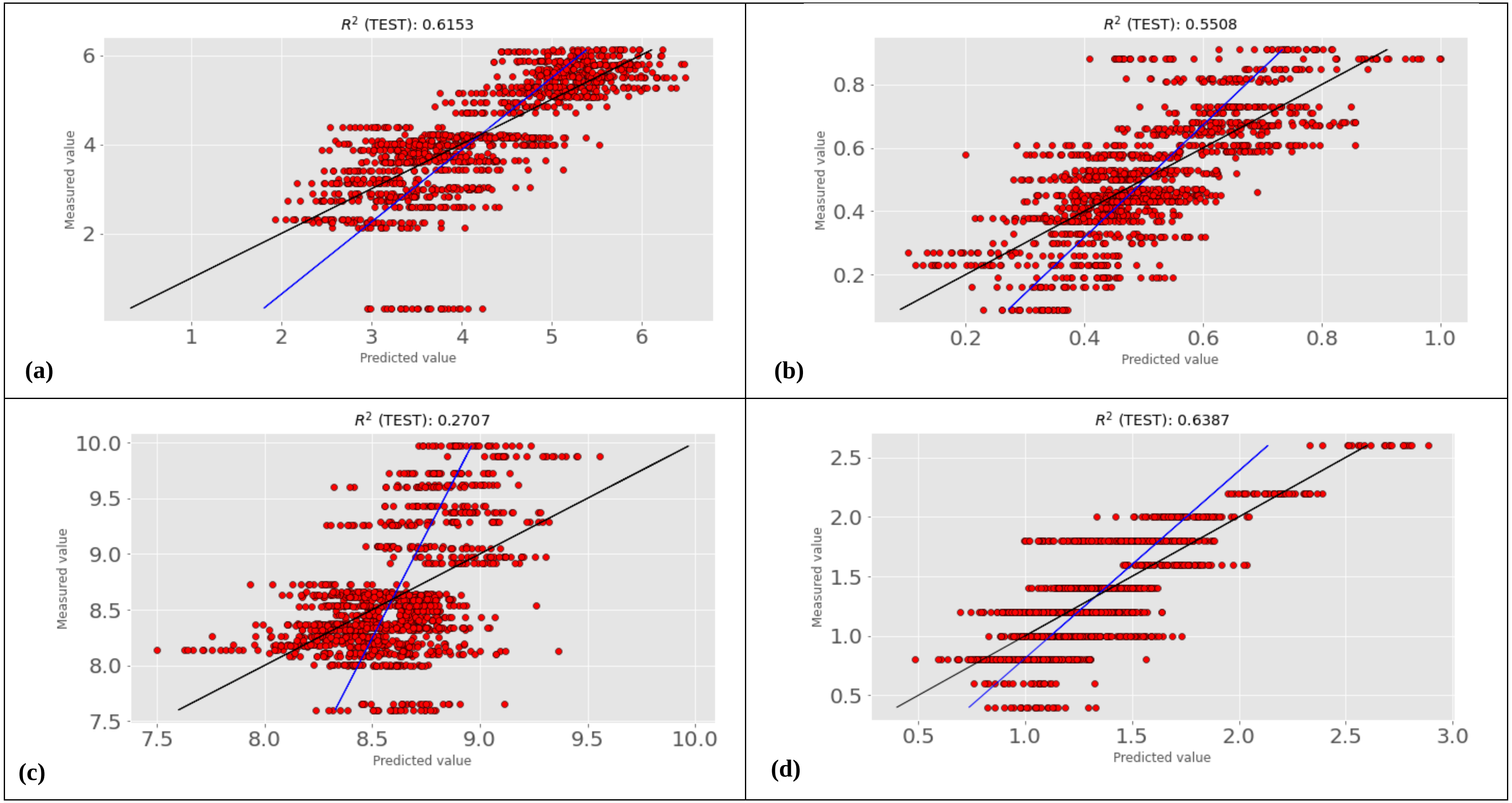
Table C.3: Training and validation results of DL CNN models with different model types for prediction of chemical and biological analytical properties

Indicator	**Pre-processing**	**Chemical**						**Microbiological**	
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
R^2 Training	**Raw**	0.6648	0.6878	0.4977	0.0151	0.6093	0.3678	0.4202	0.2345
	SG	0.7262	0.7063	0.5869	0.1888	0.6739	0.4742	0.4885	0.1795
	SG-SNV	0.7135	0.7152	0.6257	0.0029	0.6805	0.4336	0.4725	0.1428
	SG-1st	0.7664	0.7472	0.6019	0.1484	0.6765	0.4837	0.4695	0.2062
	SG-2nd	0.7492	0.7467	0.6576	0.2401	0.6895	0.4604	0.5529	0.1603
RMSE Training	**Raw**	0.9412	0.6918	0.1328	0.5178	0.2797	0.2436	99139.87	22178.17
	SG	0.8507	0.6710	0.1204	0.4699	0.2555	0.2222	93113.68	22961.66
	SG-SNV	0.8702	0.6607	0.1146	0.5210	0.2529	0.2306	94559.78	23468.38
	SG-1st	0.7857	0.6225	0.1182	0.4815	0.2545	0.2202	94831.71	22584.25
	SG-2nd	0.8141	0.6231	0.1096	0.4548	0.2494	0.2251	87054.92	23228.42
MAPE Training (%)	**Raw**	8.81	26.47	26.27	4.85	18.56	17.57	132.71	113.88
	SG	7.98	23.90	21.50	3.84	15.60	15.37	126.67	126.88
	SG-SNV	7.73	24.22	21.38	4.29	15.19	16.32	105.83	129.75
	SG-1st	7.14	20.54	23.11	4.11	16.45	15.30	110.07	109.90
	SG-2nd	7.17	22.21	19.39	3.96	16.21	17.26	100.16	121.78
R^2 Validation	**Raw**	0.6877	0.6924	0.5054	-0.0621	0.6125	0.3226	0.3927	0.2120
	SG	0.7132	0.6755	0.5757	0.0687	0.6687	0.4114	0.4235	0.1453
	SG-SNV	0.6686	0.6528	0.6080	-0.0703	0.6609	0.4133	0.3821	0.0938
	SG-1st	0.7057	0.6752	0.5550	0.0972	0.6782	0.4074	0.4284	0.1742
	SG-2nd	0.7022	0.7148	0.6431	0.0341	0.6849	0.4224	0.4476	0.1264
RMSE Validation	**Raw**	0.9462	0.6888	0.1396	0.5486	0.2830	0.2651	96820.09	22336.58
	SG	0.9069	0.7075	0.1293	0.5137	0.2617	0.2471	94335.92	23263.18
	SG-SNV	0.9748	0.7319	0.1243	0.5507	0.2648	0.2467	97664.68	23953.58

	SG-1st	0.9187	0.7078	0.1324	0.5058	0.2580	0.2480	93929.21	22866.77
	SG-2nd	0.924	0.6633	0.1186	0.5232	0.2553	0.2448	92340.42	23519.13
MAPE Validation (%)	Raw	8.60	28.64	28.04	5.14	18.83	17.71	140.84	120.22
	SG	8.48	26.78	24.14	4.19	16.46	15.86	141.69	138.00
	SG-SNV	8.65	28.78	23.73	4.47	16.56	16.36	117.86	140.67
	SG-1st	8.19	24.72	25.74	4.28	17.14	15.88	109.87	119.53
	SG-2nd	8.33	22.72	22.53	4.49	17.32	17.83	108.88	131.87

Table C.4: Training and validation results of DL SAE models with different model types for prediction of chemical and biological analytical properties

Indicator	Pre-processing	Chemical						Microbiological	
		1	2	3	4	5	6	7	8
R^2 Training	Raw	0.5404	0.5551	0.4600	-0.4474	0.4910	0.2239	0.4392	-0.5255
	SG	-63.234	-4.7152	-7.4684	-1388.15	-5.2296	-4.0925	-9.9866	-30.128
	SG-SNV	0.6197	0.5827	0.5008	0.1880	0.5681	0.3778	0.4877	-0.4517
	SG-1st	0.6099	0.5849	0.4944	0.0205	0.5965	0.3829	0.6795	-0.0869
	SG-2nd	-0.4983	-0.1435	0.0366	-0.0270	-0.2153	-0.0008	0.2908	-0.0532
RMSE Training	Raw	1.1022	0.8258	0.1377	0.6277	0.3193	0.2700	97502.10	31308.64
	SG	13.029	2.9598	0.5452	19.448	1.1169	0.6915	431560.91	141425.20
	SG-SNV	1.0026	0.7998	0.1324	0.4702	0.2941	0.2417	93190.91	30541.50
	SG-1st	1.0153	0.7977	0.1332	0.5164	0.2843	0.2407	73708.49	26426.70
	SG-2nd	1.9899	1.3239	0.1839	0.5288	0.4933	0.3065	109650.11	26014.04
MAPE Training (%)	Raw	11.02	31.49	25.07	5.66	21.65	21.78	158.52	132.21
	SG	161.91	90.13	100.00	223.81	65.08	53.62	1597.91	542.76
	SG-SNV	9.26	30.50	28.87	4.01	20.42	16.75	190.52	131.03
	SG-1st	9.24	29.27	25.47	4.68	18.05	15.08	132.82	155.11
	SG-2nd	23.27	52.06	43.23	4.94	38.67	23.03	169.20	137.58
R^2 Validation	Raw	0.5805	0.5480	0.4718	-0.3692	0.5299	0.2258	0.4108	-0.4515
	SG	-57.052	-4.4974	-6.7210	-1345.96	-4.8770	-4.1484	-11.319	-30.125
	SG-SNV	0.6149	0.5694	0.4993	0.0870	0.5647	0.3478	0.4266	-0.4196
	SG-1st	0.5969	0.5740	0.4869	0.0077	0.5813	0.2660	0.5926	-0.1337
	SG-2nd	-0.3640	-0.1551	0.0782	-0.0349	-0.3106	-0.0007	0.2882	-0.0882
RMSE Validation	Raw	1.0967	0.8350	0.1443	0.6229	0.3118	0.2834	95362.44	30315.85
	SG	12.902	2.9121	0.5516	19.537	1.1023	0.7309	436071.94	140384.29
	SG-SNV	1.0507	0.8150	0.1405	0.5087	0.3000	0.2601	94082.27	29980.82
	SG-1st	1.0751	0.8107	0.1422	0.5303	0.2942	0.2760	79304.12	26792.67
	SG-2nd	1.9776	1.3348	0.1906	0.5415	0.5206	0.3222	104817.92	26249.69
MAPE Validation (%)	Raw	10.77	34.25	27.57	5.79	21.96	22.04	167.96	135.44
	SG	158.43	92.60	100.00	225.35	63.53	56.12	1949.06	567.36
	SG-SNV	9.65	32.88	31.14	4.37	21.55	16.88	226.32	135.64
	SG-1st	9.76	31.10	27.63	4.84	18.96	16.22	136.09	166.58
	SG-2nd	22.77	55.53	46.97	5.14	40.23	23.64	201.22	147.52


R^2 (TEST): 0.6153
Measured value
Predicted value
6
4
2
1
2
3
4
5
6
(a)
R^2 (TEST): 0.5508
Measured value
Predicted value
0.8
0.6
0.4
0.2
0.2
0.4
0.6
0.8
1.0
(b)
R^2 (TEST): 0.2707
Measured value
Predicted value
10.0
9.5
9.0
8.5
8.0
7.5
7.5
8.0
8.5
9.0
9.5
10.0
(c)
R^2 (TEST): 0.6387
Measured value
Predicted value
2.5
2.0
1.5
1.0
0.5
0.5
1.0
1.5
2.0
2.5
3.0
(d)

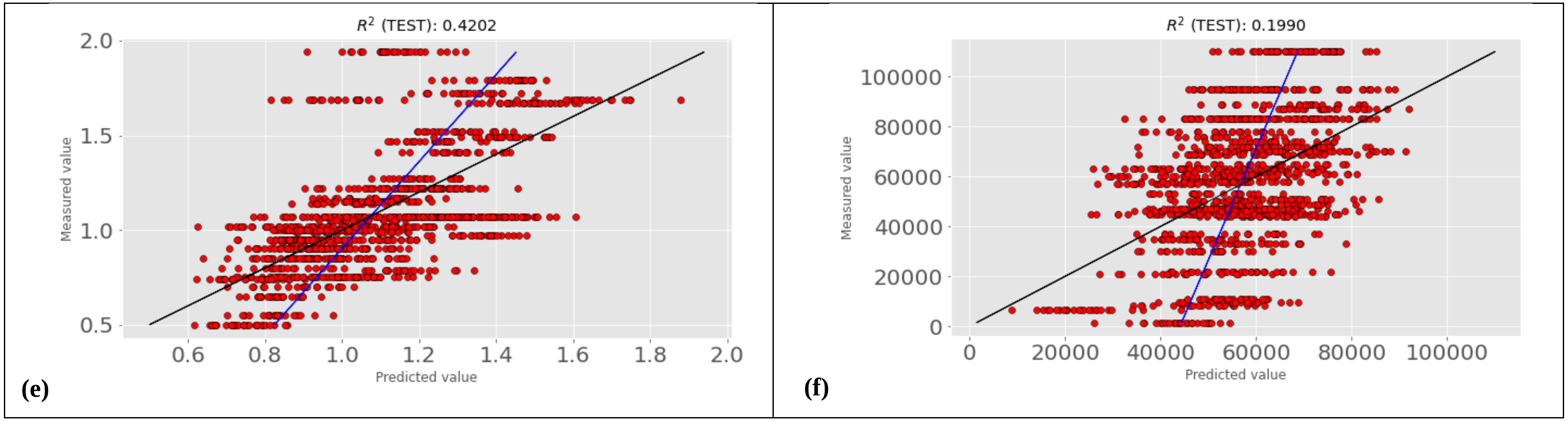


Figure C.1: Parity plots of best PLS models for different chemical and microbiological analytical properties - (a) ash content, (b) acid insoluble ash content, (c) non-volatile ether extract, (d) volatile oil content, (e) lead content and (f) total yeast count

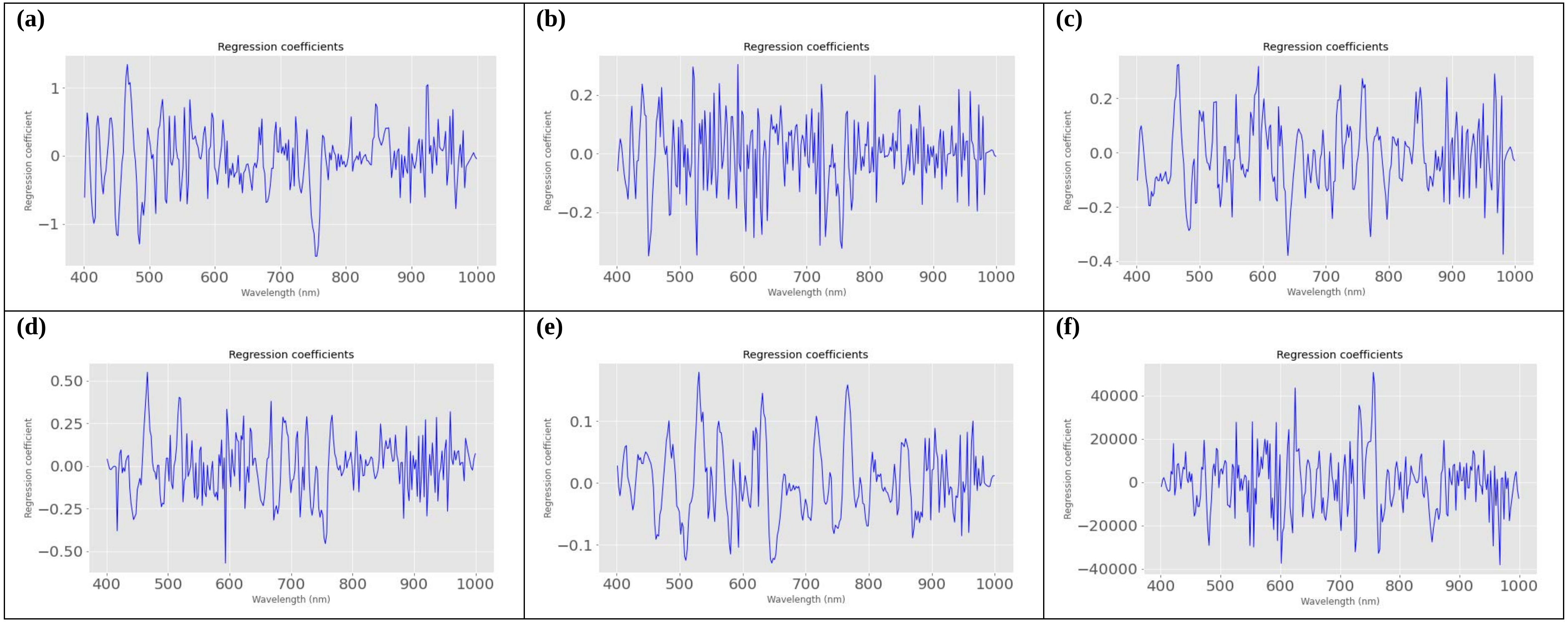
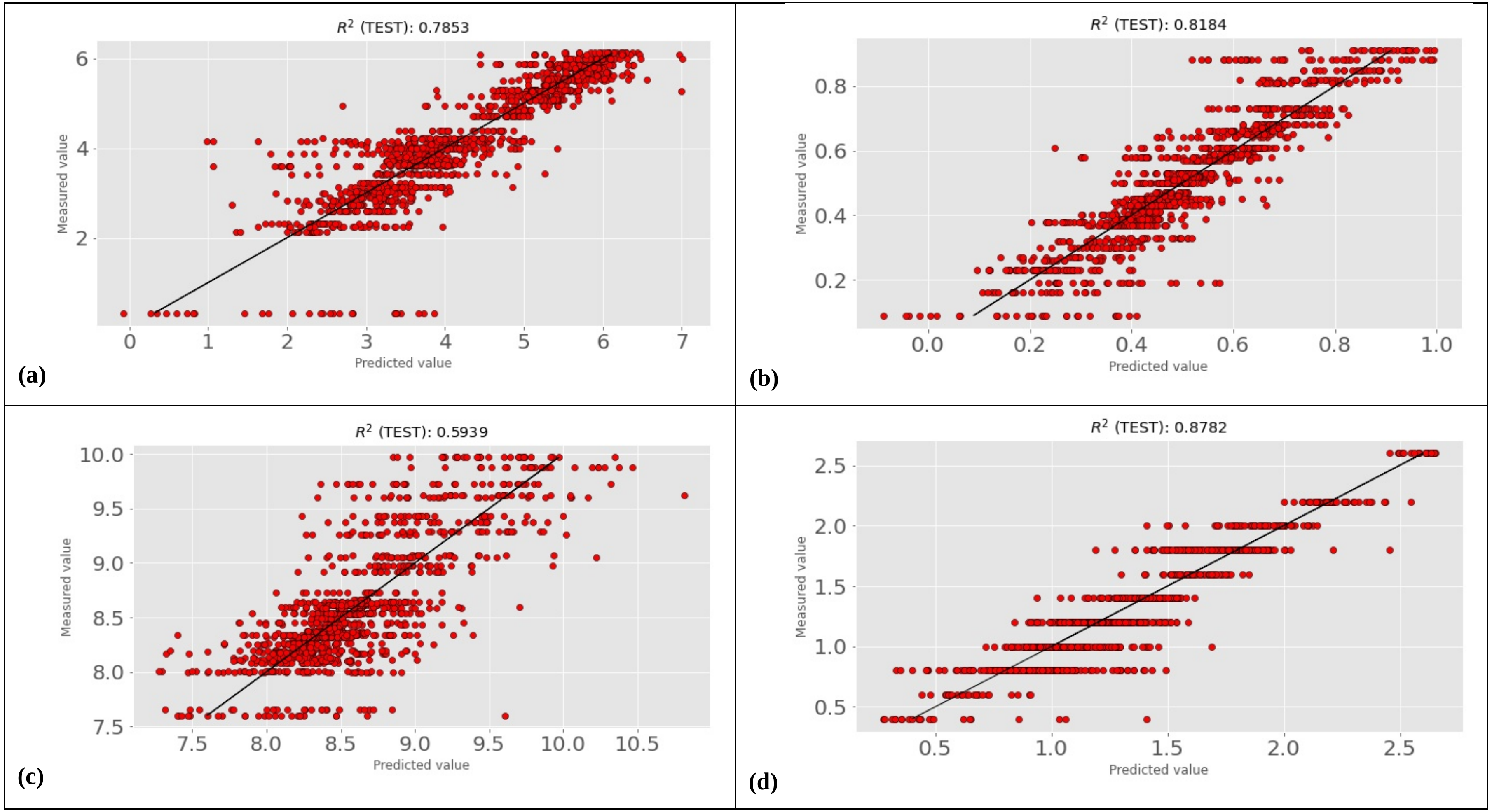


Figure C.2: Plots of regression coefficients of PLS models for prediction of chemical and biological analytical properties, (a) ash content, (b) acid insoluble ash content, (c) non-volatile ether extract, (d) volatile oil content, (e) lead content and (f) total yeast count


R^2 (TEST): 0.7853
Measured value
Predicted value
0
1
2
3
4
5
6
7
2
4
6
(a)
R^2 (TEST): 0.8184
Measured value
Predicted value
0.0
0.2
0.4
0.6
0.8
1.0
0.2
0.4
0.6
0.8
(b)
R^2 (TEST): 0.5939
Measured value
Predicted value
7.5
8.0
8.5
9.0
9.5
10.0
10.5
7.5
8.0
8.5
9.0
9.5
10.0
(c)
R^2 (TEST): 0.8782
Measured value
Predicted value
0.5
1.0
1.5
2.0
2.5
0.5
1.0
1.5
2.0
2.5
(d)

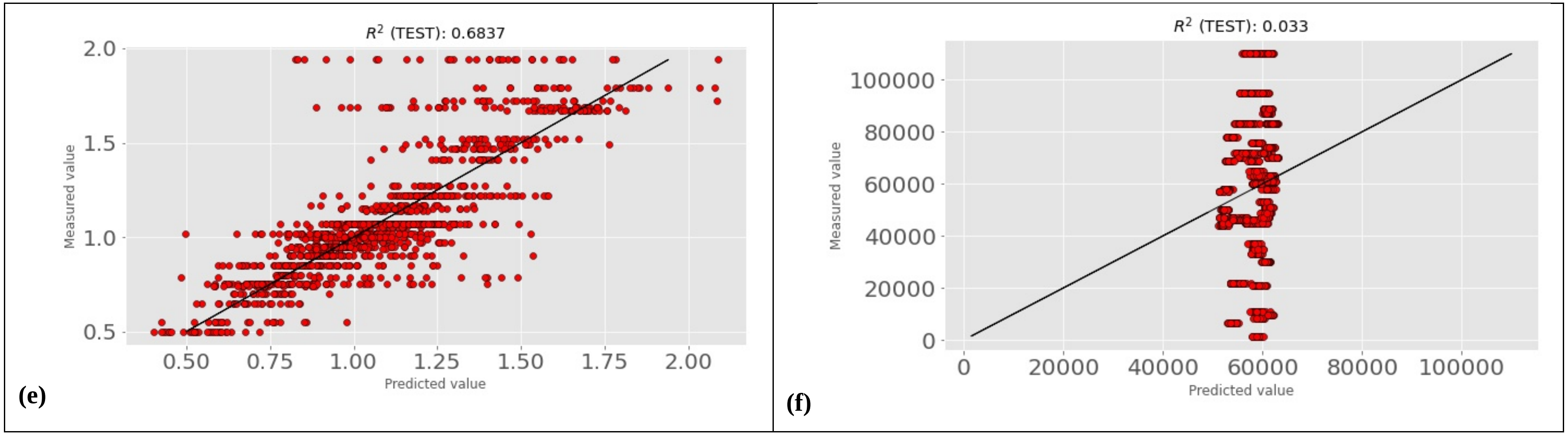
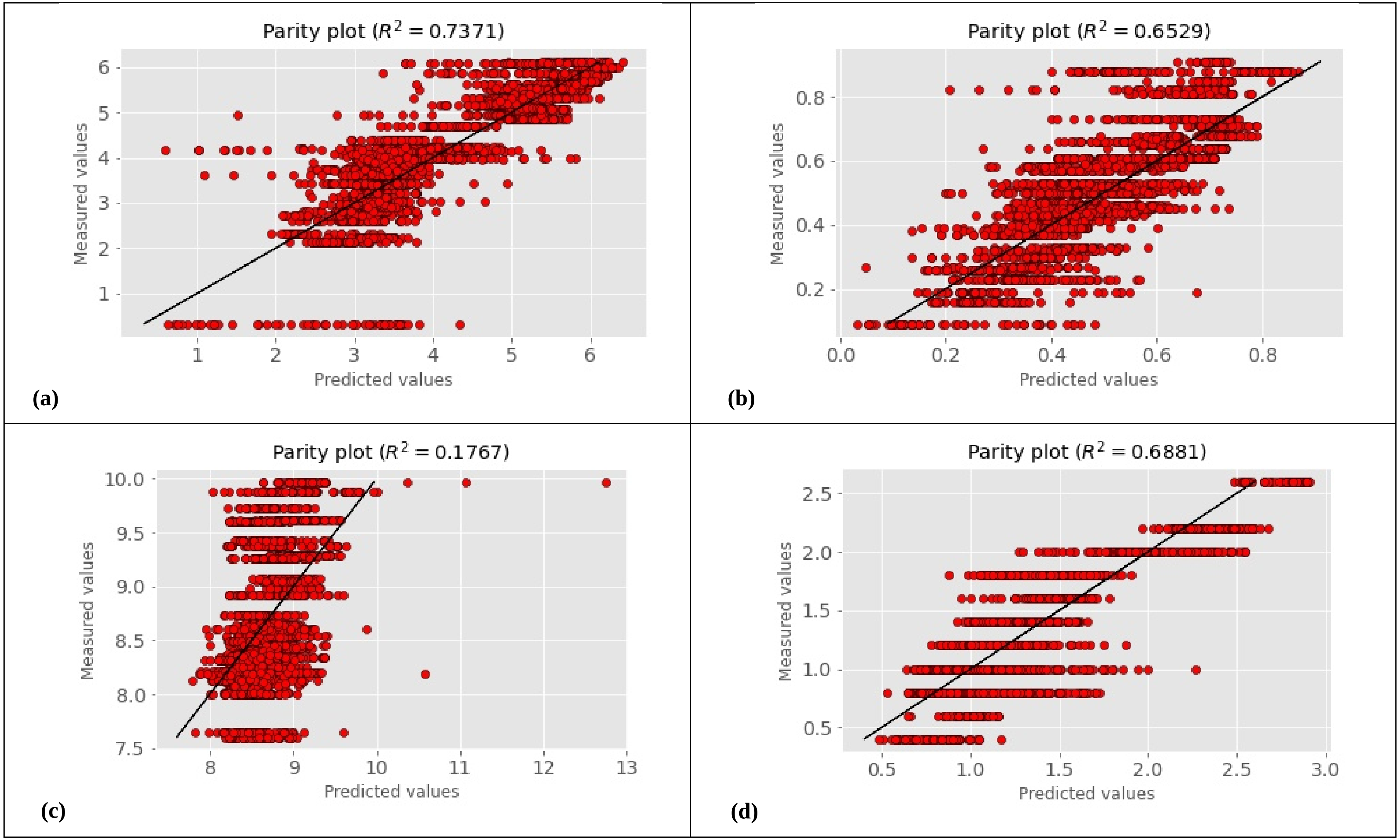


Figure C.3: Parity plots of best SVR models for different chemical and microbiological analytical properties - (a) ash content, (b) acid insoluble ash content, (c) non-volatile ether extract, (d) volatile oil content, (e) lead content and (f) total yeast count


Parity plot ($R^2 = 0.7371$)
Measured values
Predicted values
1
2
3
4
5
6
(a)
Parity plot ($R^2 = 0.6529$)
Measured values
Predicted values
0.0
0.2
0.4
0.6
0.8
(b)
Parity plot ($R^2 = 0.1767$)
Measured values
Predicted values
7.5
8.0
8.5
9.0
9.5
10.0
8
9
10
11
12
13
(c)
Parity plot ($R^2 = 0.6881$)
Measured values
Predicted values
0.5
1.0
1.5
2.0
2.5
3.0
(d)

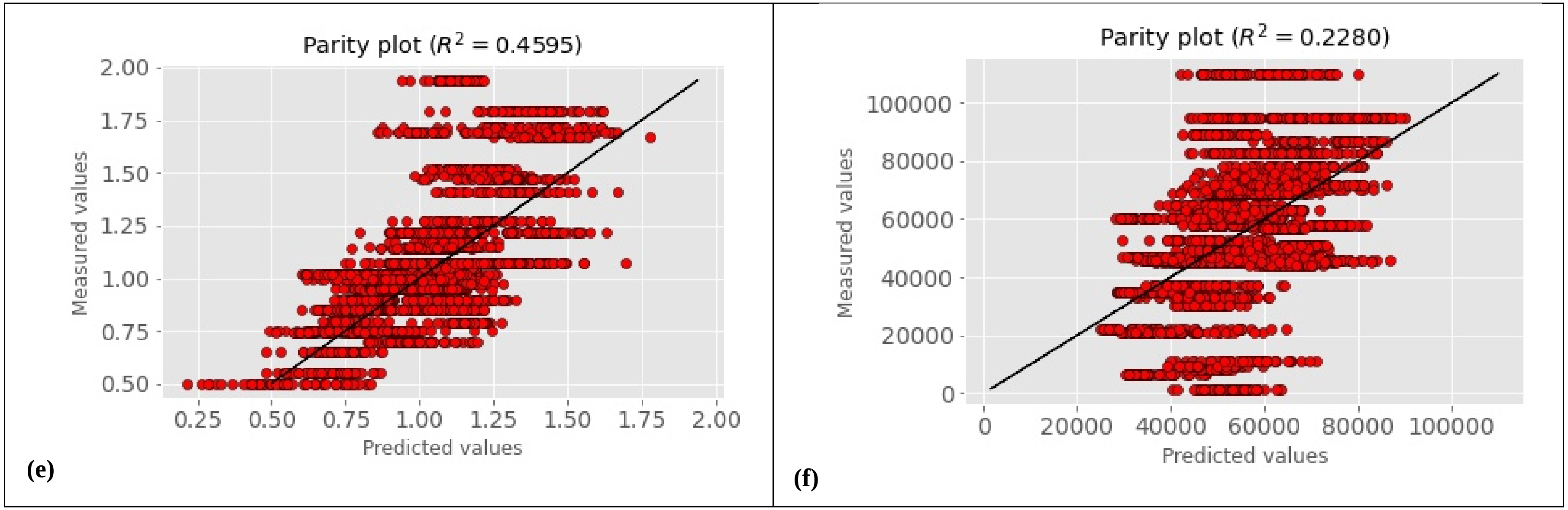
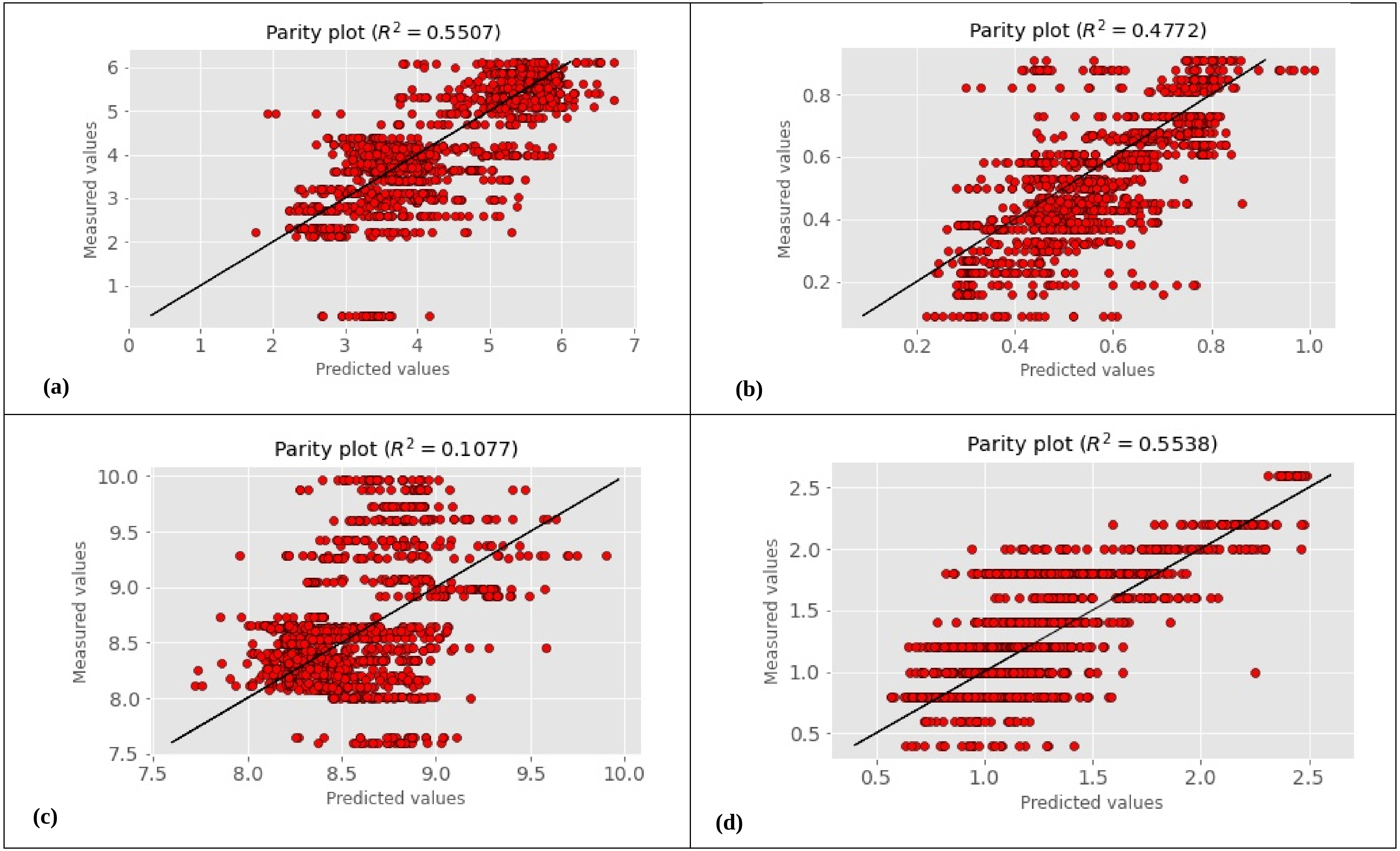


Figure C.4: Parity plots of best DL CNN models for different chemical and microbiological analytical properties - (a) ash content, (b) acid insoluble ash content, (c) non-volatile ether extract, (d) volatile oil content, (e) lead content and (f) total yeast count


Parity plot ($R^2 = 0.5507$)
Measured values
Predicted values
0
1
2
3
4
5
6
7
(a)
Parity plot ($R^2 = 0.4772$)
Measured values
Predicted values
0.2
0.4
0.6
0.8
1.0
(b)
Parity plot ($R^2 = 0.1077$)
Measured values
Predicted values
7.5
8.0
8.5
9.0
9.5
10.0
(c)
Parity plot ($R^2 = 0.5538$)
Measured values
Predicted values
0.5
1.0
1.5
2.0
2.5
(d)

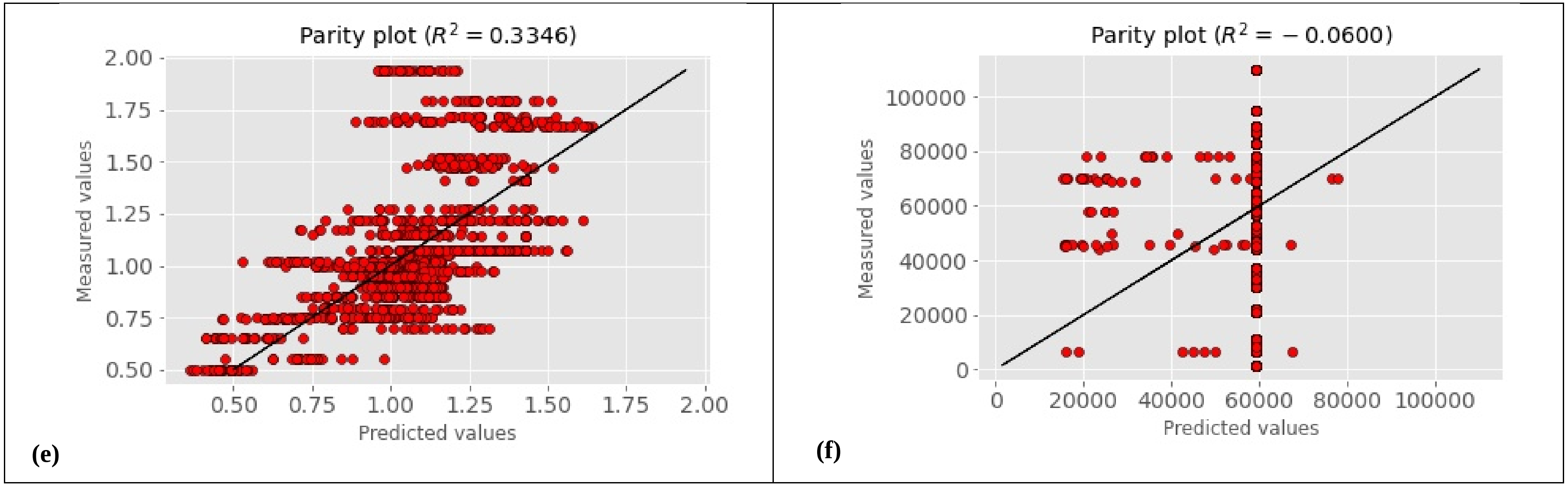


Figure C.5: Parity plots of best DL SAE models for different chemical and microbiological analytical properties - (a) ash content, (b) acid insoluble ash content, (c) non-volatile ether extract, (d) volatile oil content, (e) lead content and (f) total yeast count