DETECTION OF ADULTERATION AND ITS CLASSIFICATION BASED ON THE GEOGRAPHICAL AREA OF COCONUT MILK USING A COMBINATION OF NIR SPECTROSCOPY, MACHINE LEARNING AND DEEP LEARNING APPROACH

AGUSTAMI SITORUS

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF ENGINEERING IN
FOOD AND AGRICULTURAL INTELLIGENCE ENGINEERING
SCHOOL OF ENGINEERING
KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG
2024

KMITL-2024-EN-D-108-218

Thesis title	:	Detection of adulteration and its classification based on the geographical area of coconut milk using a combination of NIR spectroscopy, machine learning and deep learning approach
Student name	:	Agustami Sitorus
Student ID.	:	64601001
Degree	:	Doctor of Engineering
Program	:	Food and Agricultural Intelligence Engineering
Year	:	2024
Thesis advisor	:	Asst. Prof. Dr. Ravipat Lapcharoensuk

ABSTRACT

In this thesis, the research and development of chemometric tools based on machine learning (ML) and deep learning (DL) for near-infrared spectroscopy (NIRs) techniques to detect coconut milk adulteration and classify it based on its geographical area of origin were demonstrated. It is divided into 2 supporting section as a General Introduction (Chapter 1), and Conclusions, Recommendations, and Future Works (Chapter 8); 5 main sections (Chapters 2 to 6); and 1 section to challenge our model with a cross-over unknown dataset (Chapter 7). Chapter 1 briefly outlines the background of the samples, chemical content analysis of coconut milk, NIR instruments, and analysis tools used in this work. It also provides research problems, limitations of the research, and objectives of this thesis. Chapter 2 overviews the previous research paper on applying NIRs and IRs to detect and discriminate against the adulteration of food and agro-products based on recent research. Key findings from this study find that NIRs and IRs are non-destructive, rapid, simple-preparation, analytical rapidity, and straightforward methods for predicting adulteration in food and agro-products, so it is suitable for large-scale screening and on-site detection. Chapter 3 uses NIR and ML approaches to classify coconut milk types (fresh coconut milk, FCM; instant coconut milk, ICM; adulterated fresh coconut milk, A-FCM) and predicts distilled water (DW) adulteration in fresh coconut milk. A data sciences approach that combines appropriate preprocessing discovery and hyperparameter optimization concurrently is presented in this study. Partial least squares (PLS), linear discriminant analysis (LDA), support vector machine

(SVM), and multilayer perceptron (MLP) with its hyperparameter were employed together with combining 18 preprocessing types and evaluated by 5-fold cross-validation (5f-CV). All regressors obtained the same satisfactory results to distinguish FCM, ICM, and A-FCM. Reressor from SVM obtained acceptable results, with R_c^2 and R_p^2 over 0.93, RMSEc and RMSEp below 8.30%, and RPD over 3.80. In Chapter 4, a novel approach to automatically select preprocessing (single up to multiple) and tuning hyperparameters simultaneously of ML algorithms based on their best performance in 5f-CV for FT-NIR and Micro-NIR spectroscopy data of coconut milk adulteration by DW and coconut water (CW) in the range 0 to 50%. This uses as many as 9 single preprocessing types and 3 types of ML classifier (LDA, KNN, MLP) and regressor (PLS, KNN, MLP). The performance strategy demonstrates that our proposed approach effectively addressed and produced satisfactory outcomes in classification and regression challenges and problems from coconut milk adulteration using NIRs. In Chapter 5, we explore DL algorithms that are only standardized using SNV preprocessing to identify the level of adulterated coconut milk using FT-NIR and Micro-NIR. Coconut milk adulteration samples came from intentional adulteration with corn flour and tapioca starch in the 1 to 50% range. Four types of DL algorithm architecture that were self-modified to a 1D framework were developed and tested, including CNN, S-AlexNET, ResNET, and GoogleNET. The results confirmed the feasibility of DL algorithms for predicting the degree of coconut milk adulteration by corn flour and tapioca starch with reliable performance (R^2 of 0.886–0.999, RMSE of 0.370–6.108%, and Bias of –0.176–1.481). In Chapter 6, we explore the discrimination model using FT-NIR and Micro-NIR for geographical source areas of coconut milk in tandem with the classical (PCA, PLS-DA, LDA) to modern chemometrics classifier, including classifiers from ML (SVM, KNN, ANN) and DL (S-CNN, S-AlexNET, ResNET). Three sources as geographical areas of coconut milk originally from Thailand were used, including Chumphon, Samut Songkhram, and Chonburi Province. Our findings showed that an SVM and ResNET classifier could yield the optimal performance for discriminating the geographical source area of coconut milk, with an accuracy of 99.1% for the training and 100% for the testing using FT-NIR. Furthermore, when using Micro-NIR, the LDA, SVM, and KNN, the ResNET classifier delivered the highest accuracy of 99.5% for the training and 100% for the testing.

ACKNOWLEDGEMENTS

The author would like to praise and thank God, "*Allah Subhanahu Wa Ta'ala*" for all His gifts, which enabled the successful completion of this thesis. I am very grateful for such a meaningful learning experience. The success of this thesis came about due to many sources of assistance and contributions. First of all, I would like to express my deepest gratitude to my advisors, Assistant Professor Dr. Ravipat Lapcharoensuk, for his valuable time and advice, as well as their encouragement throughout the study period. Particularly, I wish to acknowledge the financial support from the KMITL Doctoral Scholarship [KDS2020/049] for tuition fees, monthly expenses, as well as research funding.

I express my gratitude to all of the Professors in the Department of Agricultural Engineering at King Mongkut's Institute of Technology Ladkrabang (KMITL). They provided not only academic relationships but also everything for living, motivation, and support. I would like to special thank the Near-Infrared Spectroscopy Research Center for Agricultural Products and Food (www.nirsresearch.com) at King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand, for providing the necessary equipment and space for the experiments.

Also, many thanks are due to all Dr. Ravipat Lapcharoensuk laboratory members, including Mr. Wutthiphon Boodnon, Mr. Thayanont Lunvongsa, Mr. Phanchay Suntisakoonwong, Mr. Chen Moul, and Mr. Pich Khoem for assistance and friendship who took the time to support me in my hard research.

Finally, I would like to dedicate all the successes of this thesis to my family, especially my wife—Dewi Sartika Thamren, my mother—Alm Rokinim Silalahi, and my father—Muhamad Haidir Sitorus. They have consistently given me unconditional love and support as long as my studies.

Agustami Sitorus

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

No table of figures entries found.

LIST OF FIGURES

Page

***

CHAPTER 1 – GENERAL INTRODUCTION

1.1 Coconut Fruit for Coconut Milk

Coconut (*Cocos nucifera*) is a plant from the Arecaceae/Palmae and subfamily Arecoideae, widely cultivated in coastal and island environments. According to Nampoothiri *et al.* (2019) the top ten countries that produce coconuts are Indonesia, Philippines, India, Brazil, Sri Lanka, Vietnam, Papua New Guinea, Mexico, Thailand, and Malaysia. Coconut is an important source for both edible and industrial applications, most of which are used for culinary purposes. These products have become important as agro-based raw materials for many industries. As a result, it has become possible to encourage product diversification and the development of value-added products in the industry. Products like packed coconut milk, coconut creams, spray-dried coconut milk powder, and virgin coconut oil, which have good commercial potential, have been developed. According to the Asian and Pacific Coconut Community (APCC) Statistical Yearbook for coconut (2015) in Nampoothiri *et al.* (2019), Sri Lanka, Malaysia, and Indonesia together contribute more than 90% of the total export share of coconut milk.

Coconuts are harvested at different stages of maturity for specific uses. Only fully mature coconuts are harvested for the production of kernel-based products (coconut milk, coconut cream, desiccated coconut powder, and coconut oil). The coconuts reach full maturity in 11 until 12 months after the inflorescence is opened. At this stage, the output of oil and brown fiber would be the maximum. Coconut palm produces an average of 12 inflorescences in a year, some of the inflorescences are likely to get aborted or fail to develop into fruit bunches due to environmental factors. Thus, the number of bunches available for harvest is less than 12 in many areas.

Focusing on obtaining primary products from coconuts, such as coconut milk, several stages of postharvest must be passed (Figure 1.1). Coconut dehusking is the first postharvest operation in any coconut processing. The main objective of this process is to separate coconut with shell from coconut husk. Traditionally coconut is dehusked manually using a spike. However, several types of dehusking coconut machines have been developed for the industry. The second process of postharvest

handling is coconut deshelling. Traditionally coconut shell is removed using a knife. This process separates coconut kernel with testa with coconut shelling and coconut water. In addition, obtaining high-value coconut products requires the removal of tests. Removal of testa can use peeler tools. The main objective of this process is to separate the coconut kernel from the testa of coconut to obtain white coconut meat. The third process of postharvest handling is grating and grinding. The primary purpose of this process is to reduce the size of white coconut meat. The product obtained from this process is known as grated coconut meat. The fourth process of postharvest handling is extracting. This process involves pressing and straining grated coconut meat. The main purpose of this process is to separate coconut milk from coconut fiber or dregs. According to Cercado and Flat (2019), 100 g of grated coconut meat will yield a maximum of 63.7 g of coconut milk when extracted using an extractor machine without adding water.

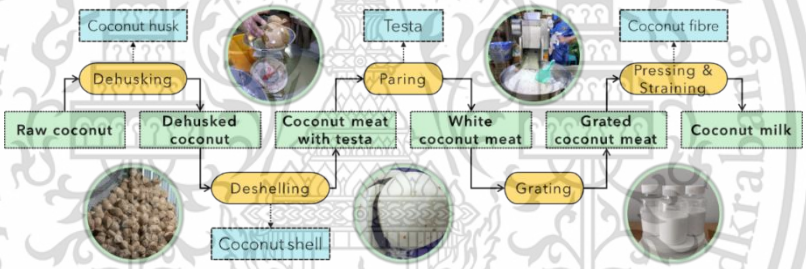


Figure 1.1. The steps involved in getting coconut milk.

The composition of coconut milk as the oil-in-water emulsion is also subject to the variety of coconut palms, geographical origin sources, and the maturity of coconut fruits. (Nampoothiri *et al.*, 2019; Patil and Benjakul, 2018). In generally, fresh coconut milk have pH between 5.6 to 6.3 and brix between 5.4 to 9.0. In addition, major nutrition components in coconut milk include moisture, fat, carbohydrate, protein, and ash. Based on the geographical origin sources and variety of coconut fruits from Asia, the main components of coconut milk are presented in Table 1.1. As the coconuts mature, the fresh endosperm's moisture, protein, and ash content slowly decrease, whereas the fat content increases. After the 10^{th} month, the protein content undergoes practically no change, but the fat content continues to increase until it reaches its peak in the 12^{th} and 13^{th} months. Subsequently, the fat content decreases until the 15^{th} month (Manikantan *et al.*, 2018).

Table 1.1. Composition of fresh coconut milk in tropical countries

Origin	Fat (%)	Protein (%)	Moisture (%)	Ash (%)	Carbohydrate (%)
Thailand[a]	35	4.02 ± 0.01	55.26 ± 0.12	1.02 ± 0.02	4.70
Malaysia[b]	15.44 ± 1.53	3.40 ± 0.59	73.57 ± 0.24	0.71 ± 0.01	n.a
Malayasia[c]	18.21 ± 1.25	4.20 ± 0.20	71.04 ± 0.65	n.a	2.26 ± 0.06
Indonesia[d]	36.93	3.88	55.79	0.99	2.93
Philippines[e]	33.4	4.1	56.3	1.2	5.0
Sri lanka[e]	39.8	3.0	50	1.2	6.0

Source: [a]-Tansakul and Chaisawang (2006); [b]-Alyaqoubi *et al.* (2015); [c]-Azlin-hashim *et al.* (2019); [d]-Karouw and Santosa (2018); [e]-Manikantan *et al.* (2018).

According to CODEX-STAN-240 (2003), the liquid fresh coconut milk can be categorized into light coconut milk, coconut milk, coconut cream, and coconut cream concentrate, based on total solids, non-fat solids, total fat content, and moisture content (Table 1.2). Light coconut milk shall be the product obtained from either the bottom portion of centrifuged coconut milk or by further dilution of coconut milk (it refers to coconut milk after dilution with some material). Coconut milk is the dilute emulsion of comminuted coconut endosperm (kernel) in water with the soluble. Coconut cream is an emulsion extracted from the matured endosperm (kernel) of coconut fruit, with or without adding coconut water. Coconut milk and coconut cream refers to fresh coconut milk. Coconut cream concentrate is the product obtained after partially removing water from coconut cream and complies (it' refers to instant coconut milk).

Table 1.2. Standard composition for coconut milk products.

Product	Total solids (% m/m)	Non-fat Solids (% m/m)	Fat (% m/m)	Moisture (% m/m)	pH
	Min — max	Min	Min	Max	Min
Light coconut milk	6.6 – 12.6	1.6	5.0	93.4	5.9
Coconut milk	12.7 – 25.3	2.7	10.0	87.3	5.9
Coconut cream	25.4 – 37.3	5.4	20.0	74.6	5.9
Coconut cream concentrate	> 37.4	8.4	29.0	62.6	5.9

Source: CODEX-STAN-240 (2003).

Major nutrition components in coconut milk include moisture, fat, carbohydrate, protein, and ash, which are responsible for the spectral absorption features. All Major nutrition components in coconut milk are typically related to hydrogen bonds such as O-H, C-H, S-H, and N-H. Several components in coconut milk that correlate with bonds in the NIR area are presented in Table 1.3.

Table 1.3. Wavelength NIR region potentially correlated with structure in coconut milk

Wavelength (nm)*	Functional group	Note
927, 1206, 1200[a]	C–H stretching vibration of $–CH_2$ or $–CH_3$. C–H 3rd and 2nd overtone stretching.	Fat or fatty acids
1202[b]	2nd overtone of CH stretching	Carbohydrates
944, 1450[a]	O–H 2nd and 1st stretching overtones	Water
1400 – 1550[a]	Vibrations of O–H, C–H, and N–H groups.	Protein
1555[b]	1st overtone N–H stretch	Protein
2099[b]	Combination bands [CO stretch and OH bend]	Carbohydrates
2100, 2200[a]	Vibration of N–H bending and C=O stretching	Protein
2136, 2346[a]	C–H stretching overtones	Fat or fatty acid content
2190[c]	Double bond in the molecule (C=C)	Ash
2207[b]	Combination bands [CO and NH stretching and bending]	Protein
2282[b]	Combination bands [OH, NH_2, stretching and bending]	Carbohydrates
2313[b]	Combination bands CH stretching and bending	Carbohydrates

*Source: [a]-Pandiselvam *et al.* (2022); [b]-Nallan Chakravartula *et al.* (2022); [c]-Wattanapahui *et al.* (2013).

1.2 Chemical Content of Coconut Milk

This sub-chapter presents the own results of direct laboratory analysis of the coconut milk samples used in this study. The main content analysis of coconut milk includes moisture, fat, protein, ash, and carbohydrate content. In general, the method used is proximate analysis. Total carbohydrates were measured using an analysis method for nutrient labeling (1993). Moisture and ash content are subject to test method AOAC (2019) from code 920.151A and 940.26, respectively. For fat and

protein content (Nx6.25) use the in-house method from code T967 base on AOAC (2019) 989.05 and T927 base in AOAC 991.20, respectively. All units measured in g/100g.

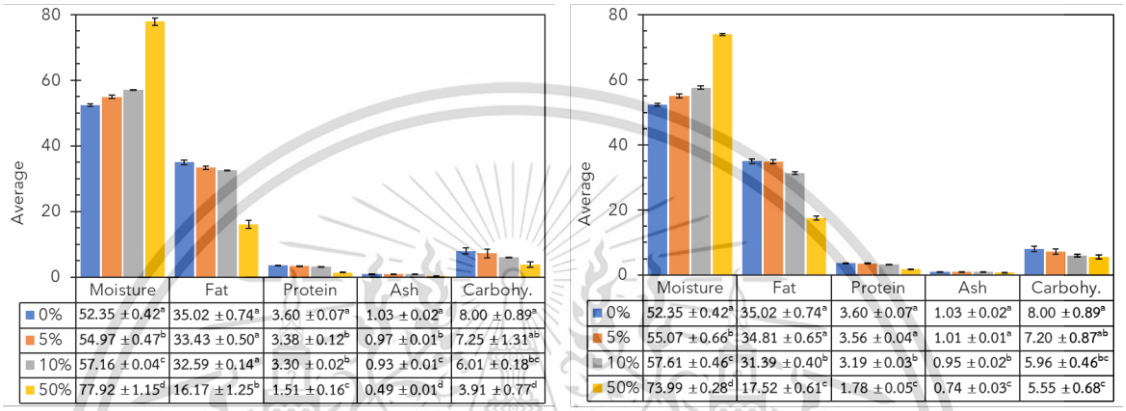
The sample analysis consisted of adulterated coconut milk with liquid adulterant (distilled and mature coconut water) and solid adulterant (corn flour and tapioca starch). Adulteration levels are 0%, 5%, 10%, and 50%. After that, samples from several areas or provinces in Thailand were also analyzed for their main contents. Each analysis sample was triplicated. Finally, the analysis of variance (ANOVA) with a significance level of 0.05 ($p<0.05$) was verified if at least one sample is different from the others, and a least significant difference (LSD) test compares the bias between models.

1.2.1 Effect of Liquid Adulteration Type and its Adulteration Level

ANOVA results from laboratory testing data for the main content of coconut milk on the type of adulterant from the liquid material and the level of adulteration are presented in Supplementary Material Chapter 1 from Table S1-1 to Table S1-5. Based on this analysis, it is known that all the main content of coconut milk that were analyzed have a significant effect on the variance in adulteration levels. However, the variable of adulteration type from liquid adulterant showed insignificant results in fat and carbohydrate content. This can be expected because the adulterant material is water and mature coconut water, which is low in fat and carbohydrate content from both the glucose and fructose groups.

The result from least significant difference (LSD) test with level of significance 5% with to each main compotent of coconut milk from sampels adulterant by destilated water and mature coconut water present in Figure 1.2. It can be seen that samples adulterated with distilled water and mature coconut water can be clearly differentiated, at least by moisture content, for all levels of adulteration. However, adulteration by mature coconut water can be significantly different for all levels of adulteration with ash content as well. Meanwhile, the fat, protein, and ash contents tend to differ significantly after being adulterated at the 10% level, except for adulteration by mature coconut water. Moreover, at the 50% adulteration level, the main content of coconut milk is significantly different. This can be an initial indication that the presence of different moisture content resulting from the adulteration of

coconut milk with distilled water or mature coconut milk should also be detected by NIR via O–H 2nd and 1st stretching overtones at 944 and 1450 nm (Pandiselvam *et al.*, 2022). Based on this information, a case study to develop a qualitative adulteration discrimination model and a quantitative prediction model for adulteration levels from coconut milk products adulterated with liquid adulterant material is feasible in Chapter 3 and Chapter 4.



	Moisture	Fat	Protein	Ash	Carbohy.
0%	52.35 ±0.42^{a}	35.02 ±0.74^{a}	3.60 ±0.07^{a}	1.03 ±0.02^{a}	8.00 ±0.89^{a}
5%	54.97 ±0.47^{b}	33.43 ±0.50^{a}	3.38 ±0.12^{b}	0.97 ±0.01^{b}	7.25 ±1.31^{ab}
10%	57.16 ±0.04^{c}	32.59 ±0.14^{a}	3.30 ±0.02^{b}	0.93 ±0.01^{c}	6.01 ±0.18^{bc}
50%	77.92 ±1.15^{d}	16.17 ±1.25^{b}	1.51 ±0.16^{c}	0.49 ±0.01^{d}	3.91 ±0.77^{d}

(a)

	Moisture	Fat	Protein	Ash	Carbohy.
0%	52.35 ±0.42^{a}	35.02 ±0.74^{a}	3.60 ±0.07^{a}	1.03 ±0.02^{a}	8.00 ±0.89^{a}
5%	55.07 ±0.66^{b}	34.81 ±0.65^{a}	3.56 ±0.04^{a}	1.01 ±0.01^{a}	7.20 ±0.87^{ab}
10%	57.61 ±0.46^{c}	31.39 ±0.40^{b}	3.19 ±0.03^{b}	0.95 ±0.02^{b}	5.96 ±0.46^{bc}
50%	73.99 ±0.28^{d}	17.52 ±0.61^{c}	1.78 ±0.05^{c}	0.74 ±0.03^{c}	5.55 ±0.68^{c}

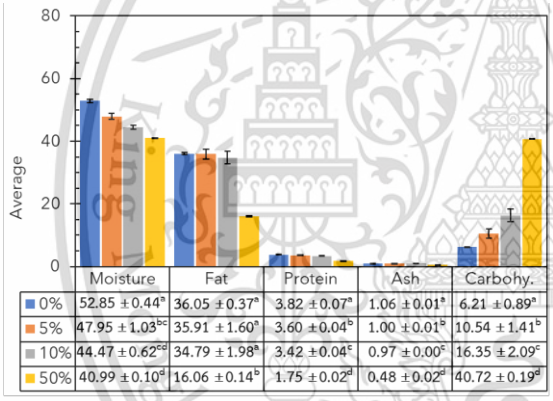
(b)

Figure 1.2. LSD analysis of coconut milk samples adulterated by (a) distilled water and (b) mature coconut water.

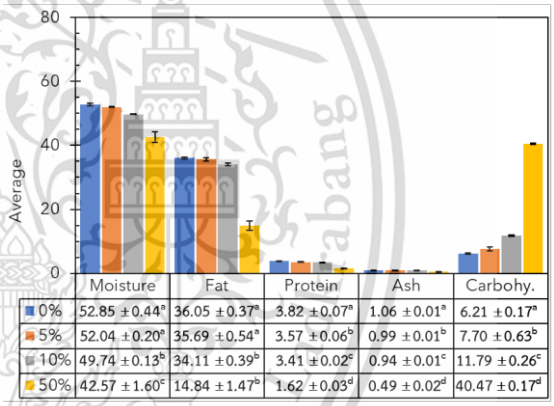
1.2.2 Effect of Solid Adulteration Type and its Adulteration Level

This sub-chapter notes the effect of solid adulteration type (corn flour and tapioca starch) and its adulteration level (0%, 5%, 10%, and 50%) on testing chemical laboratory results from the main content of coconut milk. The results of the ANOVA for moisture, fat, protein, ash, and carbohydrate content, are presented in Supplementary Material Chapter 1 from Table S1-6 to Table S1-10. Based on the results of the ANOVA analysis, the type of solid adulteration significantly affects the moisture and carbohydrate content of adulterated coconut milk. On the other hand, the adulteration level applied in this sub-chapter is known to significantly affect all the main content of coconut milk. It is reasonable to assume that this is influenced by differences in the moisture and starch content of the adulterant material originating from the powder. This causes the samples to achieve their equilibrium points for the two adulterants after mixing. The level of adulteration applied will also impact other coconut milk contents (fat, protein, and ash).

The result from least significant difference (LSD) test with level of significance 5% with to each main compotent of coconut milk from sampels adulterant by corn starch and tapioca flour present in Figure 1.3. In contrast to the case of adulteration with liquid adulterant, which causes the moisture content in coconut milk to increase and other contents to tend to decrease, adulteration with solid ingredients such as flour causes all the main contents of coconut milk to decrease except carbohydrate content. This is believed to be caused by the main content of the adulterant material, such as corn flour and tapioca flour, which is rich in starch content, which can increase the carbohydrate content in adulterated coconut milk samples. A similar finding was reported by Azlin-hashim *et al.* (2019) that coconut milk in the traditional market may have been possibilities adulterated with corn flour to increase carbohydrates.



	Moisture	Fat	Protein	Ash	Carbohy.
0%	52.85 ±0.44^{a}	36.05 ±0.37^{a}	3.82 ±0.07^{a}	1.06 ±0.01^{a}	6.21 ±0.89^{a}
5%	47.95 ±1.03^{bc}	35.91 ±1.60^{a}	3.60 ±0.04^{b}	1.00 ±0.01^{b}	10.54 ±1.41^{b}
10%	44.47 ±0.62^{cd}	34.79 ±1.98^{a}	3.42 ±0.04^{c}	0.97 ±0.00^{c}	16.35 ±2.09^{c}
50%	40.99 ±0.10^{d}	16.06 ±0.14^{b}	1.75 ±0.02^{d}	0.48 ±0.02^{d}	40.72 ±0.19^{d}



	Moisture	Fat	Protein	Ash	Carbohy.
0%	52.85 ±0.44^{a}	36.05 ±0.37^{a}	3.82 ±0.07^{a}	1.06 ±0.01^{a}	6.21 ±0.17^{a}
5%	52.04 ±0.20^{a}	35.69 ±0.54^{a}	3.57 ±0.06^{b}	0.99 ±0.01^{b}	7.70 ±0.63^{b}
10%	49.74 ±0.13^{b}	34.11 ±0.39^{b}	3.41 ±0.02^{c}	0.94 ±0.01^{c}	11.79 ±0.26^{c}
50%	42.57 ±1.60^{c}	14.84 ±1.47^{b}	1.62 ±0.03^{d}	0.49 ±0.02^{d}	40.47 ±0.17^{d}

(a) (b)

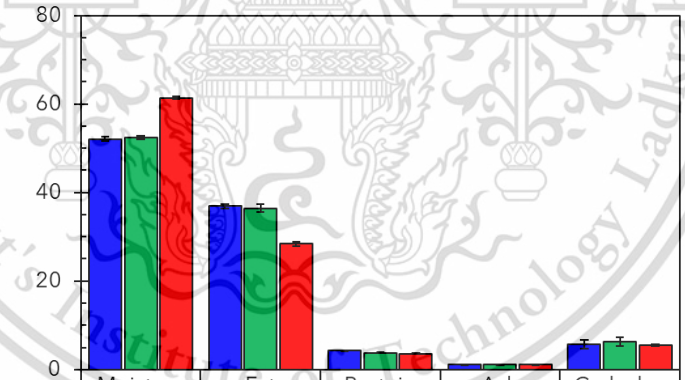
Figure 1.3. LSD analysis of coconut samples adulterated by (a) corn flour and (b) tapioca starch.

Because significant differences in the moisture and carbohydrate content of coconut milk after adulteration with solid materials such as corn starch and tapioca flour have been marked, we hypothesize that the type of adulteration by solid materials should also be discriminated by NIR spectroscopy. According to Pandiselvam *et al.* (2022), due to the difference in content, NIR will be able to detect through the functional group O–H 2nd and 1st stretching overtones related to moisture in the wavelength range of 944 and 1450 nm. Besides that, at a wavelength of 1202 nm, which is related to the 2nd overtone of CH stretching, and at

wavelengths 2099, 2282, and 2313 nm, which are related to the combination band of carbohydrates (Nallan Chakravartula *et al.*, 2022; Pandiselvam *et al.*, 2022). According to the information above, conducting a case study in Chapter 5 is feasible for developing a quantitative prediction model for determining the level of adulteration in coconut milk products contaminated with solid adulterant material.

1.2.3 Effect of Source of Coconut Fruits

This sub-chapter notes the effect of source coconut fruits representative of different geographical areas of Thailand (south, middle, east) were collected from local coconut farms located in three provinces, including Chumphon (CHP), Samut Songkhram (SSK), and Chonburi (CHB) present in Figure 1.4. The moisture and fat content indicate that the source area of coconut fruits produced into coconut milk does not differ significantly between coconuts originating from CHB and CHP except SSK. Even the carbohydrate content of the coconut fruit used to make coconut milk is no different. However, the protein and ash content was significantly different from the three coconut milk origins at a significant level of 5%.



	Moisture	Fat	Protein	Ash	Carbohy.
CHP	52.10 ±0.46^{a}	36.92 ±0.50^{a}	4.22 ±0.12^{a}	1.05 ±0.01^{a}	5.71 ±1.06^{a}
CHB	52.40 ±0.40^{a}	36.44 ±0.88^{a}	3.80 ±0.04^{b}	1.09 ±0.01^{b}	6.27 ±0.99^{a}
SSK	61.39 ±0.37^{b}	28.38 ±0.49^{b}	3.58 ±0.08^{c}	1.18 ±0.02^{c}	5.48 ±0.25^{a}

Figure 1.4. LSD analysis of coconut milk samples based on the source of coconut fruits.

Because coconut milk from different geographical areas of Thailand has significantly different protein and ash content, we hypothesize that NIR spectroscopy can be employed to differentiate the sources or geographical origins of coconut fruit

used to produce coconut milk. This is because different protein contents should be able to be identified by NIR via several wavelengths starting from 1440-1550 nm as vibrations of O–H, C–H, and N–H groups and 1555 nm as the 1st overtone N–H stretch of proteins. Besides that, it is also reported that it can be identified at 2100 nm and 2200 nm through the vibration of N–H bending and C=O stretching and at 2207 nm, which is associated with combination bands also from proteins (Nallan Chakravartula et al., 2022; Pandiselvam et al., 2022). Meanwhile, the ash content reported by Wattanapahui et al. (2013) can be identified in the wavelength of 2190 nm, related to double bonds in the molecule (C=C). Based on this information, a case study to develop a qualitative discrimination model for the source of coconut milk products is feasible in Chapter 6.

1.3 Theory of Near-Infrared Spectroscopy

The near–infrared (NIR) region of the electromagnetic spectrum extends from about 700 to 2500 nm (or 14286 to 4000 cm^{-1}). Of this range, the 700–1100 nm region is referred to as the short wave NIR (SWNIR), or Herschel region, while the 1100–2500 nm region is considered the NIR region proper (Pu *et al.*, 2020). NIR region is the region that is closest to the visible right, thus, it is called 'near' infrared. The layout of the electromagnetic spectrum, showing the position of the infrared region relative to the rest of the frequency bands, is illustrated on a wavelength scale in Figure 1.5. As seen, the wavelength progressively increases from the visible region through infrared to radio waves. The characteristics of a vast range of wavelengths, including NIR at the transition between the visible and infrared regions, allow their potential use in spectroscopic analysis.

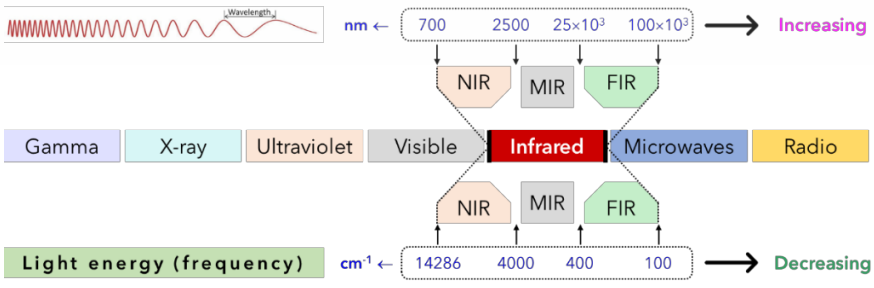


Figure 1.5. The electromagnetic spectrum (modified from Chu *et al.* (2022)).

Spectroscopy studies the electromagnetic interaction energy interacts with matter. The incident radiation may be reflected, absorbed, and transmitted in which each phenomenon depends on the chemical constitution and physical parameters of the matter when light emits on the matter (Figure 1.6). The Beer-Lambert law for a dilute (Equation 1.1), non-scattering solution relates absorbance (*A*) to the path length of the sample in cm (*b*), molar extinction coefficient in L/mol.cm (*a*), and concentration of the absorbing analyte in mol/L (*c*).

$$A = abc \tag{1.1}$$

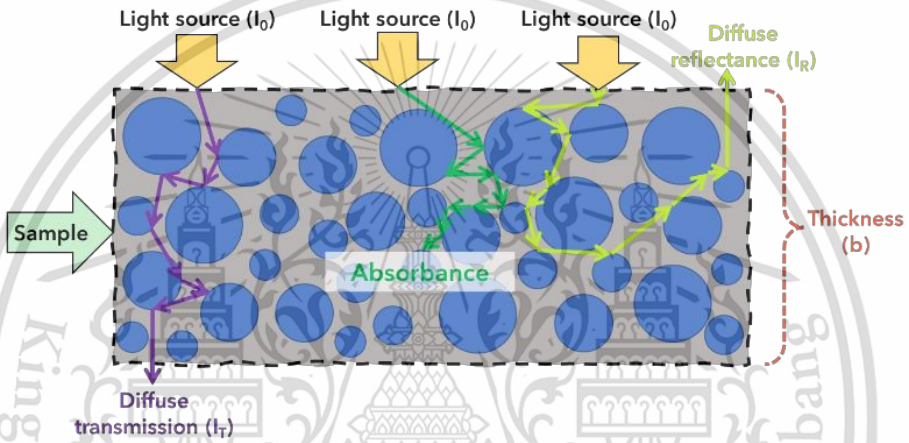


Figure 1.6. Interactions between light and matter.

Incident radiation passing through a medium undergoes several changes, the extent of which depends on the physical and chemical properties of the medium. Typically, part of the incident beam is reflected, another part is absorbed and transformed into heat by interaction with the material, and the rest passes through the medium. Transmittance (*T*) is defined as the ratio of the transmitted light intensity (I_T) to the incident light intensity (I_0) (Equation 1.2). Absorbance (*A*) is defined as the logarithm of the inverse of the transmittance (Equation 1.3). Absorbance is a positive value without units. Due to their inverse relationship, absorbance is greater when the transmitted light is low (Holden *et al.*, 2021). The absorbance of each wavelength often plotted as wavelength versus log (1/R).

$$T = \frac{I_T}{I_0} \tag{1.2}$$

$$A = -\log(T) = \log\left(\frac{1}{T}\right) = \log\left(\frac{I_0}{I_T}\right) = \log\left(\frac{I_0}{I_R}\right) \tag{1.3}$$

NIR absorbance corresponds to overtones and combinations of vibrations band of molecular bonds, primarily of O-H, C-H, N-H, and C=O groups that have their fundamentals in NIR spectroscopy. These overtones are multiple frequencies of fundamental vibrations and anharmonic that do not behave in a simple fashion, making NIR spectra complex and not directly interpretable as in other spectral regions. Energy (E) is proportional to the frequency (υ) absorbed (Equation 1.4), which in turn is proportional to the wavenumber ($1/\lambda$), the first overtone that appears in the spectrum will be twice the wavenumber of the fundamental. That is, the first overtone is (approximately) twice the energy of the fundamental (Eldin and Akyar, 2011). NIR wavelengths from 700 to 2500 nm are absorption bands from high (3^{rd}) to low (1^{st}) overtone. The SW-NIR region (700–1100 nm) is regarded as the absorption band of high overtones (3^{rd}), while the 1100–2500 nm region belongs to the 2^{nd} and 1^{st} overtones. Therefore, the absorption intensity will increase when the overtone decrease. This makes the 700–1100 nm wavelengths more suitable for transmission analysis with long path lengths due to the low absorption intensity, and wavelengths 1100–2500 nm are used in diffuse reflection due to the high absorption intensity. The fundamental vibrations modes are stretching and bending (Figure 1.7). Table 1.4 show a brief overview of the frequency ranges which are normally observed in near-infrared spectroscopy.

$$E = h\upsilon = h\frac{c}{\lambda} = hc\bar{\nu} \rightarrow \upsilon = \frac{c}{\lambda}; \upsilon = c\bar{\nu} \qquad (1.4)$$

Where, E=energy of photons (J); h=Plank's constant (6.63×10^{-34} J·s); υ=frequency (Hz); λ=wavelength (m); $\bar{\nu}$=wavenumber (1/cm); c=speed of light (3×10^{8} m/s).

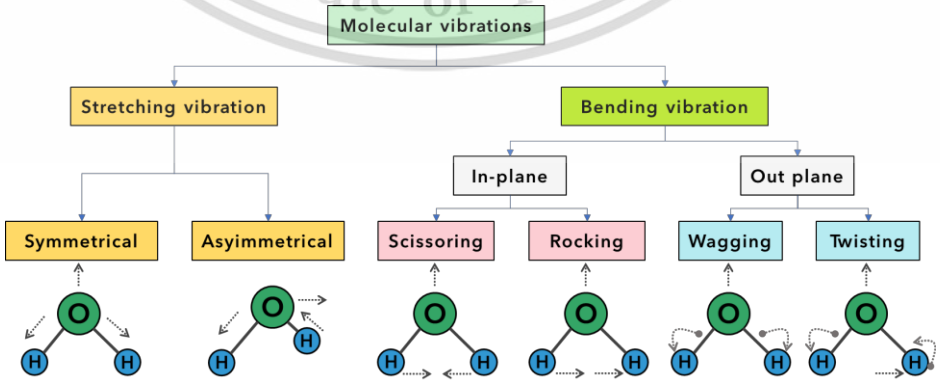


Figure 1.7. Fundamental vibrations modes (Stuart, 2004)

Several steps are needed to build a robust calibration model utilizing near-infrared spectroscopy, including starting with careful sample preparation (such as sample temperature and ambient temperature) to acquire NIR spectral data that can represent sample contents. Furthermore, it is necessary to preprocess the spectral data to eliminate noises, baseline shifts, and multiplicative errors from some sources such as instruments, samples, and backgrounds. In addition, destructive testing after scanning using NIR instruments needs to be carried out immediately to obtain good reference data. Then a calibration model (correlation between the data spectral and the reference data) in the form qualitative model or quantitative model is developed using some algorithms, and each algorithm is evaluated with various error parameters. Finally, the model with the lowest error can be tested using an unknown set of samples to get the deployment of the model.

Table 1.4 Common near-infrared bands of organic compounds.

Wavelength (nm)	Assignment
2200 – 2450	Combination C–H stretching
2000 – 2200	Combination N–H stretching, Combination O–H stretching
1650 – 1800	1st overtone C–H stretching
1400 – 1500	1st overtone N–H stretching, 1st overtone O–H stretching
1100 – 1225	2nd overtone C–H stretching
950 – 1100	2nd overtone N–H stretching, 2nd overtone O–H stretching
850 – 950	3rd overtone C–H stretching
775 – 850	3rd overtone N–H stretching

Source: Eldin and Akyar (2011); Workman Jr and Weyer (2007).

1.4 Preprocessing of NIR Spectra Data

Before further analysis, different preprocessing steps were carried out on the collected spectra to enhance the useful features in the spectra, remove noise, and restrict the wavelength range to include only valid information. In addition to the chemical information of the sample itself, the collected near-infrared spectroscopy data may contain other irrelevant information and noise, including interfering physical and/or chemical factors, imperfections in the experimental apparatus,

and/or other random factors (Xu *et al.*, 2008). Spectral preprocessing is key to extracting accurate information in near-infrared spectroscopy data analysis and enhancing subtle differences between different samples. Figure 1.8 shows a flowchart for selecting the preprocessing technique to use.

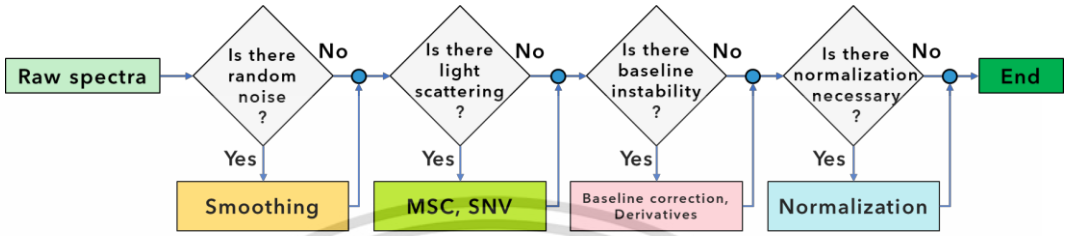


Figure 1.8. Selecting the preprocessing technique (Xu *et al.*, 2020)

Smoothing the near-infrared spectral can retain useful spectral information and remove random noise at the same time. At present, the used smoothing methods of near-infrared spectrum include moving window average method, moving window median method, and Savitzky-Golay filtering method (SG). The most common smoothing technique is the SG filtering method, which is based on a polynomial equation fitted in a least squares sense within a predefined interval of spectral points. This interval is then displaced to the next point of the spectrum, and the fitting procedure is repeated (Equation 1.5).

$$X_t = \frac{\sum_{i=t-(ss-1)}^{t} (X_i)}{SS} \qquad (1.5)$$

Where, X_t=moving average value of X for time-t; X_i=smoothed value of X for time-i; SS=segment size value.

NIR spectral data may include baseline shift and multiplicative effects induced by physical effects, such as the non-uniform scattering throughout the spectrum, as the degree of scattering depends on the wavelength of the radiation, the particle size, and the refractive index. MSC corrects scattering, maintaining the original spectral shape and the same spectral scale. The idea behind MSC is that the two effects including amplification (multiplicative, scattering) and offset (additive, chemical), should be removed from the data table to avoid that they dominate the information (signal) in the data. Two correction coefficients, intercept (a) and slope (b), are calculated from a reference from the average spectrum in the data set and

used in these computations to determine MSC preprocessing using equation 1.6. Since the MSC normalizes based on the mean spectrum in a data set, it is best suited for similar sample sets (CamoASA, 2014).

$$X_t = \frac{Xi - a}{b} \quad (1.6)$$

Standard normal variate (SNV) is mainly used to eliminate the influence of solid particle size, surface scattering and optical path change on diffuse reflectance spectrum. The practical difference with MSC preprocessing is that SNV standardizes each spectrum using only the data from the self-spectrum; it does not use the mean spectrum of any set (Equation 1.7). Therefore, the average value and standard deviation of each spectrum obtained from preprocessing using SNV are zero and one, respectively. SNV preprocessing not only retains the original spectrum shape but also creates an artificial absorbance scale with negative values.

$$X_t = \frac{Xi - \overline{X}}{SD_X} \quad (1.7)$$

Baseline offset correction was used to adjust the spectral offset by adjusting the data to the minimum point (Equation 1.8). Baseline correction is used to separate true spectroscopic signals from interference effects or remove background effects, stains, compound traces, instrument noise effects, and light-scattering particulates in the sample that cause an offset in the overall sample absorbance.

$$X_t = X_i - \min(X) \quad (1.8)$$

Derivatives are applied to correct baseline effects in spectra for the purpose of removing nonchemical effects and creating robust calibration models. Derivatives may also help in resolving overlapped bands which can provide a better understanding of the data, emphasizing small spectral variations not evident in the raw data.

First derivative is a very effective method for removing such baseline offsets. Second derivative is very effective method for removing both the baseline offset and slope from a spectrum. The second derivative can help resolve nearby peaks and sharpen spectral features. Peaks in raw spectra change sign and turn to negative peaks on either side in the second derivative.

The Savitzky-Golay method is the most popular derivatives method for preprocessing near-infrared spectral data. The Savitzky-Golay is based on a localized

linear regression of several neighboring points to determine the best fit polynomial. This polynomial can be mathematically differentiated and evaluated at the x values coincident with wavelength collection points. In practice, a mathematical equivalent of the regression and differentiation procedure is performed by a convolution with a set of derived coefficients (Delwiche and Reeves, 2010).

Spectra data normalization is to scale and translate the data in proportion to make the data fall into a small specific interval. Numerous normalization techniques available, mainly including area normalization, vector normalization, mean normalization, maximum normalization, range normalization, and peak normalization. However, because the normalization might hide important spectral bands, which may be discriminative features between samples, this procedure must be carried out only when needed and with care.

Area normalization attempts to correct the transmission spectra for indeterminate path length when there is no way of measuring it or isolating a band of a constant constituent or of an internal standard. This transformation divides each spectral data point (X_i) by the area under the curve (A_X) for the observation. (Equation 1.9).

$$X_t = \frac{X_i}{A_X} \tag{1.9}$$

Vector normalization can be used for pattern normalization, which is useful for preprocessing in some pattern recognition applications. Vector transformation is to divide each spectral data point (X_i) by unit vector (A_v) for the observation (Equation 1.10).

$$X_t = \frac{X_i}{A_v} = \frac{X_i}{\sqrt{\sum_1^n x_i^2}} \tag{1.10}$$

Mean normalization can be used in chromatography to express the results in the same units for all samples, regardless of which volume was used for each. The mean transformation divides each spectral data point (X_i) by the observation's unit average (A_m) (Equation 1.11).

$$X_t = \frac{X_i}{A_m} \tag{1.11}$$

Maximum normalization can be used to normalize the spectral so that it becomes at the same level as the highest intensity peak in other spectra (Deininger

et al., 2011). Maximum transformation is to divide each spectral data point (X_i) by the maximum value of the spectral itself (A_{max}) (Equation 1.12).

$$X_t = \frac{X_i}{A_{max}} \quad (1.12)$$

Range normalization can be used to handle differences in the baseline, possibly due to variations in shape, thickness, surface texture, and physical damage affecting the scattering (Sundaram *et al.*, 2010). Range transformation is to divide each spectral data point (X_i) by the difference between the maximum and minimum values of the spectral itself ($A_{max-min}$) (Equation 1.13).

$$X_t = \frac{X_i}{A_{max-min}} \quad (1.13)$$

Peak normalization can be used to correct spectra for indeterminate path length. By peak normalizing the spectrum to the intensity of the peak, the path length variation is effectively removed. Peak transformation is to divide each spectral data point (X_i) by peak normalization (A_P), which is the total number of variables (Equation 1.14).

$$X_t = \frac{X_i}{A_P} \quad (1.14)$$

1.5 Machine Learning

Machine learning is a branch of artificial intelligence that allows computer systems to learn from data without being explicitly programmed. This approach enables computers to identify complex patterns and make decisions or predictions based on the given data. Machines can learn to perform specific tasks like classification and regression using algorithms and statistical models. One of machine learning's advantages is its ability to process vast and complex data sets and discover patterns that may be difficult to recognize. Therefore, machine learning has been used in various fields, including spectroscopy in chemometrics. Its ability to improve its performance over time by learning from experience makes it a valuable tool in today's era of big data.

Machine learning has been extensively used in processing and analyzing Near-Infrared (NIR) spectroscopy data. NIR spectroscopy is a powerful analytical technique used to determine the chemical composition of substances by measuring the absorption of near-infrared light. One critical application of machine learning in NIR

spectroscopy is in chemometric modeling. Chemometrics uses statistical and mathematical methods to extract meaningful information from chemical data. Machine learning algorithms, such as support vector machines (SVM), K-Nearest neighbors (KNN), and artificial neural networks (ANN), can be trained on NIR spectra and corresponding reference values to build predictive models for various properties in samples.

1.6 Deep Learning

Deep learning is a subset of machine learning that uses artificial neural networks to model and solve complex problems. Unlike traditional machine learning approaches that rely on handcrafted features, deep learning algorithms learn hierarchical representations of data, automatically allowing them to discover intricate patterns and features from raw input. One of the critical advantages of deep learning is its ability to scale with large datasets, as neural networks can effectively learn from vast amounts of data to improve their performance. However, deep learning models are computationally intensive and require substantial resources for training, limiting their applicability in resource-constrained environments.

The application of deep learning in NIR spectroscopy has shown great promise in various aspects of spectral data analysis. Deep learning models, particularly convolutional neural networks (CNNs), can effectively learn complex patterns and relationships within NIR spectra. This allows for more accurate and robust chemical properties or sample component predictions. These models excel in feature extraction and selection, automatically identifying relevant spectral features without manual intervention. Additionally, deep learning enables the development of models that can handle non-linear relationships and interactions in the data, which are common in NIR spectroscopy. By leveraging the large amounts of data typically available in spectroscopic applications, deep learning algorithms can improve models' prediction accuracy and generalization, leading to enhanced performance.

Compared to deep learning, traditional machine learning approaches have limitations when applied to NIR spectroscopy data processing. One key challenge is requiring a large amount of data for practical training, particularly for deep neural networks. In the context of NIR spectroscopy, collecting a sufficiently large dataset

can be challenging, which may hinder the application of deep learning models. Additionally, with their complex architectures, deep learning models require intensive computational resources, posing a challenge for users with limited computing capabilities. Moreover, deep learning models are often considered "black boxes," making it difficult to interpret the relationship between the spectral data and chemical properties, which is crucial in NIR spectroscopy analysis. Furthermore, deep learning models are prone to overfitting, especially in cases where the data has complex variations, which can be a significant concern in NIR spectroscopy data processing. Despite these challenges, deep learning can substantially benefit NIR spectroscopy analysis when correctly applied and tailored to specific data conditions.

1.7 Research Problem

The fundamental challenge addressed in this thesis is how to develop a robust approach model to carry out rapid and non-destructive detection and screening based on NIR spectroscopy for adulteration of coconut milk products. As has been reported in previous research, at least by Azlin-hashim *et al.* (2019), coconut milk products, especially fresh coconut milk, are very easily adulterated by adding some adulterant from liquids (including water and mature coconut water) or solid (including corn flour and tapioca starch). On the other hand, how many previous research reports to date have there been extensive applications of NIR spectroscopy in detecting adulteration for agricultural and food products.

Besides, the main composition of coconut milk will be different due to the diversity of geographical areas that cultivate coconut into coconut milk. Therefore, the next research problem is how to explore a robust discrimination model to classify fast and non-destructive based on NIR spectroscopy from coconut milk with different geographical origins, mainly from Thailand in some provinces.

Another major research problem related to the difficulty of obtaining a robust model from the application of NIR spectroscopy is the presence of noise and scattering. This causes NIR spectroscopy to depend on preprocessing before conducting modeling. Many preprocessing techniques have been developed for NIR spectroscopy, so finding the best preprocessing method currently requires trial and

error. Therefore, the main problem being addressed in this study is how to streamline the process of getting the best preprocessing.

Recently, the development of chemometrics for NIR spectroscopy has also directed to the application of machine learning and deep learning. These approaches are known to increase models' robustness in qualitative and quantitative case studies and can work more deeply. Therefore, how to incorporate these more advanced chemometric techniques (the first is machine learning) in developing calibration and discrimination models combined with preprocessing selection automation. Furthermore, how to employ other more advanced techniques (the second is deep learning) that are highly adaptive and have almost minimal preprocessing approaches.

1.8 Research Limitation

Some research limitation this research are as follows. We have conducted a literature review study to answer the question of how many previous research reports have extensively applied NIR spectroscopy to detect adulteration in agricultural and food products, as shown in Chapter 2. However, this research limitation is the time range from 1990 to February 2022, based on the Scopus electronic database (www.scopus.com). The keywords for finding the research papers are only “NIR” or “near-infrared” and “adulteration.”

In chapters 3 to 4, we have developed a method to answer questions about how to select the best preprocessing for NIR spectroscopy data. In Chapter 5, we used deep learning that has minimum preprocessing stages. Also, in chapters 3 to 5 answer how to develop a robust approach model to carry out rapid and non-destructive detection and screening based on NIR spectroscopy for adulteration of coconut milk products. However, the limitations of each chapter are as follows. Chapter 3 is limited to only carrying out an automatic selection of 18 preprocessing singles using algorithms (classifier and regressor) from machine learning. Besides, adulteration uses a single material (distilled water) with an adulteration range of 0-100% and uniform increasing levels of 10% (w/w). Chapter 4 tries to cover the weaknesses of Chapter 3 by implementing an automatic selection of 9 preprocessing multiples (4 layers) using algorithms (classifier and regressor) from machine learning.

Again, this study case has limitations: the adulteration range is also reduced to a maximum of 50% (w/w) with non-uniform increasing levels. However, adulteration in this study was still binary, using liquid materials, including distilled water and mature coconut water. Chapter 5 switches from the ML to DL approach with the research limitation of only one type of processing (SNV) for the regression case, while the adulteration range is also reduced to a maximum of 50% (w/w) with non-uniform increasing levels.

Lastly, we have developed a method in Chapter 6 to answer the equation about how to explore a robust discrimination model to classify fast and non-destructive based on NIR spectroscopy from coconut milk with different geographical origins, mainly from Thailand in some provinces. However, the research limitations include using samples from three provinces in Thailand and only from one harvesting season.

1.9 Research Objectives

The main objective of this thesis is to research and develop chemometric tools based on machine learning and deep learning for NIR spectroscopy techniques to detect coconut milk adulteration and classify it based on its geographical area of origin. Meanwhile, the specific objectives of this research are as follows.

1. To overview of the previous research paper from the application of near-infrared and infrared spectroscopy in detecting and discriminating the adulteration of food and agro-products is based on recent research.
2. To develop a combination of single appropriate preprocessing discovery and hyperparameter optimization concurrently from a machine learning algorithm to classify the type of coconut milk (fresh coconut milk—FCM, instant coconut milk—ICM, adulteration of fresh coconut milk—A-FCM) and predict the level of adulteration of water in fresh coconut milk using NIR spectroscopy.
3. To develop a strategy for ensembling between selecting single up to multiple preprocessing and tuning hyperparameters of machine learning algorithms for cases of qualitative (classification of type adulteration) and quantitative (regression of level adulteration) from two types of NIR instruments, including benchtop FT-NIR and portable Micro-NIR.

4. To explore and test the performance of four types of regressor architecture CNN of deep-learning algorithms with minimum preprocessing to identify the level of adulterated coconut milk from corn flour and tapioca starch using two kinds of NIR spectrophotometer, including benchtop FT-NIR and portable Micro-NIR.
5. To explore the potential of classifier algorithms for NIR spectroscopy (benchtop FT-NIR and portable Micro-NIR) from classical approach (PCA, PLS-DA, LDA) to modern chemometrics, including classifiers from machine learning (SVM, KNN, ANN) and deep learning (S-CNN, S-AlexNET, ResNET) focusing specifically on geographical area classification from coconut milk in Thailand.

1.10 Navigation of the Thesis

This thesis is organized into six chapters. A brief discussion of every chapter is presented below.

Chapter 1 presents a brief general overview of near-infrared spectroscopy and samples used in this work that initiated this thesis, defines the research problem, research limitations, and research objectives, and shows the maps of this thesis book as well.

Chapter 2 presents a literature review of this work.

Chapter 3 presents the first full article from this work as a case study 1.

Chapter 4 presents the second full article from this work as a case study 2.

Chapter 5 presents the third full article from this work as a case study 3.

Chapter 6 presents the fourth full article from this work as a case study 4.

Chapter 7 presents the model deployment from case studies 1 until 4 with an unknown crossover dataset.

Chapter 8 presents the conclusion, recommendations, and future work from this thesis book.

1.11 References

Alyaqoubi, S., Abdullah, A., Samudi, M., Abdullah, N., Addai, Z. R., & Musa, K. H. (2015). Study of antioxidant activity and physicochemical properties of coconut milk (Pati santan) in Malaysia. *Journal of Chemical and Pharmaceutical Research, 7*(4), 967-973.

Azlin-hashim, s., Siang, q. l., Yusof, F., Zainol, M. K., & Yusof, H. M. (2019). Chemical composition and potential adulterants in coconut milk sold in Kuala Lumpur. *Malaysian Applied Biology, 48*(3), 27-34.

CamoASA. (2014). The Unscrambler X v10.3 User Manual: CamoASA Oslo.

Cercado, A. P., & Flat, X. (2019). Design and fabrication of equipment for extraction of coconut milk from shells. *International Research Journal of Advanced Engineering and Science, 4*(3), 27-29.

Chu, X., Huang, Y., Yun, Y.-H., & Bian, X. (2022). *Chemometric methods in analytical spectroscopy technology*: Springer.

CODEX-STAN-240. (2003). Standard for Aqueous Coconut Products-Coconut Milk and Coconut Cream.: FAO/WHO Food Standards Programme.

Deininger, S.-O., Cornett, D. S., Paape, R., Becker, M., Pineau, C., Rauser, S., Walch, A., & Wolski, E. (2011). Normalization in MALDI-TOF imaging datasets of proteins: practical considerations. *Analytical and bioanalytical chemistry, 401*(1), 167-181.

Delwiche, S. R., & Reeves, J. B. (2010). A graphical method to evaluate spectral preprocessing in multivariate regression calibrations: Example with Savitzky–Golay filters and partial least squares regression. *Applied spectroscopy, 64*(1), 73-82.

Eldin, A. B., & Akyar, I. (2011). Near infra red spectroscopy. *Wide spectra of quality control. InTech, Rijeka, Croatia*, 237-248.

Holden, N. M., Wolfe, M. L., Ogejo, J. A., & Cummins, E. J. (2021). Introduction to Biosystems Engineering *Introduction to Biosystems Engineering* (pp. 0): American Society of Agricultural and Biological Engineers.

Karouw, S., & Santosa, B. (2018). Stability of Coconut Milk on Various Addition of Sodium Caseinate as Emulsifier. *Buletin Palma Volume, 19*(1), 27-32.

Manikantan, M. R., Pandiselvam, R., Beegum, S., & Mathew, A. C. (2018). Harvest and Postharvest Technology. In V. Krishnakumar, P. K. Thampan, & M. A. Nair (Eds.), *The Coconut Palm (Cocos nucifera L.) - Research and Development Perspectives* (pp. 635-722). Singapore: Springer Singapore.

Nallan Chakravartula, S. S., Moscetti, R., Bedini, G., Nardella, M., & Massantini, R. (2022). Use of convolutional neural network (CNN) combined with FT-NIR

spectroscopy to predict food adulteration: A case study on coffee. *Food Control, 135*, 108816.

Nampoothiri, K., Krishnakumar, V., Thampan, P. K., & Nair, M. A. (2019). *The Coconut Palm (Cocos Nucifera L.)--Research and Development Perspectives*: Springer.

Pandiselvam, R., Mahanti, N. K., Manikantan, M. R., Kothakota, A., Chakraborty, S. K., Ramesh, S. V., & Beegum, P. P. S. (2022). Rapid detection of adulteration in desiccated coconut powder: vis-NIR spectroscopy and chemometric approach. *Food Control, 133*, 108588.

Patil, U., & Benjakul, S. (2018). Coconut milk and coconut oil: their manufacture associated with protein functionality. *Journal of food science, 83*(8), 2019-2027.

Pu, Y.-Y., O'Donnell, C., Tobin, J. T., & O'Shea, N. (2020). Review of near-infrared spectroscopy as a process analytical technology for real-time product monitoring in dairy processing. *International Dairy Journal, 103*, 104623.

Stuart, B. H. (2004). *Infrared spectroscopy: fundamentals and applications*: John Wiley & Sons.

Sundaram, J., Kandala, C., & Butts, C. (2010). Classification of in-shell peanut kernels nondestructively using VIS/NIR reflectance spectroscopy. *Sensing and Instrumentation for Food Quality and Safety, 4*(2), 82-94.

Tansakul, A., & Chaisawang, P. (2006). Thermophysical properties of coconut milk. *Journal of Food Engineering, 73*(3), 276-280.

Wattanapahui, S., Suwonsichon, T., Jirapakkul, W., & Kasermsumran, S. (2013). *Prediction of total fat content, lauric acid, palmitic acid and oleic acid of coconut milk products by near infrared spectroscopy (NIRS).* Paper presented at the Proceedings of the 51st Kasetsart University Annual Conference, Bangkok, Thailand, 5-7 February 2013.

Workman Jr, J., & Weyer, L. (2007). *Practical guide to interpretive near-infrared spectroscopy*: CRC press.

Xu, L., Zhou, Y.-P., Tang, L.-J., Wu, H.-L., Jiang, J.-H., Shen, G.-L., & Yu, R.-Q. (2008). Ensemble preprocessing of near-infrared (NIR) spectra for multivariate calibration. *Analytica Chimica Acta, 616*(2), 138-143.

Xu, Y., Zhong, P., Jiang, A., Shen, X., Li, X., Xu, Z., Shen, Y., Sun, Y., & Lei, H. (2020). Raman spectroscopy coupled with chemometrics for food authentication: A review. *TrAC Trends in Analytical Chemistry, 131*, 116017.

CHAPTER 2 – LITERATURE REVIEW
A COMPREHENSIVE OVERVIEW OF NEAR INFRARED AND INFRARED SPECTROSCOPY FOR DETECTING THE ADULTERATION ON FOOD AND AGRO-PRODUCTS—A CRITICAL ASSESSMENT[1]

2.1 Abstract

In the past decade, fast and non-destructive methods based on spectroscopy technology have been studied to detect and discriminate against food adulteration and agro-products. Numerous linear and nonlinear chemometric approaches have been developed for spectroscopy analysis. Recently, various approaches have been developed for spectroscopic calibration modeling to detect and discriminate adulteration food and agro-products. This article discusses the application of spectroscopy technology, including near infrared and infrared, in detecting and discriminating the adulteration of food and agro-products based on recent research and delivered a critical assessment on this topic to serve as lessons from current studies and future outlooks. The current state-of-the-art techniques, including detection and classification of various adulteration in food and agro-products, have been addressed in this paper. Key findings from this study, near infrared and infrared spectroscopy is a non-destructive, rapid, simple-preparation, analytical rapidity, and straightforward method for classification and determination of adulteration in the food and agro-products so it is suitable for large-scale screening and on-site detection. Although there are still some unsatisfactory research results, especially in detecting tiny adductors, these technologies can potentially detect any adulteration in the various food and agro-products at an economically viable level, at least for the initial screening process. In that respect, near infrared and infrared spectroscopy should be expanded to cover all food and agro-products sold in the market. Only

[1]This chapter constituted the publication article: Sitorus, A., & Lapcharoensuk, R. (2022). A comprehensive overview of near infrared and infrared spectroscopy for detecting the adulteration on food and agroproducts—a critical assessment. **INMATEH-Agricultural Engineering**, 67(2). https://doi.org/10.35633/inmateh-67-47.

then will there be an acceptable deterrent in place to stop adulteration activity in widely consumed food and agro-products ingredients.

Keywords: agro-product; food; fraud; near infrared; infrared.

2.2 Introduction

In today's worldwide economy, concerns about food authenticity are a top priority. Customers' primary focus has changed to the originality of food and agroproducts commodities, due to the growing desire for local products (Amirvaresi *et al.*, 2021; Tao *et al.*, 2021; Wongsaipun *et al.*, 2021). As a result, indigenous food and agro-products are frequently chosen over imported ones. Consumers consider freshness and geographical origin when selecting high-quality food products to consume daily, such as meat, flour, flavoring, herbs, and spices.

The increasing population and high cost of produced food and agro-products have created opportunities to use adulteration in postharvest processing. The quality control of these products still relies on laboratory testing based on chemical analysis. Regrettably, these methods seem expensive, complicated to use, usually time-consuming and require a sample preparation step before analysis, in turn, they need many kinds of chemical solvent. In that respect, the option of spectroscopy technology, including near infrared and infrared, offers a valid key to overcoming some of the abovementioned disadvantages since they allow performing a non-destructive evaluation, rapid, easy, eco-friendly, and directly in situ (Galvin-King *et al.*, 2021a; Ndlovu *et al.*, 2019; Silva *et al.*, 2020). This is why researchers have worked over the years to find another application as standard analysis in various fields, especially food science (Ozaki *et al.*, 2021).

According to the recent literature, many studies have been using spectroscopy technology, including near infrared and infrared, to detect and classify the adulteration of food and agro-products. Yet, to date, no comprehensive study has reported on it or provided a critical assessment on this topic. Therefore, the article presents an overview of the application of near infrared and infrared spectroscopy in detecting and discriminating the adulteration of food and agro-products based on recent research.

2.3 Methods

Applications of spectroscopy technology, including near infrared and infrared, to assess fraud, particularly in food and agro-products, have increased each year (Figure 2.1). Research papers were searched in February 2022 via the electronic database Scopus (www.scopus.com). The keyword for finding the research papers using “NIR” or “near-infrared” and “adulteration”. From the first search, research papers can be categorized into an article (447), conference article (56), review (41), book chapter (15), conference review (5) and short survey (1). Most of the articles published come from China (33.6%), followed by Brazil (11.7%), the United States (8.3%), Spain (6.2%), the UK (4.8%), India (4.4%), Italy (4.2%), Ireland (4.1%), Malaysia (3.2%), and France (3.0%). The most popular keywords were infrared device (50.4%), near infrared spectroscopy (50.4%), adulteration (29.9%), least squares approximations (23.7%), chemometrics (20.4%), principal component analysis (19.6%), and spectroscopy, near infrared (18.8%).

Subsequently, the abstracts of the paper were investigated to include or exclude them in this article. From there, 447 documents were further examined, and inappropriate documents were excluded. Excluded research papers were carried out because they did not use near infrared or infrared spectroscopy to detect adulteration, papers that did not use food and agro-products as the main object of the study, conference papers, book chapters, conference reviews, short survey, and review articles. A total of 126 documents were used in the further study. An overview of the research papers is shown in Table 2.1 to Table 2.3.

2.4 NIR and IR Spectroscopy for Food and Agro-product

Infrared (IR) spectroscopy uses the spectral range between 800 and 500,000 nm, which can be further subdivided into the far IR (FIR: 25,000 to 500,000 nm), the mid IR (MIR: 2500 to 25,000 nm), the near IR (NIR: 800 to 2500 nm), and ultraviolet-visible (UV-VIS: 200 to 780) (Ozaki *et al.*, 2021; Reich, 2016). The application of near infrared and infrared spectroscopy for food and agro-products has long been known in the industrial world and continues to expand today (Wesley *et al.*, 1995). In general, this technology is utilized to evaluate food and agro-products in the form of quantitative and qualitative analysis. The wavelengths used vary widely from near

infrared spectroscopy (780–2500 nm) to MIR spectroscopy (2500–25,000 nm) (Alamar *et al.*, 2020; Pereira *et al.*, 2019; Santos *et al.*, 2021). Meanwhile, some researchers combine the wavelength of the near infrared spectroscopy region with the wavelength of the visible region wavelength (340–2500 nm) or commonly known as VIS-NIR spectroscopy (Ndlovu *et al.*, 2021b; Pandiselvam *et al.*, 2022; Valinger *et al.*, 2021b).

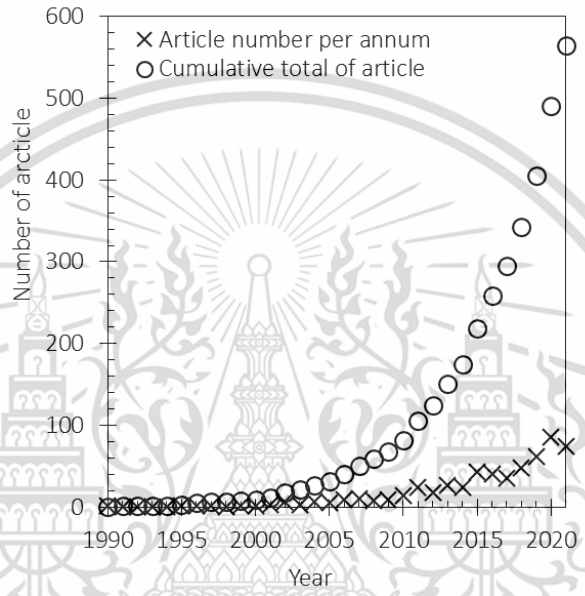


Figure 2.1. Metadata Scopus record of research paper per annum and cumulative total of articles until 2021.

Likewise, several wavelength ranges in near infrared and infrared spectroscopy for food and agro-products that have been studied are shown in Figure 2.2. Unfortunately, although it has limitations in the spectral range, visible near infrared technology (340–780 nm) is still used to detect and discriminate adulteration in food and agro-products. However, full-wavelength near infrared (780–2500 nm) and infrared (2500–16,000 nm) spectroscopy with wider wavelengths are more commonly used for detecting adulterations of food and agro-products. On the other hand, some studies also combine ultraviolet, visible, and near infrared wavelength ranges known as UV-VIS-NIR (325–2500 nm).

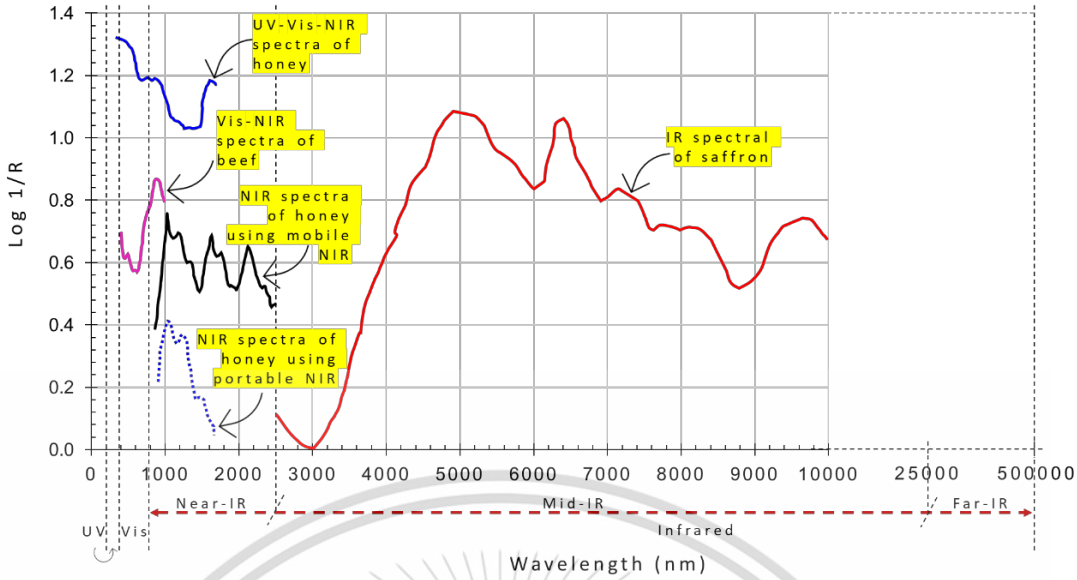


Figure 2.2. Wavelength range of NIR and IR spectroscopy technology.

2.4.1 NIR Spectroscopy Technology (780–2500 nm)

The spectral band represents the interaction of molecules with the near infrared wavelength. The chemical content on the samples tends to absorb specific frequencies of light when a sample is irradiated with near infrared spectroscopy. Thus, near infrared spectroscopy can provide a fingerprint of the content in a sample, especially in food and agro-products. Near infrared spectroscopy has been used in a wide range of investigations to find adulteration in foods and agro-products such as livestock (dos Santos Pereira *et al.*, 2021a; Mabood *et al.*, 2020; Teixeira *et al.*, 2021a), flour (Ayvaz *et al.*, 2021b; Ndlovu *et al.*, 2021a; Tao *et al.*, 2021), liquid agro-product (Du *et al.*, 2021b; Tan *et al.*, 2021; Valinger *et al.*, 2021b), and herbs and spices (Cantarelli *et al.*, 2020; Castro *et al.*, 2021; Rukundo and Danao, 2020).

Near infrared spectroscopy offers a fast, effective, and low-cost alternative procedure that can provide clues about the chemical content and physical properties of the samples. The more affordable near infrared spectroscopy technology is due to the fact that more and more mechatronic industries are developing spectrometer packages that are simpler, more portable, and smaller in size than the benchtop types available in the laboratory. Several studies have reported that it detects adulteration in food and agro-products using portable near infrared spectroscopy in the wavelength range of 908–1676 nm, 950–1650 nm, 1351-

2551 nm and 1600–2400 nm (Aykas and Menevseoglu, 2021; Correia *et al.*, 2018; dos Santos Pereira *et al.*, 2021b; Oliveira *et al.*, 2020; Santos *et al.*, 2013; Silva *et al.*, 2020; Torres *et al.*, 2021). Although many industries have developed near infrared spectroscopy technology packages, unfortunately, they will still be relatively expensive over the next few years. On the other hand, near infrared spectroscopy instruments generate a large amount of data that require an adequate method to build useful analytical information. Combining chemometric and near infrared spectroscopy techniques is required to collect as much associated information from the spectral data as possible (Genis *et al.*, 2021). In this case, chemometrics is the science of extracting information from a chemical system through data-driven methods.

The use of a wider spectral region allowed them to obtain more information related to the stretching and deformation vibrations of the C–H, O–H, and N–H groups that are abundant in a sample. For example, from a honey sample, wavelengths in the visible region up to near infrared (400–2500 nm) are related to those compounds in the honey that absorb in the blue-violet range, giving the characteristic orange-amber color of the honey (Yang *et al.*, 2020). In the near infrared region, the wavelength at 1451 nm is related to the first overtone of the vibrational mode of the O–H stretch from water (Huang *et al.*, 2020a). Therefore, signal regions of near-infrared and infrared spectra are needed to understand the compound in the samples with greater precision. With that in mind, the next step is to focus only on the few wavelength regions that can provide the information that correlates with the compounds in our sample. In addition, portable near infrared spectroscopy with a narrow wavelength region can be utilized, while providing high accuracy.

2.4.2 IR Spectroscopy Technology (2500–16,000 nm)

Infrared spectroscopy data cover the 2500 to 16,000 nm range used to represent fundamental vibrations, molecular overtones, and combination vibrations. The absorption areas are predominantly composed of hydrogen-containing groups related to the acid, oil content, protein, sugar, and water of food and agro-products. Consequently, the spectral contains chemical information by reflecting the molecular structures from the samples.

Several recent studies have been carried out using infrared spectroscopy technology to detect and discriminate adulteration of food and agricultural products for livestock products, including milk and eggs (Botelho *et al.*, 2015; Hosseini *et al.*, 2021; Uysal and Boyaci, 2020). In addition, flour products have been investigated for products including pistachios and peppers (Aykas and Menevseoglu, 2021; Galvin-King *et al.*, 2020a). Liquid products have also been studied for products including yogurt, guava pulp, durum wheat pasta, and butter oil (Alamar *et al.*, 2020; De Girolamo *et al.*, 2020b; Pereira *et al.*, 2019; Temizkan *et al.*, 2020b). For herbs and spices, products have been studied, including those of black pepper, garlic, and saffron (Amirvaresi *et al.*, 2021; Galvin-King *et al.*, 2021a; Wilde *et al.*, 2019). Nevertheless, the most challenging thing for researchers in adulteration studies in this range spectral is to explain the connection between absorption in the spectral region with the chemical content of food and agro-products. Occasionally, the various intrinsic properties to be determined usually lead to non-linear patterns. Finally, many linear and non-linear chemometric approaches have been developed for quantitative and qualitative analyses to tackle this problem.

2.5 Analysis Data

Spectral data analysis is the most important part of obtaining the information contained therein. In general, the procedure that must be followed in extracting the information in the near infrared and infrared spectra, especially related to the purity of food and agro-products, is presented in Figure 2.3. Food and agro-products that have been adulterated with an adulterating agent will create different infrared spectra data as a result of the various functional groups in the material. However, this will not necessarily produce information without developing a calibration model, which is followed by testing to build a predictive model. Furthermore, the predictive model performance should also be tested with several unknown datasets to create a proven model.

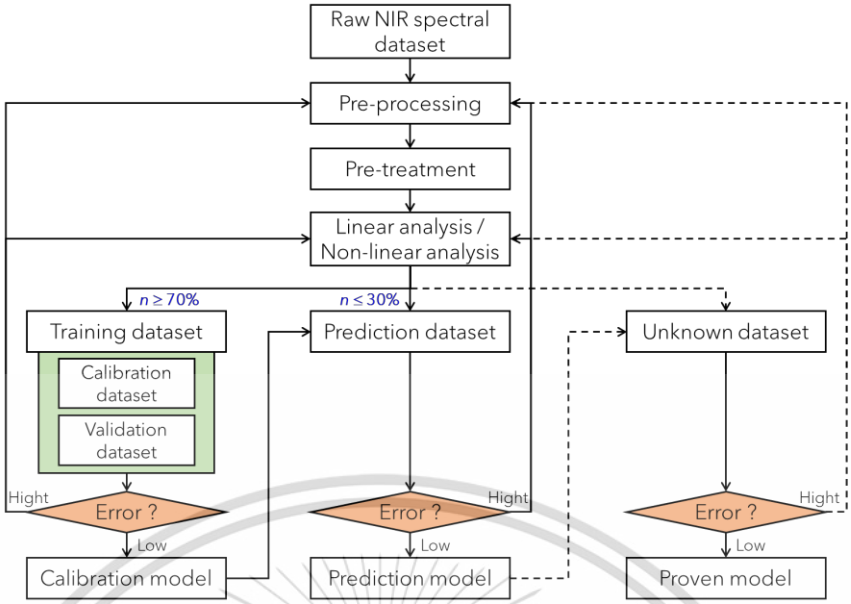


Figure 2.3. Procedure of model construction and performance evaluation.

In many cases of adulteration of food and agro-products, the processing and pretreatment steps are very important to reduce noise spectra data. Furthermore, many linear and nonlinear chemometric approaches, including partial least squares regression (PLSR), principal component regression (PCR), support vector machine (SVM), and artificial neural network (ANN), have been developed to quantify the physical and chemical properties of food and agricultural products to acquire information from spectral data. The last two algorithms are the newest, along with the k-nearest neighbor (k-NN), the convolutional neural network (CNN), and the radial basis function neural networks (RBFNN) based on machine learning, which are reported to produce the best predictive models compared to PLSR and PCR (Alamar *et al.*, 2020; Liu *et al.*, 2021; Xie *et al.*, 2008).

2.5.1 Preprocesing Data

The difficulty of using spectral data for food and agro-products quality assessment stems from the need for a strong and accurate model with low sensitivity and low-intensity spectral data. Almost all studies involving near infrared and infrared spectroscopy use preprocessing data to avoid noise from light scattering, instrumental drift, particle size variation, and also high overlaps between combination bands and overtones to address this problem. Preprocessing is a method used to go from raw data to clean data ready for analysis including removing

baseline artifacts, peak selection, or alignment. Pretreatment is to transform the preprocessed data to make them suitable for analysis, including normalization, scaling, transformations, and removing any outliers in the data.

The application of preprocessing does not always provide the best results. For example, Valinger *et al.* (2021b) did not apply preprocessing or pretreatment to its spectral data. However, they could provide an RPD value greater than 3 using the PLSR algorithm to detect fructose corn in honey. However, Santos *et al.* (2021) reported that preprocessing of SNV to detect adulteration of cocoa solids gave better results than without the application of preprocessing. Therefore, we conclude that applying preprocessing to near infrared and infrared spectroscopy data is a procedure that must be tested regardless of the results obtained.

2.5.2 Linear Approach

A linear approach in near infrared and infrared spectroscopy data analysis will be successful if a linear association exists between the absorbance spectra and predicted content, more commonly referred to as the Beer-Lambert law. It is capable of conducting qualitative and quantitative analyses of adulteration in food and agro-products. The linear chemometric methods that were used most frequently to formulate a qualitative and quantitative analysis of adulteration in food and agro-products were PLSR, PCR, partial least squares discriminant analysis (PLS-DA) and principal component analysis-linear discriminant (PCA-DA) (Gayo *et al.*, 2006; Kazazić *et al.*, 2021; Paradkar *et al.*, 2002a).

In general, linear chemometric methods from IR spectroscopic data can be evaluated with several parameters. The parameters most used, including calibration and cross-validation (CV), are the determination coefficients (R^2), the coefficients correlation (r), the root mean square error (RMSE) and the standard error (SE). In addition, some use difference average value between predicted and measured values (Bias), range error ratio (RER), and predicted deviation ratio (RPD). Each parameter has its own purpose in evaluating the model. Coefficient determination indicates how well a model performs in terms of the proportion of variance in the dependent variable predicted by the independent variables. The RPD shows the robustness of the model. SE and RMSE indicate the level of precision and accuracy of the developed model.

2.5.3 Non-linear Approach

Another method to analyze near infrared and infrared spectroscopy data of adulteration in food and agro-products associated with chemometrics is a non-linear approach. This approach is required when the connection between the spectral absorption region of the IR spectroscopy is non-linear. The origin of these non-linear relationships is diverse and challenging to identify, but according to Ramírez-Morales *et al.* (2016), in some cases, due to the disparities in viscosity, temperature, pH, particle dimensions, and chemical content. Calibration is generally achieved utilizing non-linear methods and multivariate analysis for this reason. A reasonable variable selection aimed at collecting a small sub-group with lower sensitivity to non-linear or excluding the most wavelengths is usually effective in enhancing the model's performance (Kaufmann *et al.*, 2022; Pandiselvam *et al.*, 2022).

The research that applies a non-linear approach in chemometrics for detecting and authenticating adulteration on food and agro-product is currently in constant expansion. As mentioned before, a non-linear approach to analyzing near infrared and infrared spectroscopy data can also perform qualitative analysis and quantitative prediction of adulteration in food and agro-products. Machine learning-based chemometric research is rapidly expanding at the moment. ANN, CNN, k-NN, RBFNN, RF, SVM are also more reported to analyze IR spectroscopy data of adulteration in food and agro-product as these techniques are based on pattern recognition (Ding and Xu, 2000; Le Nguyen Doan *et al.*, 2021; Weng *et al.*, 2020).

2.6 Some Case Adulteration on Food and Agro-products

Near infrared and infrared spectroscopy analysis has been applied to both detecting and discriminating adulteration of food and agro-products. Qualitative evaluation can be the detecting of adulteration in livestock products, flour products, liquid agro-product, and herbs and spices (Table 2.1). In contrast, the quantitative study concentrates on predicting multiple contents adulteration of food and agro-products has been reported quite a lot recently (Table 2.2). In the present studies, various IR spectroscopy ranges are utilized for the quantitative and qualitative analysis of food and agro-products, including near infrared and infrared spectroscopy data (Table 2.3).

Table 2.1. Some qualitative study of food and agro-product adulteration.

#	Source	Objective (Sample number)	Adulterant material	Range of spectral (nm)	The best of		Prediction results
					Pretreatment	Algorithm	
1	de Araújo *et al.* (2021)	Gourmet ground roasted coffees (90)	Traditional and superior coffees	1205 – 2128	Offset correction	SIMCA	Specificity = 100%
2	Srinuttrakul *et al.* (2021)	Hom Mali rice (170)	Rice from northern and northeastern regions of Thailand	740 – 1070	MSC+ 1st dev	PLS-DA	Accuracy =84.85 – 86.96%
				2500 – 22222			Accuracy = 96.97 –100%
3	Tan *et al.* (2021)	Stingless bee honey (30)	High fructose corn syrup	900 – 1700	Cutting + Gaussian smoothing	LR	Accuracy = 98.2%
4	dos Santos Pereira *et al.* (2021a)	Goat milk (146)	Cow milk	900 – 1650	Moving mean + Baseline offset	iSPA-PLS-DA	Accuracy = 98.3%
5	Shannon *et al.* (2021)	Basmati rice (1399)	Other varieties basmati rice	740 – 1070	Raw	PLS-DA	F1_score = 0.93
6	Tao *et al.* (2021)	Wheat flour (48)	Eight varieties of cassava flour	1150 – 2150	Raw	PLS-DA	Accuracy = 97.53%
7	Galvin-King *et al.* (2021b)	Garlic (117)	12 types of white powder	833 – 2500	SNV + 1st dev SG	OPLS-DA	Youden index = 0.98
				2500 – 18182			Youden index = 1
8	Teixeira *et al.* (2021b)	Yogurt and Cheese from goat milk (576)	Cow milk	1000 – 2500	Smoothing + 2nd dev SG	PLS-DA	Sensitivity = 99.2 – 100% Specificity = 99.2 – 100%
9	Torres *et al.* (2021)	Sweet almonds (216)	Bitter almonds	950 – 1650	SNV + 1st dev SG	PLS-DA	Non-error rate = 86 – 100%
10	Le Nguyen Doan *et al.* (2021)	High-quality rice (200)	Low-quality rice	740 – 1070	1st dev SG + mean centered	PLS-DA	Accuracy = 82.6%
11	Cantarelli *et al.* (2020)	Cinnamon verum (120)	Cinnamon cassia	940 – 1640	Raw	PNN	Accuracy = 99.25%
12	Huang *et al.* (2020b)	Honey (224)	Syrup	1000 – 2500 2222 – 12500	2nd dev SG	SVMC	Accuracy = 100%
13	Galvin-King *et al.* (2020b)	Powdered paprika (159)	Varying seed/pod	833 – 2500	SNV + 1st + 2nd dev SG	OPLS-DA	R2 = 0.85
				2500 – 18182	SNV + 1st dev SG		R2 = 0.94
14	Alamar *et al.* (2020)	Guava pulp (240)	Sugar and water	1000 – 2500	MSC	k-NN	Accuracy = 100%
				2500 – 25000			Accuracy = 100%
15	De Girolamo *et*	Durum wheat	Durum wheat pasta from	1000 –	Mean baseline +	PLS-DA	Accuracy =

	al. (2020a)	pasta from Italy (280)	Argentina	2500	detrending		97 – 100%
				2500 – 25000	MSC + detrending		Accuracy = 96 – 97%
16	Teixeira *et al.* (2020)	Goat milk (600)	Water, urea, bovine whey, and cow's milk	1000 – 2500	1st dev SG + SNV	PLS-DA	Precision = 100%
17	Visconti *et al.* (2020)	Grated cheese (196)	Microcrystalline cellulose, silicon dioxide, wheat-flour, wheat-semolina, sawdust	1000 – 2500	1st dev SG	PLS-DA	Precision = 100%
18	Jahani *et al.* (2020)	Lime juices (56)	Water and citric acid	900 – 1700	MSC	k-NN	Precision = 100%
19	Wilde *et al.* (2019)	Black pepper (126)	papaya seeds, chili and non-functional black pepper material	833 – 2500	SNV + 1st dev SG	OPLS-DA	Precision = 90 – 100%
				2500 – 25000			Precision = 92 – 100%
20	Karunathilaka *et al.* (2018)	Milk powder (383)	11 potential adulterants	800 – 2500	SNV + 1st dev SG	SIMCA	Accuracy = 100%
21	Chen *et al.* (2017)	Milks (102)	Melamine	1000 – 2500	SNV	OC-PLS	Accuracy = 89%
22	Shen *et al.* (2016)	Soybean meal (88)	Six types of non-protein nitrogen	1282 – 2500	1st dev SG + SNV	PLS-DA	Sensitivity = 100%
23	Ziegler *et al.* (2016)	Wheat kernels and flours (1225)	Bread wheat, spelt, durum, emmer, and einkorn	1200 – 2400, 650 – 2500	1st dev SG	PLS-DA	Accuracy = 80 – 100%
24	Xu *et al.* (2015)	Tea (100)	Exogenous amino acids	833 – 2500	SNV	PLS-DA	Accuracy = 0.936
25	Schmutzler *et al.* (2015)	Pork meat (84)	Pork fat	833 – 2500	2nd dev SG	SVMC	Accuracy = 83.3%
26	Botelho *et al.* (2015)	Raw cow milk (155)	Water, starch, sodium citrate, formaldehyde, and sucrose	2500 – 16667	1st dev SG + Smoothing	PLS-DA	Sensitivity = 88.5 – 100%
27	Ding *et al.* (2015)	Sweet potato powder (116)	purple and white sweet potato	700 – 2500	Selection wavelength using GA-PLS	LDA	Accuracy = 100%
28	López *et al.* (2014)	Hazelnut paste (135)	Almond paste and Chickpea flour	1000 – 2740	Offset correction	SIMCA	Accuracy = 96.3%
29	Zhang *et al.* (2014)	Raw cow milk (800)	pseudo proteins (urea, ammonium nitrate, melamine) and thickeners (dextrin and Starch)	1000 – 2500	SNV	SVMC	Precission = 96.62%
30	Xu *et al.* (2013a)	Chinese glutinous rice flour (215)	Extraneous adulterants, unwanted variations	1000 – 2500	2nd dev SG	OC-PLS	Specificity = 0.92
31	Xu *et al.* (2013b)	Chinese yogurt (257)	Edible gelatin, industrial gelatin, soy protein powder	833 – 2500	SNV	OC-PLS	Specificity = 0.95
32	Xu *et al.* (2013c)	Lotus root powder (85)	Four cheaper starches	833 – 2500	SNV	SIMCA	Specificity = 0.94
33	Chen *et al.* (2011)	Honey (144)	High fructose corn syrup	1000 – 2500	1st dev SG + smoothing +	PLS-DA	Accuracy = 96.88%

					mean centering		
34	Zhu *et al.* (2010)	Honey (135)	Sweeteners materials	1000 – 2500	SNV + Smoothing SG	SVM	Accuracy = 95.1%
35	Xie *et al.* (2008)	Pure bayberry Juice (129)	Water	800 – 2400	SNV	RBFNN	Accuracy = 97.62
36	Downey *et al.* (2003)	Honey (300)	Fructose and glucose	400 – 2498	2nd dev SG	PLS-DA	Accuracy = 96%

1st dev SG=First derivatives Savitzky-Golay; 2nd dev SG=Second derivatives Savitzky-Golay; iSPA-PLS-DA=Intervals SPA-partial least squares-algorithm discriminant analysis; k-NN=k-nearest neighbor; LDA=Linear discriminant analysis; LR=Logistic regression; MSC=Multiplicative scatter correction; OC-PLS=One class-partial least squares; OPLS-DA=Orthogonal partial least squares-discriminant analysis; PLS-DA=Partial least squares-discriminant analysis; PNN=Probabilistic neural network; RBFNN=Radial basis function neural networks; SNV=Standard normal variate; SIMCA=Soft independent modelling of class analogy; SVMC=Support vector machines classification.

Table 2.2. Some quantitative study of food and agro-products adulteration.

#	Source	Objective (Sample number)	Adulterant material	Range of spectral (nm)	The best of		Prediction results
					Pretreatment	Algorithm	
1	Ndlovu *et al.* (2021a)	Green banana flour (72)	Wheat flour	400 – 2500	SNV + Baseline	PLSR	RPD = 3.9
2	Ndlovu *et al.* (2021b)	Green banana flour (66)	Wheat flour	400 – 2500	2nd dev + Detrend	PLSR	RPD = 6.24
3	Ayvaz *et al.* (2021b)	Einkorn flour (64)	Wheat flour	1000 – 2500	MN + MSC + 1st dev	PLSR	RPD=19.3
4	Santos *et al.* (2021)	Cocoa solids (110)	Cocoa solids content	1100 – 2500	SNV	PLSR	RPD = 31.09
				2500 – 16667			RPD = 17.28
5	Valinger *et al.* (2021b)	Acacia honey (135)	Fructose corn syrup	325 – 900; 904 – 1699	Raw	PLSR	RPD = 3.32
6	Wongsaipun *et al.* (2021)	Thai Jasmine Rice (423)	3 type rice	400 – 2498	Normalization	PLSR	RMSEP = 2.6; R^2p = 0.98
7	Castro *et al.* (2021)	Saffron (38)	Onion, Calendula, Pomegranate and Turmeric	1000 – 2500	2nd dev SG + SNV	MCR-ALS	RMSEP = 0.8 – 2.3
8	Liu *et al.* (2021)	Infant formula (200)	Hydrolyzed leather protein and melamine	900 – 1700	1st dev	CNN	R^2p=0.96 – 0.99
9	Aykas and Menevseoglu (2021)	Powdered Pistachio (19)	Powdered green pea and peanut	2500 – 15385	2nd dev SG + Smoothing	PLSR	rval = 0.99
				1351 – 2551			rval = 0.99
10	Masithoh *et al.* (2021)	Arenga pinnata sugar (187)	Coconut sugar	1000 – 2500	MSC	PLSR	RMSEP = 12.42
				2500 – 15385	Normalization		RMSEP = 6.95
11	Genis *et al.* (2021)	Pistachio nut (143)	Green pea and spinach nut	908 – 1695	Raw	PLSR	RMSEP = 4.69 – 7.87
12	Silva *et al.* (2020)	Ground meat	Beef, pork	908 –	1st dev SG + MSC	SVMR	RMSEP = 3.5

		chicken (150)		1676			– 4.7
13	Yang *et al.* (2020)	Manuka honey (93)	Five different syrups	400 – 2500 1100 – 2500	2nd dev SG	PLSR	RMSEP = 3.61
14	Rukundo *et al.* (2020)	Dried turmeric powder (120)	Metanil yellow	780 – 2500	1st dev SG	PLSR	RPD = 10.3
15	Uysal and Boyaci (2020)	Liquid egg (100)	Water	1000 – 2500	Baseline, autoscale, smoothing, 1st dev SG	PCR	RMSECV = 0.8 – 0.74
				2500 – 25000		PCR	RMSECV = 0.12 – 17.4
16	Ndlovu *et al.* (2019)	Unripe banana flour (82)	Wheat flour	447– 1005	2nd dev SG	PLSR	RPD = 12.02
17	Kar *et al.* (2019)	Turmeric powder (200)	Corn starch	1000 – 2500	SNV + 1st dev SG	PLSR	RMSEP = 0.26; R^2p = 0.99
18	Pereira *et al.* (2019)	Butter oil (33)	Soybean oil	833 – 2500	Raw	PLSR	RPD = 21.68
				2500 – 25000			RPD = 12.27
19	Yasmin *et al.* (2019)	Cinnamon Powder (195)	Lower quality cinnamon Powder	1000 – 2500	2nd dev SG	PLSR	R^2p = 0.97; RMSEP = 2.2
				2857 – 15385			R^2p = 0.96; RMSEP = 2.5
20	Lukacs *et al.* (2018)	Whey protein powder (279)	Urea, L-taurine, L-histidine	800 – 2750	Smoothing, SNV, 2nd dev SG	PLSR	$R^2p > 0.98$
21	Da Silva Dias *et al.* (2018)	Raw milk (50)	Water	1200, 1450, 1530,	Raw	MLR	R^2_p = 0.96, RMSEP = 0.018
22	Picouet *et al.* (2018)	Sunflower oil (138)	Mineral oil	1000 – 2200	Baseline, MSC, SNV	PLSR	RMSEP = 0.23 – 1.26
23	Kar *et al.* (2018)	Turmeric Powder (248)	Metanil yellow	1000 – 2500	1st dev SG	PLSR	R^2p = 0.91
24	Correia *et al.* (2018)	Arabica coffee (125)	Robusta coffee, corn, peels, and sticks	908 – 1676	1st dev SG	PLSR	RPD = 64.23
25	Liu and Zhou (2017)	Apple juice (31)	Water	830 – 2490	MSC	SPA-PSO-PLS	R^2p = 0.99; RMSEP = 0.063
26	Bázár *et al.* (2016)	Honey (492)	High fructose corn syrup	1100 – 2500	Smoothing + SNV + 2nd dev SG	PLSR	R^2cv = 0.987; RMSECV = 1.48
27	Dvorak *et al.* (2016)	Goat milk for cheeses (48)	Cow's milk	1000 – 2500	Raw	PLSR	R^2cv = 0.783
28	Winkler-Moser *et al.* (2015)	Coffea arabica (84)	Corn	400 – 2500	1st dev SG	PLSR	R^2cv = 0.974
29	Kumaravelu and Gopal (2015)	Honey (160)	Jaggery	400 – 2500	Smoothing + SNV	PLSR	R^2_p = 0.99
	Mouazen and Al-Walaan (2014)	Honey (345)	Glucose syrup	305 – 2200	SNV + 1st dev SG + Smoothing	PLSR	R^2p = 0.78, RPD = 2.06
30	Lohumi *et al.* (2014)	Onion powder	Corn starch	1000 –	SNV	PLSR	R^2p = 0.90

31		(180)		2500			
				2500 – 15385			R^2p = 0.98
32	Vichasilp and Poungchompu (2014)	Beaf and chicken Meatballs (140)	Pork meat	1000 – 2500	Raw	PLSR	R^2v = 0.88 – 0.83
	Wang *et al.* (2014)	Oat flour (220)	Wheat flour	833 – 2500	2nd dev SG	PLSR	RMSEP = 1.975
33	Santos *et al.* (2013)	Bovine milk (744; 372 – 837)	Tap water, whey, synthetic milk, synthetic urine, urea, and hydrogen peroxide	1600 – 2400	Raw	PLSR	R^2v = 0.92
34				2500 – 15385			R^2v = 0.92 – 0.98
35	Öztürk *et al.* (2010)	Olive oil (160)	Soybean, cotton, corn, canola and sunflower oils	1000 – 2500	Raw	GILS	SEP = 2.93 – 5.86 rv = 0.90 – 0.99
36	Mishra *et al.* (2010)	Honey (56)	Jaggery syrup	1380 – 1960	Raw	PLSR	R^2v = 0.66
37	Pizarro *et al.* (2007)	Arabica coffea powder (191)	Robusta coffea powder	1100 – 2500	1st dev SG + OWAVEC	PLSR	R^2p = 1
38	Özdemir and Öztürk (2007)	Olive oil (52)	Sunflower and corn oil	1000 – 2500	Raw	GILS	R^2p = 0.99
39	Gayo and Hale (2007)	Atlantic blue crabmeat (110)	Blue swimmer crabmeat	400 – 2498	1st dev SG	PLSR	R^2p = 0.98
40	Cocchi *et al.* (2006)	Durum wheat flour (58)	Bread wheat flour	400 – 2498	SNV	PLSR	RMSEP = 0.38
41	Gayo *et al.* (2006)	Crab meat (66)	Surimi-based imitation crab meat	400 – 2498	1st dev SG	PCR	R^2p = 0.99; SEP = 0.24
42	Jha and Matsuoka (2004)	Cow Milk (125)	Urea, NaOH, Oil, shampoo	700 – 1124	MSC	MLR	R^2v = 0.58 – 0.98
43	Uddin and Okazaki (2004)	Fresh (162)	Frozen-thawed fish	1920 – 2350	2nd dev SG	MLR	R^2c = 0.95 – 0.99
44	Maraboli *et al.* (2002)	Milk powder (155)	Vegetable proteins	1100– 2500	1st dev SG	MLR	R^2p = 0.993
45	Rodriguez-Saona *et al.* (2001)	Fruit juices (60)	Sugars	1000 – 2500	2nd dev SG	PLSR	R^2p = 0.99
46	Wesley *et al.* (1995)	Olive oil (310)	Corn oil, sunflower oil, raw olive residue oil	800 – 2500	1st dev SG	PLSR	rv = 0.8

CNN=Convolutional neural network; GILS=Genetic inverse least squares; MCR-ALS=Multivariate curve resolution-alternating least squares; MLR=Multiple linear regression; PCR=Principal component regression; PLSR=Partial least squares regression; SVMR=Support vector machines regression

Table 2.3. Combine qualitative and quantitative analysis of food and agro-products adulteration.

#	Source	Objective (Sample number)	Adulterant material	Range of spectral (nm)	The best of		Prediction results
					Pretreatment	Algorithm	
1	Kazazić *et al.*	Butter (36)	Pork fat, Margarine	900 – 1700	Raw	PLS-DA	Accuracy =

	(2021)						100%
						PLSR	RPD = 5.24 – 37.51
2	Amirvaresi *et al.* (2021)	Saffron (120)	C. sativus style, safflower, rubia and calendula	833 – 2500	MN + 2nd dev	PLS-DA	Acuracy = 95.4 –100%
						PLSR	R^2 = 0.95 – 0.99
				2500 – 25000		PLS-DA	Acuracy = 81.3 – 100%
3	Hosseini *et al.* (2021)	Sterilized milk (11)	Sodium dodecyl sulfate	769 – 2500	MN + SD scaled	PLS-DA	R^2cv = 0.98
					2nd dev+SNV	PLSR	R^2p = 0.96
				2500 – 16667	Smoothing SG	PLS-DA	R^2cv = 0.94
						PLSR	R^2p = 0.98
4	Du *et al.* (2021a)	Camellia oil (130)	Corn oil, rapeseed oil and sunflower oil	1000 – 2381	1st dev SG	DA	Accuracy = 96.7%
					SNV + 1st dev SG	PLSR	RMSEP = 4.98
5	Le Nguyen Doan *et al.* (2021)	Green tea (475)	Sugar and glutinous rice flour	900 – 1700	SNV	SVMC	Accuracy = 97.47%
						SVMR	rp > 0.94
6	Vitalis *et al.* (2020)	Tomato paste (57)	Ground paprika seed, Corn starch, Sucrose, Salt	740 – 1700	1st dev SG + MSC	LDA	Precission = 78.64% – 97.65%
						PLSR	RMSECV = 0.23 – 0.89
7	Temizkan *et al.* (2020a)	Yoghurt (100)	Several fat-free UHT	1000 – 2500	MN + MSC	SIMCA	Specificity = 100%
					MN + 1st dev SG + MSC	PLSR	RPD = 4.35
				2500 – 15385	MN + 2nd dev SG	SIMCA	Specificity = 100%
						PLSR	RPD = 4.65
8	Mabood *et al.* (2020)	Fresh milk samples (162)	Urea	1000 – 2500	Baseline	PLS-DA	R^2 = 0.97
						PLSR	R^2 = 0.98
9	Leng *et al.* (2020)	Minced beef (150)	Pork and Duck meat	800 – 1852	Raw	DA	Accuracy = 91.5 –100%
					Raw	PLSR	RMSEP = 7.27 – 9.27
10	Pereira *et al.* (2020)	Goat milk (112)	Cow milk	1000 – 2500	Raw	PLS-DA	Accuracy = 100%
					Moving mean + Baseline offset	SPA	RPD = 10
11	Weng *et al.* (2020)	Minced beef (240)	Beef loin, beef heart, beef tallow, and pork loin	1000 – 2500	SG smoothing	CNN	Accuracy = 99%
					CARS	RF	RMSEP = 2.145
12	Biancolillo *et al.* (2020)	Egg pasta (100)	Turmeric	1000 – 2500	MSC	PLS-DA	Precision = 97.5%
					SNV	PLSR	RMSEP = 0.11
13	Oliveira *et al.* (2020)	Paprika powder (315)	Potato starch, acacia gum and annatto	900 – 1700	Auto-scaling	PLS-DA	Specificity = 90%
					Smoothing + 1st dev SG	PLSR	RMSEP = 0.95 – 1.74

14	Kene Ejeahalaka and On (2020)	Fat-filled milk powder (150)	Melamine, urea and 4 different vegetable oils	850 – 2500	2nd dev SG + EMSC	SIMCA	Sensitivity = 85%
						PLSR	R^2p = 0.96
15	Lima *et al.* (2020)	Black pepper and Cumin (130)	Starch cassava, corn flour	1100 – 2500	Raw	O-PLS-DA	Specificity = 100%
						PLSR	RPD = 2.24 – 7.01
16	Aliaño-González *et al.* (2019)	Honey (68)	Inverted sugar, rice syrup, brown cane sugar and fructose syrup	400 – 2500	Raw	LDA	Precision = 100%
						PLSR	RMSEP = 3.89
17	Zaukuu *et al.* (2019)	Paprika powder (54)	Corn flour	750 – 1700	Smoothing + MSC	LDA	Accuracy = 95.55%
						PLSR	R^2cv = 0.98; RMSECV = 1.71
18	Ferreiro-González *et al.* (2018)	Honey (22)	High fructose corn syrup	400 – 2500	Raw	PCA-LDA	Accuracy = 100%
						PLSR	R^2p = 0.99, RMSEP = 4.71
19	Quelal-Vásconez *et al.* (2018)	Cocoa powder (234)	Carob flour	1100 – 2500	2nd dev SG + OSC	PLS-DA	Accuracy = 100%
					OSC	PLSR	R^2p = 0.97, RMSEP = 3.2
20	Mabood *et al.* (2018)	Fruit juice (198)	Saccharin	1000 – 2500	Baseline correction + Smoothing SG	PLS-DA	R^2cv = 0.98
						PLSR	R^2p = 0.97
21	Rady and Adedeji (2018)	Minced beef (1697)	Another beef	200 – 1100, 900 – 1700	Normalization + 1st dev SG	SVMC	Precission = 100%
						PLSR	RPD = 1.64 – 1.98
22	Mabood *et al.* (2017b)	Camel milk (54)	Cow milk	1000 – 2500	1st dev SG	PLS-DA	R^2 = 0.97
						PLSR	R^2 = 0.92; RMSEP = 1.32
23	Mabood *et al.* (2017a)	Camel milk (54)	Goat milk	700 – 2500	Baseline correction + Smoothing SG	PLS-DA	R^2 = 0.97
						PLSR	R^2 = 0.94
24	Liu *et al.* (2017)	Honey (360)	High-fructose corn syrup, maltose syrup	1000 – 2500	Norris + 2nd dev	PLS-DA	Accuracy = 86.3% – 96.1%
					Norris + 1st dev	PLSR	R^2p = 0.9 – 0.98
25	Liu and Zhou (2017)	Infant formula (170)	Hydrolyzed leather protein powder	900 – 1700	MSC + 1st dev SG	SIMCA	Accuracy = 98.21%
						SVMR	RPD = 7.42
26	Alamprese *et al.* (2016)	Minced beef meat (198)	Turkey meat	800 – 2667	SNV	PLSDA	Sensitivity = 0.84
						PLSR	R^2p = 0.884; RMSEP = 10.8
27	Capuano *et al.* (2015)	Skim milk powder (384)	Whey, starch, maltodextrin,	400 – 2498	SNV + 2nd dev SG + mean centering	SIMCA	Accuracy = 82.42%
						PLSR	R^2_p = 0.93 – 0.98

28	Kuswandi *et al.* (2015)	Beef meatball (162)	Pork meat	850 – 2000	1st dev SG	LDA	Accuracy = 100%
						PLSR	R^2_p = 0.97
29	Luqing *et al.* (2015)	Roasted green tea (150)	Sugar and glucose syrup	800 – 2500	SLB, Min/max	PLS-DA	Accuracy = 96 – 100%
					SNV	PLSR	R^2p = 0.99
30	Teye *et al.* (2014)	Fermented cocoa beans (132)	Unfermented cocoa beans	1000 – 2500	SNV	SVMC	Accuracy = 100%
					Selection wavelength using Si-PLS	PLSR	rp = 0.98; RMSEP = 1.68
31	Alamprese *et al.* (2013)	Minced beef (242)	Turkey meat	800 – 2667	SNV	LDA	Accuracy = 71.2%
						PLSR	R^2 = 98.13
32	Morsy and Sun (2013)	Minced beef (191)	Pork, fat trimming and offal	400 – 2500	2nd dev SG, SNV, Moving average	PLS-DA	Accuracy = 100%
						PLSR	R^2p = 0.82 – 0.96
33	Zhao *et al.* (2013)	Beefburger (164)	Offal	850 – 1098	2nd dev SG, MSC,	PLS-DA	Accuracy = 88.9 – 95.5%
					Raw	PLSR	RPD = 1.5 – 2.3
34	Liu *et al.* (2010)	Fishmeal (276)	Melamine	833 – 2500	2nd dev SG + Smoothing	PLS-DA	Accuracy = 99.5%
					1st dev SG + Smoothing + SNV	PLSR	R^2p = 0.98 – 0.99; RMSEP = 0.38 – 0.24
35	Kasemsumran *et al.* (2007)	Cow milk (90)	Water and Whey	1100 – 2500	MSC + 2nd dev SG	PLS-DA	Accuracy = 86.73 – 100%
					MSC	PLSR	R^2 = 0.99
36	Kelly *et al.* (2006)	Honey (179)	Beet invert syrup and High fructose corn syrup	1100 – 2498	Raw	SIMCA	Accuracy = 100%
					MSC, 2nd dev SG	PLSR	R^2 = 0.72 – 0.79
37	León *et al.* (2005)	Apple Juice (450)	Fructose, glucose, sucrose	400 – 2498	MSC	PLS-DA	Accuracy = 86 – 100%
						PLSR	r = 0.77 – 0.94
38	Downey and Kelly (2004)	Strawberry and raspberry purees (305)	Apples purees	400 – 2498	SNV + 2nd dev SG	SIMCA	Accuracy = 75.1–95.1%
						PLSR	rcv = 0.90
39	Paradkar *et al.* (2002b)	Maple syrup (272)	Cane and beet invert syurps, cane and beet sugar solutions	1100 – 1660	1st dev SG	PLS-DA	Accuracy = 98.39%
						PLSR	R^2v = 0.83 – 0.98
				2500 – 25000		PLS-DA	Accuracy = 100%
						PLSR	R^2v = 0.99
40	Contal *et al.* (2002)	Strawberry and raspberry purees (344)	Apples purees	400 – 2500	Raw	SIMCA	Accuracy = 79.07 – 94.77
						PLSR	rv = 0.98 – 0.99
41	Paradkar *et al.* (2002a)	Maple syrup (54)	Corn syurps	2500 – 25000	Raw	PCA-DA	Accuracy = 96.20

						PLSR	R^2p = 0.98
42	Murray *et al.* (2001)	Fish meal (136)	Meat and bone meal	1100 – 2500	MSC	PLS-DA	Accuracy = 98.55%
					2^{nd} dv SG + SNV	PLSR	R^2 = 0.94
43	Ding and Xu (2000)	Beef hamburgers (194)	Mutton, pork, skim milk powder, or wheat flour	400 – 2500	SNV + 2^{nd} dev SG	k-NN	Accuracy = 92.7%
						PLSR	R^2v = 0.74 – 1
44	Thyholt *et al.* (1997)	Beef (350)	Pork, mutton	780 – 2500	1^{st} dev SG + Smoothing	QDA	Accuracy = 98.53 – 100%
						PLSR	r = 0.68 – 0.94

O-PLS-DA=Orthogonal-partial least squares-discriminant analysis; PCA-LDA=Principal component analysis-linear discriminant analysis; QDA=Quadratic discriminant analysis; RF = Random forest; SPA = Successive projections algorithm

2.6.1 Aduteration in Livestock Products

Adulteration of livestock products occurs often and considerably threatens human health and safety when other substances are added for specific purposes. Liu *et al.* (2021) reported machine learning in the form of a CNN architecture in tandem with near infrared spectroscopy data to predict hydrolyzed leather protein and melamine in infant formula. Their result can predict adulterated and unadulterated milk R^2 up to 0.99%. Furthermore, Mabood also developed a method using near infrared spectroscopy in tandem with multivariate analysis to detect the mixture of camel milk with goat milk. They used PLS-DA to authenticate pure and adulterated milk and PLS to quantify adulteration levels with RMSE of 0.08% and 1.10%, respectively. Unfortunately, the model of this study still found inconsistent accuracy at the adulteration limit of 0.5% for authentication and 2% for quantification.

Even more amazing, Karunathilaka *et al.* (2018) proposed a methodology to rapidly evaluate commercial milk powders to determine if they are original or may include known or unknown adulterants using SIMCA classification algorithm. They claim that the classification models produced 100% sensitivities using benchtop spectrometers to detect milk powder fraud and are not limited only to specific types of known adulterants. This shows that using near infrared spectroscopy with the appropriate processing method will provide very precise and fast evaluation results for fraudulent food and agro-products.

Another issue in the livestock product is meat adulteration. Unscrupulous traders adulterate meat products with another adulterant (cheaper meat, animal offal, spoiled meat, and non-meat chemical synthetic materials) for profiteering

purposes. Hence, Zhao *et al.* (2019) report the VIS-NIR technique to predict beef adulteration with spoiled beef using the LS-SVM algorithm. They declare that applying LS-SVM in the spectral range of 496 to 1000 nm can predict spoiled beef with an error prediction of approximately 5.67%. Weng et al. [52] conducted another research on the detection of adulteration meat using VIS-NIR spectroscopy was conducted by Weng *et al.* (2020) with minced beef samples. They used a spectral range of 350–2500 nm and claimed to detect minced beef mixed with pork and beef heart with error predictions of approximately 2.145% and 2.758%, respectively. These studies show that the application of VIS-NIR spectroscopy coupled with chemometrics can be powerful for the fast and accurate detection of adulterated livestock products.

2.6.2 Adulteration in Flour Products

The detection of fraud in flour products ingredients has become an even more important topic since flour products, such as bread and other bakery products, are widely consumed as primary foods. Many consumers lost trust in the food they were buying and the food industry identified that more rapid measures in terms of the evaluation of its product had to be put in place. Frequently adulteration is achieved in high-value food items and those that come through complex supply chains. The flour product that comes from food is likely more highly vulnerable to adulteration due to the complexity of the characteristics, and it is widely used for products such as bread. To address this, cutting-edge methods must be easy to use, fast and inexpensive, especially for the flour industry. The most interesting method today is the application of food fingerprinting as a detection method by IR technology. At least in the last five years, durum wheat flour, banana flour, einkorn flour, wheat flour, barley flour and cassava flour were among the flour products found to be the most commonly adulterated and the researchers have studied how to detect it using IR spectroscopy technology.

In old studies, Cocchi *et al.* (2006) ever studied the use of near infrared spectroscopy to quantify the adulteration level of durum wheat flour using the PLS algorithm. The authors claim near infrared spectroscopy data can show durum wheat flour adulteration using SNV pretreatment. In another study by Ndlovu *et al.* (2019) considered VIS-NIR spectroscopy to detect adulteration of unripe banana flour with

wheat flour. They found that the PCA model could successfully separate samples of pure and contaminated banana flour. PLSR model also could quantify the level of adulteration. Both results of this study indicate that NIR and VIS-NIR spectroscopy could monitor the quality of flour in retail markets for the purpose of product verification.

In a recent study by Ayvaz *et al.* (2021a), near infrared spectroscopy is suggested to detect adulteration of einkorn flour with wheat flour and presents a correlation coefficient of 0.94 to 0.99. The lowest correlation coefficient is found in the adulteration ratio of wheat flour less than 7% (w/w). IR spectroscopy was also used by Aykas and Menevseoglu (2021) to detect the mixing of powdered pistachio with powdered green pea and peanut. Infrared spectroscopy can be correctly predicted with a coefficient correlation of about 0.99.

Furthermore, Tao published a study on the detection of eight varieties of adulterants of cassava flour in wheat flour using micro-IR spectroscopy in the range of 1150–2150 nm. The classification of this study finding that the adulteration of wheat flour with cassava flour achieved 100% accuracy, yet the level adulteration of wheat flour with cassava flour (5% to 40% adulteration) only presented correct classification rates between 56.25% and 100%. The last but not least, study reported by Xu *et al.* (2013c) used near infrared spectroscopy in the 1000–2500 nm range to classify Chinese glutinous rice flour from extraneous adulterants and unwanted variations. This study found an adulteration specificity of 0.92 with one-class partial least squares algorithms.

2.6.3 Adulteration in Liquid Agro-product

Adulteration of liquid agro-products is valued in the same way as pure products, and there is a need for fast, easy, and precise analytical methods to assess their characteristics and originality. Popular liquid agro-products obtained in the form of naturally sweet and viscous products are honey, fruit juices, and vegetable oil.

According to Tan *et al.* (2021) and [18], the chemical content of wild honey is correlated with the season, geographical region, storage method and harvesting method, which makes it very difficult to compare other types of honey. It also makes honey very susceptible to adulteration and is valued similarly to pure honey. Evaluation the feasibility of near infrared spectroscopy technology in the rapid

detection and classification of adulteration of honey has been study by some researcher. Kelly *et al.* (2006) detect adulterated honey from beet invert syrup and high fructose corn syrup using near infrared spectroscopy (1100–2498 nm) with an accuracy between 9.0 and 11.9 (RMSE-CV). Furthermore, the same study was also conducted by Bázár *et al.* (2016) to detect corn syrup additives in honey using near infrared spectroscopy in the wavelength ranges 1300–1800 nm and reached an accuracy better than the previous study (RMSE-CV of 1.48). Besides, Ferreiro-González *et al.* (2018) used VIS-NIR spectroscopy (400–2500 nm) to predict honey adulteration with fructose-rich corn syrup and obtained an accuracy not yet better than Bázár *et al.* (2016) (RMSE-CV of 4.71). The most recent to conduct a similar study is Valinger *et al.* (2021a), which evaluated the feasibility of near infrared spectroscopy technology in the rapid detection of adulteration of honey with corn syrup. Unfortunately, the results indicate that the near infrared spectroscopy of adulterated honey can be modeled to detect fraud with an accuracy that is not yet better than the previous study. However, the interesting one in this study is that the adulteration of honey with water reported cannot be predicted with precision.

Fruit juice becomes a liquid food agro-product of the most common adulteration with artificial sweeteners, dilution with water, and fraud with low-quality or less-expensive fruit juice. Therefore, some researchers have developed a fast and low-cost method for inspecting fruit juice adulteration or dilution. In one study, Mabood *et al.* (2018) reported applications of near infrared spectroscopy (860–2500 nm) for classification of adulteration and nonadulteration in commercial fruit juices with precision between 0.067 to 0.169 (RMSE).

2.6.4 Adulteration in Herbs and Spices

Spices are highly valued agro-products because they are used in many in the world to flavor and preserve processed food. However, herbs and spices are extremely vulnerable to commercial gain motivated fraud including black pepper, garlic, saffron, and oregano.

Spices are high-value food components in weight units because they have desirable flavor characteristics and, therefore, are economically profitable targets for adulteration. To address this problem, Wilde and Galvin-King *et al.* (2021b) conducted a study on the feasibility of near infrared and infrared spectroscopy to

detect adulteration in black pepper and garlic of adulterants. The developed model is claimed to classify black pepper from its adulteration with a percentage of correct between 92% to 100%. Investigation of garlic adulteration detection using parameter validation in the form of fit measurement has an accuracy in the range of 98.5% to 99.4%.

Meanwhile, Amirvaresi *et al.* (2021) applied infrared spectroscopy to authentication saffron adulteration with accuracy classification between 81.3 to 100%. Unfortunately, detection limitations are only in the range of 1.0–3.1% (w/w) for each adulterant. Work has also been carried out by Galvin-King *et al.* (2020a), who have utilized infrared spectroscopy to identify the presence of adulterate powdered paprika with Varying seed or pod. Their model claims to predict component adulteration on powdered paprika with a coefficient of determination of about 0.94.

2.7 Future Perspectives

Current studies indicate the potential of near infrared and infrared spectroscopy approaches for detecting the adulteration of food and agro-products. Such a breakthrough would undoubtedly support the further implementation of near infrared and infrared spectroscopy-based quality evaluation. The availability of multiple data sources and the fusion of multi-origin data affords a perspective for future research. The fusion of UV-VIS, near infrared, and infrared spectroscopy is the process of combining some spectral information to improve data quality and produce a high quality representation model (Valinger *et al.*, 2021a). Future studies may use sample adulteration from a different origin, variety, storage temperature, or even shelf-life when developing a model. With the increasing number and high quality of accessible samples, the future perspective for detecting the adulteration of food and agro-products possibly focuses on near infrared and infrared spectroscopy tandem with machine learning. The main advantage of the machine learning approach is decreasing the dependence on human domain knowledge by end-to-end analysis and the improved precision and generalizability.

2.8 Conclusions

In this paper, the feasibility of applying a non-destructive for detecting and discriminating food adulteration and agro-products is based on near infrared and infrared spectroscopy and various types of data analysis have been represented. Besides the non-destructive, the primary advantages of the analytical method are fast and economical, directing to cost-effective quality assurance of detecting such a key worldwide food and agro-products adulteration. Actually, once the chemometric model has been correctly calibrated, the time elapsed from the scanning of IR spectroscopy on the samples and their subsequent classification would only need a few seconds. Therefore, this approach could represent a concrete and effective answer to the need, claimed by industrial and agro-product producers, as well as by the Food Control Authority, for affordable, fast, and efficient technologies to evaluate food quality and authenticity. Furthermore, the results of the variable selection establish the basis for developing portable and handheld infrared spectroscopy, customized for the detection and discrimination of adulteration food and agro-products directly “in situ” to ensure authenticity and counteract adulteration. Last but not least, the promising results performed by the numerous laboratory model validation indicate the potential transferability of a near infrared and infrared spectroscopy-based method to various production food and agro-product sites.

In the future, although optimistic results were acquired in an investigation for fraud detection for food and agro-products today, it must be pointed out that the optical for near-infrared and infrared spectroscopy technologies applied remain pricey so far. To implement routine analyses in some food and agro-products, it is necessary to develop low-cost infrared optical technologies and have the same accuracy as those currently available.

2.9 References

Alamar, P. D., Caramês, E. T. S., Poppi, R. J., & Pallone, J. A. L. (2020). Detection of Fruit Pulp Adulteration Using Multivariate Analysis: Comparison of NIR, MIR and Data Fusion Performance. *Food Analytical Methods, 13*(6), 1357-1365.

Alamprese, C., Amigo, J. M., Casiraghi, E., & Engelsen, S. B. (2016). Identification and quantification of turkey meat adulteration in fresh, frozen-thawed and cooked minced beef by FT-NIR spectroscopy and chemometrics. *Meat Science, 121*, 175-181.

Alamprese, C., Casale, M., Sinelli, N., Lanteri, S., & Casiraghi, E. (2013). Detection of minced beef adulteration with turkey meat by UV–vis, NIR and MIR spectroscopy. *LWT - Food Science and Technology, 53*(1), 225-232.

Aliaño-González, M. J., Ferreiro-González, M., Espada-Bellido, E., Palma, M., & Barbero, G. F. (2019). A screening method based on Visible-NIR spectroscopy for the identification and quantification of different adulterants in high-quality honey. *Talanta, 203*, 235-241.

Amirvaresi, A., Nikounezhad, N., Amirahmadi, M., Daraei, B., & Parastar, H. (2021). Comparison of near-infrared (NIR) and mid-infrared (MIR) spectroscopy based on chemometrics for saffron authentication and adulteration detection. *Food Chemistry, 344*.

Aykas, D. P., & Menevseoglu, A. (2021). A rapid method to detect green pea and peanut adulteration in pistachio by using portable FT-MIR and FT-NIR spectroscopy combined with chemometrics. *Food Control, 121*.

Ayvaz, H., Korkmaz, F., Polat, H., Ayvaz, Z., & Barış Tuncel, N. (2021a). Detection of einkorn flour adulteration in flour and bread samples using Computer-Based Image Analysis and Near-Infrared Spectroscopy. *Food Control, 127*, 108162.

Ayvaz, H., Korkmaz, F., Polat, H., Ayvaz, Z., & Barış Tuncel, N. (2021b). Detection of einkorn flour adulteration in flour and bread samples using Computer-Based Image Analysis and Near-Infrared Spectroscopy. *Food Control, 127*.

Bázár, G., Romvári, R., Szabó, A., Somogyi, T., Éles, V., & Tsenkova, R. (2016). NIR detection of honey adulteration reveals differences in water spectral pattern. *Food Chemistry, 194*, 873-880.

Biancolillo, A., Santoro, A., Firmani, P., & Marini, F. (2020). Identification and Quantification of Turmeric Adulteration in Egg-Pasta by Near Infrared Spectroscopy and Chemometrics. *Applied Sciences, 10*(8), 2647.

Botelho, B. G., Reis, N., Oliveira, L. S., & Sena, M. M. (2015). Development and analytical validation of a screening method for simultaneous detection of five

adulterants in raw milk using mid-infrared spectroscopy and PLS-DA. *Food Chemistry, 181*, 31-37.

Cantarelli, M. Á., Moldes, C. A., Marchevsky, E. J., Azcarate, S. M., & Camiña, J. M. (2020). Low-cost analytic method for the identification of Cinnamon adulteration. *Microchemical Journal, 159*.

Capuano, E., Boerrigter-Eenling, R., Koot, A., & van Ruth, S. M. (2015). Targeted and Untargeted Detection of Skim Milk Powder Adulteration by Near-Infrared Spectroscopy. *Food Analytical Methods, 8*(8), 2125-2134.

Castro, R. C., Ribeiro, D. S. M., Santos, J. L. M., & Páscoa, R. N. M. J. (2021). Near infrared spectroscopy coupled to MCR-ALS for the identification and quantification of saffron adulterants: Application to complex mixtures. *Food Control, 123*.

Chen, H., Tan, C., Lin, Z., & Wu, T. (2017). Detection of melamine adulteration in milk by near-infrared spectroscopy and one-class partial least squares. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 173*, 832-836.

Chen, L., Xue, X., Ye, Z., Zhou, J., Chen, F., & Zhao, J. (2011). Determination of Chinese honey adulterated with high fructose corn syrup by near infrared spectroscopy. *Food Chemistry, 128*(4), 1110-1114.

Cocchi, M., Durante, C., Foca, G., Marchetti, A., Tassi, L., & Ulrici, A. (2006). Durum wheat adulteration detection by NIR spectroscopy multivariate calibration. *Talanta, 68*(5), 1505-1511.

Contal, L., León, V., & Downey, G. (2002). Detection and quantification of apple adulteration in strawberry and raspberry purées using visible and near infrared spectroscopy. *Journal of Near Infrared Spectroscopy, 10*(4), 289-299.

Correia, R. M., Tosato, F., Domingos, E., Rodrigues, R. R. T., Aquino, L. F. M., Filgueiras, P. R., Lacerda, V., & Romão, W. (2018). Portable near infrared spectroscopy applied to quality control of Brazilian coffee. *Talanta, 176*, 59-68.

Da Silva Dias, L., Da Silva, J. C., De Souza Maudeira Felicio, A. L., & De Franca, J. A. (2018). A NIR Photometer Prototype with Integrating Sphere for the Detection of Added Water in Raw Milk. *IEEE Transactions on Instrumentation and Measurement, 67*(12), 2812-2819.

de Araújo, T. K. L., Nóbrega, R. O., Fernandes, D. D. D. S., de Araújo, M. C. U., Diniz, P. H. G. D., & da Silva, E. C. (2021). Non-destructive authentication of Gourmet ground roasted coffees using NIR spectroscopy and digital images. *Food Chemistry, 364*.

De Girolamo, A., Arroyo, M. C., Cervellieri, S., Cortese, M., Pascale, M., Logrieco, A. F., & Lippolis, V. (2020a). Detection of durum wheat pasta adulteration with common wheat by infrared spectroscopy and chemometrics: A case study. *LWT, 127*.

De Girolamo, A., Arroyo, M. C., Lippolis, V., Cervellieri, S., Cortese, M., Pascale, M., Logrieco, A. F., & von Holst, C. (2020b). A simple design for the validation of a FT-NIR screening method: Application to the detection of durum wheat pasta adulteration. *Food Chemistry, 333*.

Ding, H. B., & Xu, R. J. (2000). Near-infrared spectroscopic technique for detection of beef hamburger adulteration. *Journal of Agricultural and Food Chemistry, 48*(6), 2193-2198.

Ding, X., Ni, Y., & Kokot, S. (2015). NIR spectroscopy and chemometrics for the discrimination of pure, powdered, purple sweet potatoes and their samples adulterated with the white sweet potato flour. *Chemometrics and Intelligent Laboratory Systems, 144*, 17-23.

dos Santos Pereira, E. V., de Sousa Fernandes, D. D., de Araújo, M. C. U., Diniz, P. H. G. D., & Maciel, M. I. S. (2021a). In-situ authentication of goat milk in terms of its adulteration with cow milk using a low-cost portable NIR spectrophotometer. *Microchemical Journal, 163*.

dos Santos Pereira, E. V., de Sousa Fernandes, D. D., de Araújo, M. C. U., Diniz, P. H. G. D., & Maciel, M. I. S. (2021b). In-situ authentication of goat milk in terms of its adulteration with cow milk using a low-cost portable NIR spectrophotometer. *Microchemical Journal, 163*, 105885.

Downey, G., Fouratier, V., & Kelly, J. D. (2003). Detection of honey adulteration by addition of fructose and glucose using near infrared transflectance spectroscopy. *Journal of Near Infrared Spectroscopy, 11*(6), 447-456.

Downey, G., & Kelly, J. D. (2004). Detection and Quantification of Apple Adulteration in Diluted and Sulfited Strawberry and Raspberry Purées Using Visible and

Near-Infrared Spectroscopy. *Journal of Agricultural and Food Chemistry, 52*(2), 204-209.

Du, Q., Zhu, M., Shi, T., Luo, X., Gan, B., Tang, L., & Chen, Y. (2021a). Adulteration detection of corn oil, rapeseed oil and sunflower oil in camellia oil by in situ diffuse reflectance near-infrared spectroscopy and chemometrics. *Food Control, 121*, 107577.

Du, Q., Zhu, M., Shi, T., Luo, X., Gan, B., Tang, L., & Chen, Y. (2021b). Adulteration detection of corn oil, rapeseed oil and sunflower oil in camellia oil by in situ diffuse reflectance near-infrared spectroscopy and chemometrics. *Food Control, 121*.

Dvorak, L., Mlcek, J., & Sustova, K. (2016). Comparison of FT-NIR spectroscopy and ELISA for detection of adulteration of goat cheeses with cow's milk. *Journal of AOAC International, 99*(1), 180-186.

Ferreiro-González, M., Espada-Bellido, E., Guillén-Cueto, L., Palma, M., Barroso, C. G., & Barbero, G. F. (2018). Rapid quantification of honey adulteration by visible-near infrared spectroscopy combined with chemometrics. *Talanta, 188*, 288-292.

Galvin-King, P., Haughey, S. A., & Elliott, C. T. (2020a). The Detection of Substitution Adulteration of Paprika with Spent Paprika by the Application of Molecular Spectroscopy Tools. *Foods, 9*(7).

Galvin-King, P., Haughey, S. A., & Elliott, C. T. (2020b). The Detection of Substitution Adulteration of Paprika with Spent Paprika by the Application of Molecular Spectroscopy Tools. *Foods, 9*(7), 944.

Galvin-King, P., Haughey, S. A., & Elliott, C. T. (2021a). Garlic adulteration detection using NIR and FTIR spectroscopy and chemometrics. *Journal of Food Composition and Analysis, 96*.

Galvin-King, P., Haughey, S. A., & Elliott, C. T. (2021b). Garlic adulteration detection using NIR and FTIR spectroscopy and chemometrics. *Journal of Food Composition and Analysis, 96*, 103757.

Gayo, J., & Hale, S. A. (2007). Detection and quantification of species authenticity and adulteration in crabmeat using visible and near-infrared spectroscopy. *Journal of Agricultural and Food Chemistry, 55*(3), 585-592.

Gayo, J., Hale, S. A., & Blanchard, S. M. (2006). Quantitative analysis and detection of adulteration in crab meat using visible and near-infrared spectroscopy. *Journal of Agricultural and Food Chemistry, 54*(4), 1130-1136.

Genis, H. E., Durna, S., & Boyaci, I. H. (2021). Determination of green pea and spinach adulteration in pistachio nuts using NIR spectroscopy. *LWT, 136*, 110008.

Hosseini, E., Ghasemi, J. B., Daraei, B., Asadi, G., & Adib, N. (2021). Application of genetic algorithm and multivariate methods for the detection and measurement of milk-surfactant adulteration by attenuated total reflection and near-infrared spectroscopy. *Journal of the Science of Food and Agriculture, 101*(7), 2696-2703.

Huang, F., Song, H., Guo, L., Guang, P., Yang, X., Li, L., Zhao, H., & Yang, M. (2020a). Detection of adulteration in Chinese honey using NIR and ATR-FTIR spectral data fusion. *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy, 235*.

Huang, F., Song, H., Guo, L., Guang, P., Yang, X., Li, L., Zhao, H., & Yang, M. (2020b). Detection of adulteration in Chinese honey using NIR and ATR-FTIR spectral data fusion. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 235*, 118297.

Jahani, R., Yazdanpanah, H., van Ruth, S. M., Kobarfard, F., Alewijn, M., Mahboubi, A., Faizi, M., Aliabadi, M. H. S., & Salamzadeh, J. (2020). Novel application of near-infrared spectroscopy and chemometrics approach for detection of lime juice adulteration. *Iranian Journal of Pharmaceutical Research, 19*(2), 34-44.

Jha, S. N., & Matsuoka, T. (2004). Detection of adulterants in milk using near infrared spectroscopy. *Journal of Food Science and Technology, 41*(3), 313-316.

Kar, S., Tudu, B., Bag, A. K., & Bandyopadhyay, R. (2018). Application of Near-Infrared Spectroscopy for the Detection of Metanil Yellow in Turmeric Powder. *Food Analytical Methods, 11*(5), 1291-1302.

Kar, S., Tudu, B., Jana, A., & Bandyopadhyay, R. (2019). FT-NIR spectroscopy coupled with multivariate analysis for detection of starch adulteration in turmeric powder. *Food Additives and Contaminants - Part A Chemistry, Analysis, Control, Exposure and Risk Assessment, 36*(6), 863-875.

Karunathilaka, S. R., Yakes, B. J., He, K., Chung, J. K., & Mossoba, M. (2018). Non-targeted NIR spectroscopy and SIMCA classification for commercial milk

powder authentication: A study using eleven potential adulterants. *Heliyon, 4*(9), 1-23.

Kasemsumran, S., Thanapase, W., & Kiatsoonthon, A. (2007). Feasibility of near-infrared spectroscopy to detect and to quantify adulterants in cow milk. *Analytical Sciences, 23*(7), 907-910.

Kaufmann, K. C., Sampaio, K. A., García-Martín, J. F., & Barbin, D. F. (2022). Identification of coriander oil adulteration using a portable NIR spectrometer. *Food Control, 132*, 108536.

Kazazić, S., Gajdoš-Kljusurić, J., Radeljević, B., Plavljanić, D., Špoljarić, J., Ljubić, T., Bilić, B., & Mikulec, N. (2021). Comparison of GC and NIR spectra as a rapid tool for food fraud detection: Case of butter adulteration with different fat types. *Journal of Food Processing and Preservation, 45*(9).

Kelly, J. D., Petisco, C., & Downey, G. (2006). Potential of near Infrared Transflectance Spectroscopy to Detect Adulteration of Irish Honey by Beet Invert Syrup and High Fructose Corn Syrup. *Journal of Near Infrared Spectroscopy, 14*(2), 139-146.

Kene Ejeahalaka, K., & On, S. L. W. (2020). Effective detection and quantification of chemical adulterants in model fat-filled milk powders using NIRS and hierarchical modelling strategies. *Food Chemistry, 309*, 125785.

Kumaravelu, C., & Gopal, A. (2015). Detection and Quantification of Adulteration in Honey through Near Infrared Spectroscopy. *International Journal of Food Properties, 18*(9), 1930-1935.

Kuswandi, B., Cendekiawan, K. A., Kristiningrum, N., & Ahmad, M. (2015). Pork adulteration in commercial meatballs determined by chemometric analysis of NIR Spectra. *Journal of Food Measurement and Characterization, 9*(3), 313-323.

Le Nguyen Doan, D., Nguyen, Q. C., Marini, F., & Biancolillo, A. (2021). Authentication of rice (Oryza sativa l.) using near infrared spectroscopy combined with different chemometric classification strategies. *Applied Sciences (Switzerland), 11*(1), 1-11.

Leng, T., Li, F., Xiong, L., Xiong, Q., Zhu, M., & Chen, Y. (2020). Quantitative detection of binary and ternary adulteration of minced beef meat with pork and duck meat by NIR combined with chemometrics. *Food Control, 113*.

León, L., Daniel Kelly, J., & Downey, G. (2005). Detection of apple juice adulteration using near-infrared transflectance spectroscopy. *Applied Spectroscopy, 59*(5), 593-599.

Lima, A. B. S. d., Batista, A. S., Jesus, J. C. d., Silva, J. d. J., Araújo, A. C. M. d., & Santos, L. S. (2020). Fast quantitative detection of black pepper and cumin adulterations by near-infrared spectroscopy and multivariate modeling. *Food Control, 107*, 106802.

Liu, X., Jia, G., Wu, C., Wang, K., & Wu, X. (2010). Determination of characteristic wave bands and detection of melamine in fishmeal by Fourier transform near infrared spectroscopy. *Journal of Near Infrared Spectroscopy, 18*(2), 113-120.

Liu, Y., & Zhou, S. (2017). Rapid detection of hydrolyzed leather protein adulteration in infant formula by near-infrared spectroscopy. *Food Science and Technology Research, 23*(3), 469-474.

Liu, Y., Zhou, S., Han, W., Li, C., Huang, K., & Liu, W. (2017). Detection of adulteration by hydrolysed leather protein in infant formula based on least squares support vector machine and near-infrared spectroscopy. *Journal of Food and Nutrition Research, 56*(3), 283-291.

Liu, Y., Zhou, S., Han, W., Li, C., Liu, W., Qiu, Z., & Chen, H. (2021). Detection of adulteration in infant formula based on ensemble convolutional neural network and near-infrared spectroscopy. *Foods, 10*(4).

Lohumi, S., Lee, S., Lee, W. H., Kim, M. S., Mo, C., Bae, H., & Cho, B. K. (2014). Detection of starch adulteration in onion powder by FT-NIR and FT-IR spectroscopy. *Journal of Agricultural and Food Chemistry, 62*(38), 9246-9251.

López, M. I., Trullols, E., Callao, M. P., & Ruisánchez, I. (2014). Multivariate screening in food adulteration: Untargeted versus targeted modelling. *Food Chemistry, 147*, 177-181.

Lukacs, M., Bazar, G., Pollner, B., Henn, R., Kirchler, C. G., Huck, C. W., & Kovacs, Z. (2018). Near infrared spectroscopy as an alternative quick method for simultaneous detection of multiple adulterants in whey protein-based sports supplement. *Food Control, 94*, 331-340.

Luqing, L., Lingdong, W., Jingming, N., & Zhengzhu, Z. (2015). Detection and Quantification of Sugar and Glucose Syrup in Roasted Green Tea Using near Infrared Spectroscopy. *Journal of Near Infrared Spectroscopy, 23*(5), 317-325.

Mabood, F., Ali, L., Boque, R., Abbas, G., Jabeen, F., Haq, Q. M. I., Hussain, J., Hamaed, A. M., Naureen, Z., Al-Nabhani, M., Khan, M. Z., Khan, A., & Al-Harrasi, A. (2020). Robust Fourier transformed infrared spectroscopy coupled with multivariate methods for detection and quantification of urea adulteration in fresh milk samples. *Food Science and Nutrition, 8*(10), 5249-5258.

Mabood, F., Hussain, J., Jabeen, F., Abbas, G., Allaham, B., Albroumi, M., Alghawi, S., Alameri, S., Gilani, S. A., Al-Harrasi, A., Haq, Q. M. I., & Farooq, S. (2018). Applications of FT-NIRS combined with PLS multivariate methods for the detection & quantification of saccharin adulteration in commercial fruit juices. *Food Additives & Contaminants: Part A, 35*(6), 1052-1060.

Mabood, F., Jabeen, F., Ahmed, M., Hussain, J., Al Mashaykhi, S. A. A., Al Rubaiey, Z. M. A., Farooq, S., Boqué, R., Ali, L., Hussain, Z., Al-Harrasi, A., Khan, A. L., Naureen, Z., Idrees, M., & Manzoor, S. (2017a). Development of new NIR-spectroscopy method combined with multivariate analysis for detection of adulteration in camel milk with goat milk. *Food Chemistry, 221*, 746-750.

Mabood, F., Jabeen, F., Hussain, J., Al-Harrasi, A., Hamaed, A., Al Mashaykhi, S. A. A., Al Rubaiey, Z. M. A., Manzoor, S., Khan, A., Haq, Q. M. I., Gilani, S. A., & Khan, A. (2017b). FT-NIRS coupled with chemometric methods as a rapid alternative tool for the detection & quantification of cow milk adulteration in camel milk samples. *Vibrational Spectroscopy, 92*, 245-250.

Maraboli, A., Cattaneo, T. M. P., & Giangiacomo, R. (2002). Detection of vegetable proteins from soy, pea and wheat isolates in milk powder by near infrared spectroscopy. *Journal of Near Infrared Spectroscopy, 10*(1), 63-69.

Masithoh, R. E., Roosmayanti, F., Rismiwandira, K., & Pahlawan, M. F. R. (2021). Detection of Palm Sugar Adulteration by Fourier Transform Near-Infrared (FT-NIR) and Fourier Transform Infrared (FT-IR) Spectroscopy. *Sugar Tech*.

Mishra, S., Kamboj, U., Kaur, H., & Kapur, P. (2010). Detection of jaggery syrup in honey using near-infrared spectroscopy. *International Journal of Food Sciences and Nutrition, 61*(3), 306-315.

Morsy, N., & Sun, D.-W. (2013). Robust linear and non-linear models of NIR spectroscopy for detection and quantification of adulterants in fresh and frozen-thawed minced beef. *Meat Science, 93*(2), 292-302.

Mouazen, A. M., & Al-Walaan, N. (2014). Glucose adulteration in saudi honey with visible and near infrared spectroscopy. *International Journal of Food Properties, 17*(10), 2263-2274.

Murray, I., Aucott, L. S., & Pike, I. H. (2001). Use of discriminant analysis on visible and near infrared reflectance spectra to detect adulteration of fishmeal with meat and bone meal. *Journal of Near Infrared Spectroscopy, 9*(4), 297-311.

Ndlovu, P. F., Magwaza, L. S., Tesfay, S. Z., & Mphahlele, R. R. (2019). Rapid visible–near infrared (Vis–NIR) spectroscopic detection and quantification of unripe banana flour adulteration with wheat flour. *Journal of Food Science and Technology, 56*(12), 5484-5491.

Ndlovu, P. F., Magwaza, L. S., Tesfay, S. Z., & Mphahlele, R. R. (2021a). Rapid spectroscopic method for quantifying gluten concentration as a potential biomarker to test adulteration of green banana flour. *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy, 262*.

Ndlovu, P. F., Magwaza, L. S., Tesfay, S. Z., & Mphahlele, R. R. (2021b). Vis-NIR spectroscopic and chemometric models for detecting contamination of premium green banana flour with wheat by quantifying resistant starch content. *Journal of Food Composition and Analysis, 102*.

Oliveira, M. M., Cruz-Tirado, J. P., Roque, J. V., Teófilo, R. F., & Barbin, D. F. (2020). Portable near-infrared spectroscopy for rapid authentication of adulterated paprika powder. *Journal of Food Composition and Analysis, 87*, 103403.

Ozaki, Y., Huck, C., Tsuchikawa, S., & Engelsen, S. B. (2021). *Near-Infrared Spectroscopy: Theory, Spectral Analysis, Instrumentation, and Applications*: Springer.

Özdemir, D., & Öztürk, B. (2007). Near infrared spectroscopic determination of olive oil adulteration with sunflower and corn oil. *Journal of Food and Drug Analysis, 15*(1), 40-47.

Öztürk, B., Yalçin, A., & Özdemir, D. (2010). Determination of olive oil adulteration with vegetable oils by near infrared spectroscopy coupled with multivariate calibration. *Journal of Near Infrared Spectroscopy, 18*(3), 191-201.

Pandiselvam, R., Mahanti, N. K., Manikantan, M. R., Kothakota, A., Chakraborty, S. K., Ramesh, S. V., & Beegum, P. P. S. (2022). Rapid detection of adulteration in

desiccated coconut powder: vis-NIR spectroscopy and chemometric approach. *Food Control, 133*(Part A), 108588.

Paradkar, M. M., Sakhamuri, S., & Irudayaraj, J. (2002a). Comparison of FTIR, FT-Raman, and NIR spectroscopy in a maple syrup adulteration study. *Journal of Food Science, 67*(6), 2009-2015.

Paradkar, M. M., Sivakesava, S., & Irudayaraj, J. (2002b). Discrimination and classification of adulterants in maple syrup with the use of infrared spectroscopic techniques. *Journal of the Science of Food and Agriculture, 82*(5), 497-504.

Pereira, C. G., Leite, A. I. N., Andrade, J., Bell, M. J. V., & Anjos, V. (2019). Evaluation of butter oil adulteration with soybean oil by FT-MIR and FT-NIR spectroscopies and multivariate analyses. *LWT, 107*, 1-8.

Pereira, E. V. D. S., Fernandes, D. D. D. S., de Araújo, M. C. U., Diniz, P. H. G. D., & Maciel, M. I. S. (2020). Simultaneous determination of goat milk adulteration with cow milk and their fat and protein contents using NIR spectroscopy and PLS algorithms. *LWT, 127*.

Picouet, P. A., Gou, P., Hyypiö, R., & Castellari, M. (2018). Implementation of NIR technology for at-line rapid detection of sunflower oil adulterated with mineral oil. *Journal of Food Engineering, 230*, 18-27.

Pizarro, C., Esteban-Díez, I., & González-Sáiz, J. M. (2007). Mixture resolution according to the percentage of robusta variety in order to detect adulteration in roasted coffee by near infrared spectroscopy. *Analytica Chimica Acta, 585*(2), 266-276.

Quelal-Vásconez, M. A., Pérez-Esteve, É., Arnau-Bonachera, A., Barat, J. M., & Talens, P. (2018). Rapid fraud detection of cocoa powder with carob flour using near infrared spectroscopy. *Food Control, 92*, 183-189.

Rady, A., & Adedeji, A. (2018). Assessing different processed meats for adulterants using visible-near-infrared spectroscopy. *Meat Science, 136*, 59-67.

Ramírez-Morales, I., Rivero, D., Fernández-Blanco, E., & Pazos, A. (2016). Optimization of NIR calibration models for multiple processes in the sugar industry. *Chemometrics and Intelligent Laboratory Systems, 159*, 45-57.

Reich, G. (2016). Mid and near infrared spectroscopy *Analytical Techniques in the Pharmaceutical Sciences* (pp. 61-138): Springer.

Rodriguez-Saona, L. E., Fry, F. S., McLaughlin, M. A., & Calvey, E. M. (2001). Rapid analysis of sugars in fruit juices by FT-NIR spectroscopy. *Carbohydrate Research, 336*(1), 63-74.

Rukundo, I. R., Danao, M.-G. C., Weller, C. L., Wehling, R. L., & Eskridge, K. M. (2020). Use of a handheld near infrared spectrometer and partial least squares regression to quantify metanil yellow adulteration in turmeric powder. *Journal of Near Infrared Spectroscopy, 28*(2), 81-92.

Rukundo, I. R., & Danao, M. C. (2020). Identifying turmeric powder by source and metanil yellow adulteration levels using near-infrared spectra and PCA-SIMCA modeling. *Journal of Food Protection, 83*(6), 968-974.

Santos, I. A., Conceição, D. G., Viana, M. B., Silva, G. D. J., Santos, L. S., & Ferrão, S. P. B. (2021). NIR and MIR spectroscopy for quick detection of the adulteration of cocoa content in chocolates. *Food Chemistry, 349*.

Santos, P. M., Pereira-Filho, E. R., & Rodriguez-Saona, L. E. (2013). Application of hand-held and portable infrared spectrometers in bovine milk analysis. *Journal of Agricultural and Food Chemistry, 61*(6), 1205-1211.

Schmutzler, M., Beganovic, A., Böhler, G., & Huck, C. W. (2015). Methods for detection of pork adulteration in veal product based on FT-NIR spectroscopy for laboratory, industrial and on-site analysis. *Food Control, 57*, 258-267.

Shannon, M., Ratnasekhar, C. H., McGrath, T. F., Kapil, A. P., & Elliott, C. T. (2021). A two-tiered system of analysis to tackle rice fraud: The Indian Basmati study. *Talanta, 225*.

Shen, G., Fan, X., Yang, Z., & Han, L. (2016). A feasibility study of non-targeted adulterant screening based on NIRM spectral library of soybean meal to guarantee quality: The example of non-protein nitrogen. *Food Chemistry, 210*, 35-42.

Silva, L. C. R., Folli, G. S., Santos, L. P., Barros, I. H. A. S., Oliveira, B. G., Borghi, F. T., Santos, F. D. D., Filgueiras, P. R., & Romão, W. (2020). Quantification of beef, pork, and chicken in ground meat using a portable NIR spectrometer. *Vibrational Spectroscopy, 111*.

Srinuttrakul, W., Mihailova, A., Islam, M. D., Liebisch, B., Maxwell, F., Kelly, S. D., & Cannavan, A. (2021). Geographical differentiation of hom mali rice cultivated in different regions of thailand using ftir-atr and nir spectroscopy. *Foods, 10*(8).

Tan, S. H., Pui, L. P., Solihin, M. I., Keat, K. S., Lim, W. H., & Ang, C. K. (2021). Physicochemical analysis and adulteration detection in Malaysia stingless bee honey using a handheld near-infrared spectrometer. *Journal of Food Processing and Preservation, 45*(7).

Tao, F., Liu, L., Kucha, C., & Ngadi, M. (2021). Rapid and non-destructive detection of cassava flour adulterants in wheat flour using a handheld MicroNIR spectrometer. *Biosystems Engineering, 203*, 34-43.

Teixeira, J. L. D. P., Caramês, E. T. D. S., Baptista, D. P., Gigante, M. L., & Pallone, J. A. L. (2020). Vibrational spectroscopy and chemometrics tools for authenticity and improvement the safety control in goat milk. *Food Control, 112*.

Teixeira, J. L. D. P., Caramês, E. T. D. S., Baptista, D. P., Gigante, M. L., & Pallone, J. A. L. (2021a). Rapid adulteration detection of yogurt and cheese made from goat milk by vibrational spectroscopy and chemometric tools. *Journal of Food Composition and Analysis, 96*.

Teixeira, J. L. d. P., Caramês, E. T. d. S., Baptista, D. P., Gigante, M. L., & Pallone, J. A. L. (2021b). Rapid adulteration detection of yogurt and cheese made from goat milk by vibrational spectroscopy and chemometric tools. *Journal of Food Composition and Analysis, 96*, 103712.

Temizkan, R., Can, A., Dogan, M. A., Mortas, M., & Ayvaz, H. (2020a). Rapid detection of milk fat adulteration in yoghurts using near and mid-infrared spectroscopy. *International Dairy Journal, 110*, 104795.

Temizkan, R., Can, A., Dogan, M. A., Mortas, M., & Ayvaz, H. (2020b). Rapid detection of milk fat adulteration in yoghurts using near and mid-infrared spectroscopy. *International Dairy Journal, 110*.

Teye, E., Huang, X.-y., Lei, W., & Dai, H. (2014). Feasibility study on the use of Fourier transform near-infrared spectroscopy together with chemometrics to discriminate and quantify adulteration in cocoa beans. *Food Research International, 55*, 288-293.

Thyholt, K., Indahl, U. G., Hildrum, K. I., Ellekjær, M. R., & Isaksson, T. (1997). Meat speciation by near infrared reflectance spectroscopy on dry extract. *Journal of Near Infrared Spectroscopy, 5*(4), 195-208.

Torres, I., Sánchez, M. T., Vega-Castellote, M., & Pérez-Marín, D. (2021). Fraud detection in batches of sweet almonds by portable near-infrared spectral devices. *Foods, 10*(6).

Uddin, M., & Okazaki, E. (2004). Classification of fresh and frozen-thawed fish by near-infrared spectroscopy. *Journal of Food Science, 69*(8), C665-C668.

Uysal, R. S., & Boyaci, I. H. (2020). Authentication of liquid egg composition using ATR-FTIR and NIR spectroscopy in combination with PCA. *Journal of the Science of Food and Agriculture, 100*(2), 855-862.

Valinger, D., Longin, L., Grbeš, F., Benković, M., Jurina, T., Gajdoš Kljusurić, J., & Jurinjak Tušek, A. (2021a). Detection of honey adulteration – The potential of UV-VIS and NIR spectroscopy coupled with multivariate analysis. *LWT, 145*, 111316.

Valinger, D., Longin, L., Grbeš, F., Benković, M., Jurina, T., Gajdoš Kljusurić, J., & Jurinjak Tušek, A. (2021b). Detection of honey adulteration – The potential of UV-VIS and NIR spectroscopy coupled with multivariate analysis. *LWT, 145*.

Vichasilp, C., & Poungchompu, O. (2014). Feasibility of detecting pork adulteration in halal meatballs using near infrared spectroscopy (NIR). *Chiang Mai University Journal of Natural Sciences, 13*(1), 497-507.

Visconti, L. G., Rodríguez, M. S., & Di Anibal, C. V. (2020). Determination of grated hard cheeses adulteration by near infrared spectroscopy (NIR) and multivariate analysis. *International Dairy Journal, 104*, 104647.

Vitalis, F., Zaukuu, J. L. Z., Bodor, Z., Aouadi, B., Hitka, G., Kaszab, T., Zsom-Muha, V., Gillay, Z., & Kovacs, Z. (2020). Detection and quantification of tomato paste adulteration using conventional and rapid analytical methods. *Sensors (Switzerland), 20*(21), 1-21.

Wang, N., Zhang, X., Yu, Z., Li, G., & Zhou, B. (2014). Quantitative analysis of adulterations in oat flour by FT-NIR spectroscopy, incomplete unbalanced randomized block design, and partial least squares. *Journal of Analytical Methods in Chemistry, 2014*.

Weng, S., Guo, B., Tang, P., Yin, X., Pan, F., Zhao, J., Huang, L., & Zhang, D. (2020). Rapid detection of adulteration of minced beef using Vis/NIR reflectance spectroscopy with multivariate methods. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 230*, 1-9.

Wesley, I. J., Barnes, R. J., & McGill, A. E. J. (1995). Measurement of adulteration of olive oils by near-infrared spectroscopy. *Journal of the American Oil Chemists' Society, 72*(3), 289-292.

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.

Winkler-Moser, J. K., Singh, M., Rennick, K. A., Bakota, E. L., Jham, G., Liu, S. X., & Vaughn, S. F. (2015). Detection of Corn Adulteration in Brazilian Coffee (Coffea arabica) by Tocopherol Profiling and Near-Infrared (NIR) Spectroscopy. *Journal of Agricultural and Food Chemistry, 63*(49), 10662-10668.

Wongsaipun, S., Theanjumpol, P., & Kittiwachana, S. (2021). Development of a Universal Calibration Model for Quantification of Adulteration in Thai Jasmine Rice Using Near-infrared Spectroscopy. *Food Analytical Methods, 14*(5), 997-1010.

Xie, L. J., Ye, X. Q., Liu, D. H., & Ying, Y. B. (2008). Application of principal component-radial basis function neural networks (PC-RBFNN) for the detection of water-adulterated bayberry juice by near-infrared spectroscopy. *Journal of Zhejiang University: Science B, 9*(12), 982-989.

Xu, L., Fu, X. S., Fu, H. Y., & She, Y. B. (2015). Rapid Detection of Exogenous Adulterants and Species Discrimination for a Chinese Functional Tea (Banlangen) by Fourier-Transform Near-Infrared (FT-NIR) Spectroscopy and Chemometrics. *Journal of Food Quality, 38*(6), 450-457.

Xu, L., Yan, S.-M., Cai, C.-B., & Yu, X.-P. (2013a). Untargeted Detection of Illegal Adulterations in Chinese Glutinous Rice Flour (GRF) by NIR Spectroscopy and Chemometrics: Specificity of Detection Improved by Reducing Unnecessary Variations. *Food Analytical Methods, 6*(6), 1568-1575.

Xu, L., Yan, S. M., Cai, C. B., Wang, Z. J., & Yu, X. P. (2013b). The feasibility of using near-infrared spectroscopy and chemometrics for untargeted detection of protein adulteration in yogurt: Removing unwanted variations in pure yogurt. *Journal of Analytical Methods in Chemistry, 2013*.

Xu, L., Yan, S. M., Cai, C. B., & Yu, X. P. (2013c). Untargeted Detection of Illegal Adulterations in Chinese Glutinous Rice Flour (GRF) by NIR Spectroscopy and

Chemometrics: Specificity of Detection Improved by Reducing Unnecessary Variations. *Food Analytical Methods, 6*(6), 1568-1575.

Yang, X., Guang, P., Xu, G., Zhu, S., Chen, Z., & Huang, F. (2020). Manuka honey adulteration detection based on near-infrared spectroscopy combined with aquaphotomics. *LWT, 132*.

Yasmin, J., Ahmed, M. R., Lohumi, S., Wakholi, C., Lee, H., Mo, C., & Cho, B. K. (2019). Rapid authentication measurement of cinnamon powder using FT-NIR and FT-IR spectroscopic techniques. *Quality Assurance and Safety of Crops and Foods, 11*(3), 257-267.

Zaukuu, J. L. Z., Bodor, Z., Vitalis, F., Zsom-Muha, V., & Kovacs, Z. (2019). Near infrared spectroscopy as a rapid method for detecting paprika powder adulteration with corn flour. *Acta Periodica Technologica, 50*, 346-352.

Zhang, L.-G., Zhang, X., Ni, L.-J., Xue, Z.-B., Gu, X., & Huang, S.-X. (2014). Rapid identification of adulterated cow milk by non-linear pattern recognition methods based on near infrared spectroscopy. *Food Chemistry, 145*, 342-348.

Zhao, H.-T., Feng, Y.-Z., Chen, W., & Jia, G.-F. (2019). Application of invasive weed optimization and least square support vector machine for prediction of beef adulteration with spoiled beef based on visible near-infrared (Vis-NIR) hyperspectral imaging. *Meat Science, 151*, 75-81.

Zhao, M., O'Donnell, C. P., & Downey, G. (2013). Detection of offal adulteration in beefburgers using near infrared reflectance spectroscopy and multivariate modelling. *Journal of Near Infrared Spectroscopy, 21*(4), 237-248.

Zhu, X., Li, S., Shan, Y., Zhang, Z., Li, G., Su, D., & Liu, F. (2010). Detection of adulterants such as sweeteners materials in honey using near-infrared spectroscopy and chemometrics. *Journal of Food Engineering, 101*(1), 92-97.

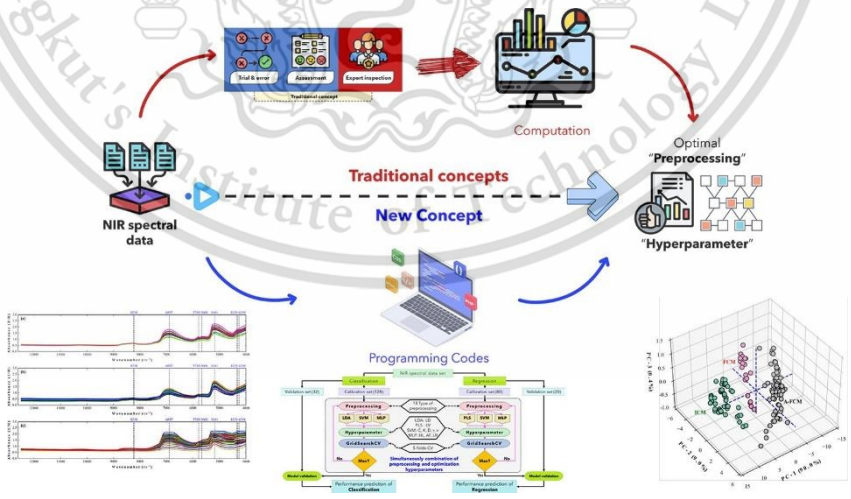
Ziegler, J. U., Leitenberger, M., Longin, C. F. H., Würschum, T., Carle, R., & Schweiggert, R. M. (2016). Near-infrared reflectance spectroscopy for the rapid discrimination of kernels and flours of different wheat species. *Journal of Food Composition and Analysis, 51*, 30-36.

CHAPTER 3 – CASE STUDY 1
A RAPID METHOD TO PREDICT TYPE AND ADULTERATION OF COCONUT MILK BY NEAR-INFRARED SPECTROSCOPY COMBINED WITH MACHINE LEARNING AND CHEMOMETRIC TOOLS[2]

3.1 Highlights

1. Rapid non-destructive methods to identify liquid coconut milk products were present in this study.
2. NIRs data followed by machine learning and chemometrics tools were employed to estimate the type and adulteration of coconut milk products.
3. A novel concept for generating model calibration that integrates preprocessing and hyperparameter optimization simultaneously was proposed for classification and regression issues in NIRs data.
4. The use of machine learning algorithms and chemometric tools for qualification and quantification purposes demonstrated satisfactory predictive models.

3.2 Graphical Abstract



[2]This chapter constituted the publication article: Sitorus, A., & Lapcharoensuk, R. (2023). A rapid method to predict type and adulteration of coconut milk by near-infrared spectroscopy combined with machine learning and chemometric tools. **Microchemical Journal**, 195, 109461. https://doi.org/10.1016/j.microc.2023.109461.

3.3 Abstract

Coconut milk is a soft target for adulterators owing to its simplicity of chemical composition. Professionals and consumers want to control the originalitas of coconut milk, while sellers can profit by mixing fresh coconut milk from low-cost products into high-value fresh coconut milk. Non-destructively and rapidly identifying coconut milk classification goods may be useful in quality assurance settings. However, no studies to date have investigated this topic. In this study, near-infrared spectra (NIRs) were collected from fresh coconut milk (FCM), instant coconut milk (ICM), and adulterated fresh coconut milk (A-FCM) in order to investigate the prospect of non-invasively discriminating coconut milk type and at the same time predicting the level of A-FCM. Partial least squares (PLS), linear discriminant analysis (LDA), support vector machine (SVM), and multilayer perceptron (MLP) were employed to establish classification and regression models using NIRs. Combining 18 preprocessing types and hyperparameter optimization of individual machine learning algorithms is carried out together and evaluated using 5-folds cross-validation. All algorithms in this study (LDA, SVM, MLP) obtained the same satisfactory results with all the precision, recall, F1-score, and perfect accuracy (100%) to distinguish FCM, ICM, and A-FCM in both calibration and prediction. Regression models using the SVM obtained acceptable results, with a determination coefficient of calibration and prediction all over 0.93, root mean square error of calibration and prediction all below 8.30%, and ratio of prediction to deviation over 3.80. Last but not least, this study would help apply NIRs to detect the originality of coconut milk in real-world conditions.

Keywords: adulteration; agro-product; chemometrics; coconut milk; FT-NIR.

3.4 Introduction

Coconut (*Cocos nucifera* Linn.) is a popular fruit in Asia due to the special features associated with its white meat, sweet water, and rich nutrition. Extraction from white meat will produce oil-in-water emulsion (coconut milk) is one of the important ingredients for traditional Asian dishes and is available worldwide in various forms. Thus, coconut milk can be easily found in traditional markets in Asia as fresh coconut milk and obtained in modern markets as instant coconut milk that has been packaged with various brands. Major nutrition component in coconut milk are

fat (15.44–38.0%), protein (2.06–3.50%), moisture (52.0–74.6%), ash (0.64–0.90%), and carbohydrate (2.7–6.88%) (Alyaqoubi *et al.*, 2015). In addition, one of the standards referred for coconut milk products is issued by Food and Agriculture Organization (FAO) through "CODEX STAN 240-2003: Codex standard for aqueous coconut milk and coconut cream products". This standard has categorized aqueous coconut products into four categories: light coconut milk, coconut milk, coconut cream, and coconut cream concentrate, depending on total solids, non-fat solids, total fat content and moisture content (CODEX-STAN-240, 2003).

Today, consumers are increasingly aware of food fraud for daily consumption. This is because food and agricultural products are more vulnerable to fraud than other products, particularly regarding their composition, origin, and processing technique. There is a gap for unscrupulous sellers to capitalize on the opportunity to tamper with the natural constituents of food and agricultural products to increase their apparent value for financial gains. Therefore, the determination of food quality, authenticity, and traceability of food and agricultural products should be monitored closely in the processing industry and its supply chain. Sampling and testing should be routinely conducted during food and agricultural production and trading to determine the major components to minimize adulteration (Faith Ndlovu *et al.*, 2022; Kucharska-Ambrozej and Karpinska, 2020; Sørensen *et al.*, 2016). In general, adulteration of coconut milk is created by adding water and mature coconut water to get more volumetric and cornflour to increase coconut milk's carbohydrate content (Azlin-hashim *et al.*, 2019; Simuang *et al.*, 2004; Sitorus *et al.*, 2021). Consequently, the adulteration of coconut milk will cause losses in the product's characteristics and limit the amount of content in coconut milk.

Traditional quality control methods of fresh coconut milk can be carried out in food testing laboratories. A laboratory based on chemical analysis testing requires many samples for each test parameter, sample preparation is not easy, lots of chemicals, labor-extensive, required skilled chemical operator, waste of energy, sophisticated instruments, and produces a lot of test residues as well as expensively (Lakshanasomya *et al.*, 2011). Compared to laboratories based on NIR spectroscopy, instruments are simple to install and operate with little or no sample preparation and can evaluate many constituents at a time (Panmanas and Chin Hock, 2019).

Besides that, the results of the quality of the fresh coconut milk tested can only be known in the next few days and yet provides a real-time evaluation. This causes the decision-making of fresh coconut milk quality cannot be made quickly because their matrix is similar, making it challenging to detect adulteration using traditional methods. Moreover, using chemicals and intensive energy in traditional methods of the testing process will also cause environmental problems if carried out continuously. In brief, as reported by Lakshanasomya *et al.* (2011), the general method for determining total solids in coconut milk is by drying it in a hot air oven or using a vacuum oven, which takes more than 2 hr from preparing until getting one result. Next, determining total fat in coconut milk using solvent extraction techniques also involves at least several chemicals such as ether, petroleum ether, hydrochloric acid, and ammonium hydroxide.

As an alternative to traditional methods, near-infrared (NIR) spectroscopy technology is a non-destructive and rapid method that can directly predict specific agricultural products' property and their chemical constituent. NIR spectroscopy is powerful because its measurement does not require complicated sample preparation and provides faster predictive results. Previously, NIR spectroscopy have been applied to evaluate the adulteration in many kinds of food and agricultural products with acceptable outcomes. Some researchers report that NIR spectroscopy can predict qualitatively and quantitatively the adulteration of liquid food and agro-products such as honey (Bázár *et al.*, 2016), milk (Mabood *et al.*, 2017), fruit juice (Mabood *et al.*, 2018), chili sauce (Lapcharoensuk *et al.*, 2020), and coriander oil (Kaufmann *et al.*, 2022). These studies show a coefficient of determination between 0.72 to 0.99. Another study used NIR spectroscopy to predict both qualitatively and quantitatively the adulteration of solid food and agro products, including black pepper with the PLS-DA algorithm (Wilde *et al.*, 2019), ginger powder by random forest and gradient boosting (Yu *et al.*, 2022), chickpea flour by PLSR (Bala *et al.*, 2022) and nutmeg with PC-MLP algorithm (Sitorus *et al.*, 2023). These studies show a coefficient of determination and or accuracy between 0.93 to 1.0. Although these technology are sensitive and accurate, spectroscopy data acquired using NIR spectroscopy frequently found in overlapping circumstances and revealing little about the overall structure (Cardoso and Poppi, 2021; Ribeiro *et al.*, 2021). This

makes disclosing the information in it challenging for a chemometrician and makes them less appealing as a technique to determine the adulteration (Subramanian *et al.*, 2011).

Most of the best performance in the last years utilize a combination of NIR spectroscopy with partial least squares (PLS) for detecting adulteration in food and agro-products. As reported by Bázár *et al.* (2016) in honey, Mabood *et al.* (2017) in camel milk, Mabood *et al.* (2018) in fruit juice, and Kaufmann *et al.* (2022) in coriander oil. All studies provide a coefficient of correlation above 0.9. However, several studies use machine learning to analyze spectral data from NIR spectroscopy and compare it with the PLS algorithm. Recently, the machine learning approach is appropriate because several spectral data NIR issues need various techniques to handle overlapping classes, diverse samples, high sample sizes and nonlinear relationships in data as well (Cardoso and Poppi, 2021). The reasons above are the initial sources of non-linearity in the NIR spectra when associated with the response. Therefore, chemometricians continue looking for alternative methods to overcome this by utilizing algorithms from machine learning that can work with both linear and non-linear datasets for efficiency.

Although PLS algorithms are popularly applied to generate calibration models from NIR data spectral, machine learning algorithms have benefits over the already investigated techniques that work with linear and non-linear NIR spectroscopy datasets, including support vector machine (SVM) and artificial neural networks (ANN). For instance, some who reported using SVM included Cardoso and Poppi (2021) in commercial green tea blends with an accuracy between 82.0 to 93.0%, Cruz-Tirado *et al.* (2021) in egg with an accuracy of 87.0%, and Fowler *et al.* (2021) in lamb loin with a coefficient of determination between 0.61 to 0.65. Likewise, Valinger *et al.* (2021) and Puttipipatkajorn and Puttipipatkajorn (2020) reported that the use of NIR spectroscopy coupled with the machine learning algorithm artificial neural network (ANN) provided the best model accuracy for honey and rubber sheets, respectively. The coefficient of determination of both studies is between 0.8 to 0.99. Because there are many advantages of the mentioned techniques above indicates that machine learning techniques have great potential to be used in analyzing NIR spectroscopy data.

Generally, a robust calibration model is generated following appropriate data preprocessing stages because NIR spectroscopy is often subject to overwhelming background, light scattering, varying noises, and other unexpected factors. Preprocessing is crucial because it can transform raw data into clean data, which can significantly affect the accuracy of the developed model. Various preprocessing methods have been developed to eliminate the interference from these effects, and the selection methods currently available include trial and error, assessment of preprocessed data quality parameters, and expert visual inspection (Engel *et al.*, 2013). The first two methods are time-consuming to accomplish manually if considering all the existing preprocessing methods and their combination iterations simultaneously. Similarly, the third method also required a long time to become an expert in this matter.

In order to fill this gap, at least until now, some chemometricians are focused on efforts to develop an automatic preprocessing selection strategy for spectroscopy data through data sciences. At least started by Gerretzen *et al.* (2015), who decided to use the design of experiments to produce 16 iterations of combined preprocessing methods with none to four preprocessing steps combined with the PLS-DA algorithm. Furthermore, Bian *et al.* (2020) developed a selective ensemble preprocessing strategy from 120 iterations of combined preprocessing methods with none to four preprocessing steps combined with the PLS algorithm. It can be seen that this research focuses on preprocessing combination iterations without considering hyperparameter tuning in the algorithm and has not tried to extend it to other machine learning algorithms.

On the other hand, to produce a robust calibration model of a machine learning algorithm based on NIR spectroscopy data, optimization of each hyperparameter of the algorithm is required from the beginning. A hyperparameter is a parameter whose value is used to control the learning process, and it has to be tuned to reach good performance (Guido et al., 2022). Therefore, each machine learning algorithm has parameter components that must be tuned up and evaluated internally according to the existing data set. Hyperparameter values can significantly impact the resulting model's performance, but their selection requires particular

expertise and many labor-intensive manual iterations. Moreover, it is added by including several preprocessing methods in the combined iteration process.

With our best knowledge, no research has distinguished fresh (FCM), instant (ICM), and adulterated fresh coconut milk (A-FCM) using NIR spectroscopy in tandem with machine learning according to the literature published until now. Therefore, this study aims to classify the type of coconut milk (FCM, ICM, A-FCM) and predict the level of adulteration of water in fresh coconut milk using NIR spectroscopy and a machine learning approach for data analysis. A data sciences approach that combines appropriate preprocessing discovery and hyperparameter optimization concurrently of each algorithm is presented in this study as well. The output of this research is to obtain a novel model based on a combination of NIR spectroscopy and machine learning that can classify the type of coconut milk and predict the impurity of fresh coconut milk with the best preprocessing and optimum hyperparameter.

3.5 Materials and Methods

3.5.1 Sample Preparation

Fresh coconut milk is liquid coconut milk directly extracted from the matured endosperm (kernel) of the old coconut fruit without adding other ingredients, such as water, as a diluent. Instant coconut milk is liquid coconut milk concentrate obtained after partially removing moisture from fresh coconut milk and adding several types of chemicals as preservatives (CODEX-STAN-240, 2003). In Asia, instant coconut milk is usually referred to as commercially packaged liquid coconut milk products. In this study, fresh coconut milk without added water and instant coconut milk was purchased from local markets in Lad Krabang (Thailand) and stored at room temperature in a glass bottle. In the following text, the fresh coconut milk samples are denoted as FCM, the instant coconut milk samples are represented as ICM, and water-adulterated fresh coconut milk samples are indicated as A-FCM. Before the experiment, FCM was passed through the cloth filters 100-mesh sieve so that no impurities, such as dregs, were included in the sample. Furthermore, as many as 20 FCM samples were prepared for spectroscopy data acquisition. Also, as many as 50

ICM samples were prepared for spectral NIR scanning from 5 different commercially packaged brands and unknown batch productions.

Liquid adulterant materials using destilated water with coconut milk were intentionally-adulterated following percentage levels including 10, 20, 30, 40, 50, 60, 70, 80, and 90% (w/w). For homogenization, after adding water to fresh coconut milk, manually shaking in the 500 mL glass bottle was carried out for about 1 minute and waiting for the sample to go to room temperature before scanning. Then, each level of coconut milk adulteration (as many as 9 levels) was prepared with as many as 10 samples, totaling 90 samples A-FCM.

Thus, in the classification study of coconut milk into FCM, ICM, and A-FCM using a total of 160 NIR spectral data. After that, as many as 110 NIR spectral data were used to perform a quantitative study to predict the level of adulteration of fresh coconut milk from 0 to 90% (w/w). The total NIR spectra samples in this study were larger than that suggested by Manley (2014), who stated that at least 100 or more samples should be collected for model development.

3.5.2 Spectra Data Acquisition

The NIR spectral was measured with benchtop FT-NIR spectrometer (Bruker Ltd.) in reflection mode on 800–2500 nm (12,500–4000 cm^{-1}). The instrument was connected to software OPUS software package. Each sample of coconut milk was taken from a 500 mL glass bottle using a micropipette of about 1 mL. After that, a reflector with a path length of 0.35 mm was also put in a test glass vial (diameter of 20 mm and heigh of 43 mm) and placed on the FT-NIR spectrometer for scanning. A background spectrum against air was recorded before acquisition to minimize the influence of temporal baseline shifts. One average spectrum was obtained by an average of 32 scans at resolution of 8 cm^{-1}. Scan results were recorded in absorption mode (log 1/R) for each sample.

3.5.3 Chemometric Analysis

Both the classification as qualitative analysis and regression models as a quantitative analysis was created and tested individually based on the absorbance NIR spectra following the flowchart shown in Figure 3.1. Methods for developing classification models using machine learning algorithms, including linear discriminant

analysis (LDA), support vector machine (SVM) and multilayer perceptron (MLP) were employed in this study. Machine learning algorithms, including partial least squares (PLS), support vector machine (SVM), and multilayer perceptron (MLP), were utilized to develop a prediction model for the level of adulteration of fresh coconut milk. Because of the massive use of this machine learning algorithm not only for processing NIR spectra data but also for another purpose, the theoretical explanation of each algorithm is not explained further in this article and can be seen in previous publications (Chu *et al.*, 2022).

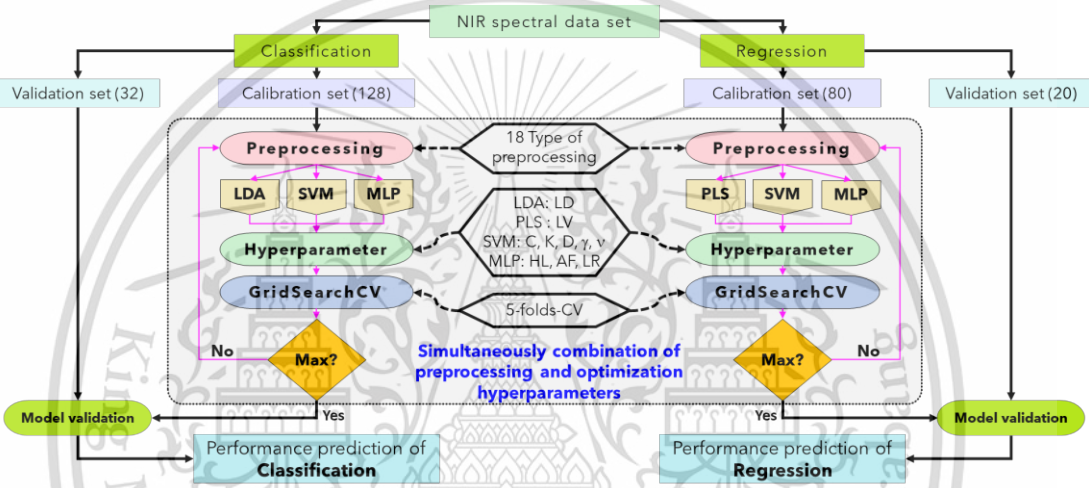


Figure 3.1. Proposed overall methodology for chemometrics analysis.

Exhaustive search over specified parameter values for optimization of each estimator through hyperparameters of each machine learning algorithm implemented using Grid-search approach (GridSearchCV). The preprocessing method and range tuning of hyperparameters of the machine learning algorithm are optimized exhaustively with the values and ranges presented in Table 3.1 and Table 3.2, respectively. Also, the auto-preprocessing strategy, a combination of 18 types of preprocessing (including original data), is used simultaneously in optimizing the hyperparameters of individual machine learning algorithms via the GridSearchCV command.

From the total dataset, a random splitting strategy was employed, i.e. 80% of the experimental samples were used in the calibration dataset (as training), and 20% for the prediction dataset (as validation) (Sankaran *et al.*, 2011). The statistical parameters of the calibration and prediction dataset are shown in Table 3.3.

The calibration dataset was applied 5-folds cross-validation to validate the models internally over a parameter grid (preprocessing type and hyperparameter) for each machine learning algorithm to avoid over and underfitting issues in a model (Setiadi *et al.*, 2022). Cross-validation results are the average validation results from five folds. Finally, the calibrated model with the best preprocessing and optimal hyperparameters was tested by the validation dataset (20%). Chemometric analyses were performed using the open-source, Scikit-learn machine learning package for Python 3.8.8, built on the scientific and numerical Python libraries Scipy and Numpy (Pedregosa *et al.*, 2011).

Table 3.1. Detailed of preprocessing methods used in this study.

Type of preprocessing	Abbreviation	Code in programming
Raw data	RD	None
Standard normal variate	SNV	SNV ()
Multiplicative scatter correction	MSC	MSC ()
Mean scaling	MnS	MeanScaling ()
Median scaling	MdS	MedianScaling ()
Max scaling	MxS	MaxScaling ()
Range scaling	RS	RangeScaling ()
Mean centering	MC	MeanCentering ()
Standard scaler	SS	StandardScaler ()
First derivative	FstD	FirstDerivative ()
Second derivative	SndD	SecondDerivative ()
Baseline Second Order (degree)	BSO (2)	BaselineSecondOrder (2)
	BSO (3)	BaselineSecondOrder (3)
	BSO (4)	BaselineSecondOrder (4)
Savitzky–Golay filter (window, polynomial order)	SGF (5, 2)	SavgolFilter (5, 2)
	SGF (9, 2)	SavgolFilter (9, 2)
	SGF (11, 2)	SavgolFilter (11, 2)
	SGF (5, 3)	SavgolFilter (5, 3)

Three groups of coconut milk samples will be studied qualitatively for distinction, including FCM, ICM, and A-FCM. The criteria to classify coconut milk samples in this study refer to the international FAO CODEX STAN 240-2003 standard

(CODEX-STAN-240, 2003). This standard has categorized coconut milk products into four categories, of which in this study, only three groups of the four groups were studied, including light coconut milk, which is represented by the A-FCM sample, coconut milk, which is represented by the FCM sample and coconut cream concentrate which is represented by the sample ICM. The standard categories of coconut milk products depend on total solids, non-fat solids, total fat, and moisture content. Model evaluation in machine learning, including precision (Pr), recall (Rc), F1-score (Fs), and accuracy (Ac) are the common criteria used to evaluate the performance of classification models, which is defined by Equation (3.1) until Equation (3.4). Classification results model were determined as false positive (FP), false negative (FN), true positive (TP), and true negative (TN).

$$Pr = \frac{TP}{TP+FP} \tag{3.1}$$

$$Rc = \frac{TP}{TP+FN} \tag{3.2}$$

$$Fs = \frac{2 \times Pr \times Rc}{Pr + Rc} \tag{3.3}$$

$$Ac = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.4}$$

Table 3.2. Range tuning of hyperparameters from machine learning algorithm.

Algorithm	Hyperparameters	Tuning range
Classification		
LDA	LD	1, 3, 5, 7, 9, 11
SVM	C	1, 10, 100
	Kernel	Linear, Polynomial, RBF, Sigmoid
	Degree	2, 3, 4
	Gamma	Scale, Auto
MLP	Hidden layer sizes	(3, 5, 7), (3, 5), (5, 7), (5), (100), (100, 100), (1000)
	Activation function	Identity, Logistic, Tanh, ReLU
	Learning rate initial	0.01, 0.1, 1.0
Regression		
PLS	LV	5, 7, 9, 11
SVM	C	1, 10, 100
	Kernel	Linear, Polynomial, RBF, Sigmoid
	Degree	2, 3, 4
	Gamma	Scale, Auto
	Nu	0.1, 0.5, 1.0

MLP	Hidden layer sizes	(3, 5, 7), (3, 5), (5, 7), (5), (100), (100, 100), (1000)
	Activation function	Identity, Logistic, Tanh, ReLU
	Learning rate initial	0.01, 0.1, 1.0

As many as ninety adulteration samples from fresh coconut milk (A-FCM) and twenty pure fresh coconut milk (FCM) will be studied quantitatively for predictive levels of its adulteration. The statistical parameter concepts used in this article include the coefficient of determination for calibration and prediction (Equation 3.5), root mean square error for calibration and prediction (Equation 3.6), and ratio of prediction to deviation (Equation 3.7). The coefficient of determination (R^2) represents the proportion of variance of the response variable obtained and how certain its models can be making predictions by the spectral features in the calibration and validation model. Accuracy is evaluated by root mean square error in calibration (RMSEc) and prediction (RMSEp). The robustness test from the model is evaluated by ratio of prediction to deviation (RPD). The optimal model should have a high R^2, a low RMSE and Bias (Equation 3.7), and high RPD (Equation 3.8) (Janse Van Vuuren and Groenewald, 2013).

Table 3.3. Statistics of the calibration and validation dataset.

Subset	Calibration			Validation		
Classify						
Samples	FCM	ICM	A-FCM	FCM	ICM	A-FCM
Number samples per class	18	40	70	2	10	20
Total number samples	128			32		
Regression						
Item	Min-max	Mean	SD	Min-max	Mean	SD
Range value	0 - 90	47.13	29.39	0 - 90	36.50	25.60
Total number samples	80			20		

$$R_c^2 \text{ or } R_{cv}^2 \text{ or } R_p^2 = 1 - \frac{\sum_{i=1}^{n}\left(E_i - P_i\right)^2}{\sum_{i=1}^{n}\left(E_i - \bar{E}\right)^2} \quad (3.5)$$

$$\text{RMSEc or RMSEp} = \sqrt{\frac{\sum_{i=1}^{n}\left(E_i - P_i\right)^2}{n}} \quad (3.6)$$

$$\text{Bias} = \frac{\sum_{i=1}^{n}\left(E_i - P_i\right)}{n} \quad (3.7)$$

$$RPD = \frac{SD}{SEp} = \frac{1}{\sqrt{1 - R_p^{\ 2}}} \quad (3.8)$$

Where R_c^2 is the coefficient of determination of calibration, R_{cv}^2 is the coefficient of determination of cross-validation, R_p^2 is the coefficient of determination of prediction, E_i is the existing value for point to-*i*, P_i is the prediction value for point to-*i*, *n* is the number of samples, RMSEc is root mean square error of calibration, RMSEp is root mean square error of prediction, $\overline{E}$ is average of existing value, RPD is ratio of prediction to deviation, SD is standard deviation, and SEp is standard error of prediction.

3.6 Results and Discussions

3.6.1 Spectra Profiles

Figure 3.2 showed the raw NIR spectra of FCM, ICM, and A-FCM. Because of coconut milk is mainly composed of moisture and fat so similar in chemical composition, it can be seen that the originating spectral waveform and absorption peak location (at 1215 nm (8230 cm^{-1}), 1450 nm (6897 cm^{-1}), 1730 nm (5780 cm^{-1}), 1765 nm (5666 cm^{-1}), 1930 nm (5181 cm^{-1}), 2310 nm (4329 cm^{-1}), and 2348 nm (4259 cm^{-1})) of the spectra look closely similar except for the subtle difference in absorbance. According to Osborne *et al.* (1993), the peaks at 1210 nm (8230 cm^{-1}), 1730 nm (5780 cm^{-1}), and 1760 nm (5682 cm^{-1}) are due to the absorption band of the second and first overtones associated with the CH_2, but in this study, they are shifted from 1215 nm (8230 cm^{-1}), and 1725 nm (5797 cm^{-1}), respectively. In contrast to the peaks at 1450 nm (6897 cm^{-1}), the absorption band of the 1st overtones is associated with the OH stretching of water where there is no shift, and at peak 1930 nm (5181 cm^{-1}), shifted from 1940 nm (5155 cm^{-1}), which is attributed to OH combination. There were also absorption peaks of the CH_2 structure at 2310 nm (4329 cm^{-1}), and 2348 nm (4259 cm^{-1}) shifted from 2347 nm (4261 cm^{-1}), which was attributed to $HC{=}CHCH_2$. Therefore, it is difficult to differentiate the type of coconut milk (FCM, ICM, A-FCM) and even determine the adulteration level directly without the chemometrics method.

As mentioned by several previous research results (Alyaqoubi *et al.*, 2015), the main components of coconut milk, consisting of fat, protein, moisture, ash, and

carbohydrates, have been confirmed through several intensity absorbance peaks in this study. The spectra show the characteristics bands of carbohydrates such as starch and sugars at 2101 nm (4760 cm^{-1}), 1490 nm (6711 cm^{-1}), 1484 nm (6739 cm^{-1}); proteins at 2050–2060 nm (4878–4854 cm^{-1}), 1500–1530 nm (6667–6536 cm^{-1}); fats at 2381 nm (4200 cm^{-1}), 2070 nm (4831 cm^{-1}) and moisture at 1947 nm (5136 cm^{-1}), 1203 nm (8313 cm^{-1}) (Workman and Weyer, 2007). If we compared the intensity of spectra, the ICM absorbance spectra seemed lower than FCM. We can assume this is due to the difference in the samples' moisture amount, as Osborne *et al.* (1993) reported that NIR spectroscopy is sensitive due to differences in moisture content. Apart from that, the thickening treatment from the FCM sample to ICM causes the possibility of a little shift in the amount of fat, protein, ash, and carbohydrate content of the coconut milk. However, a slight shift in the amount should indicate that the two can be differentiated based on spectroscopy, where the naked eye cannot differentiate between the two.

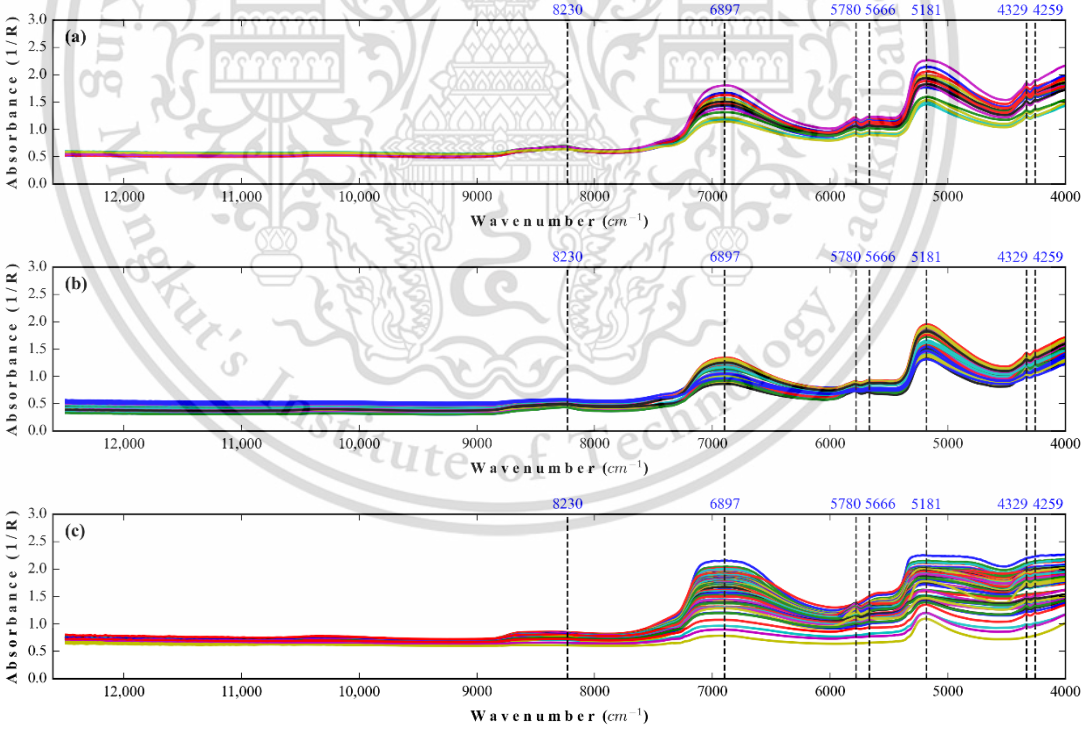
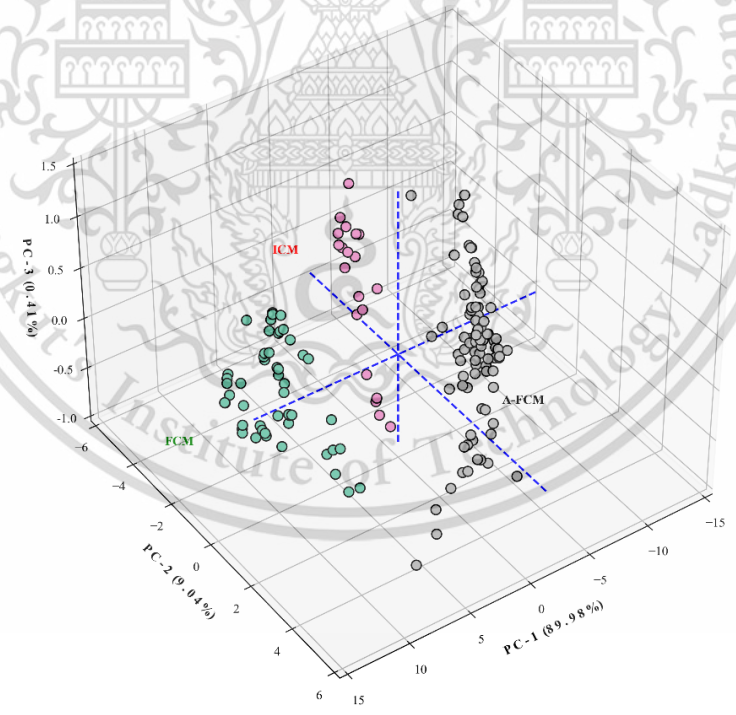


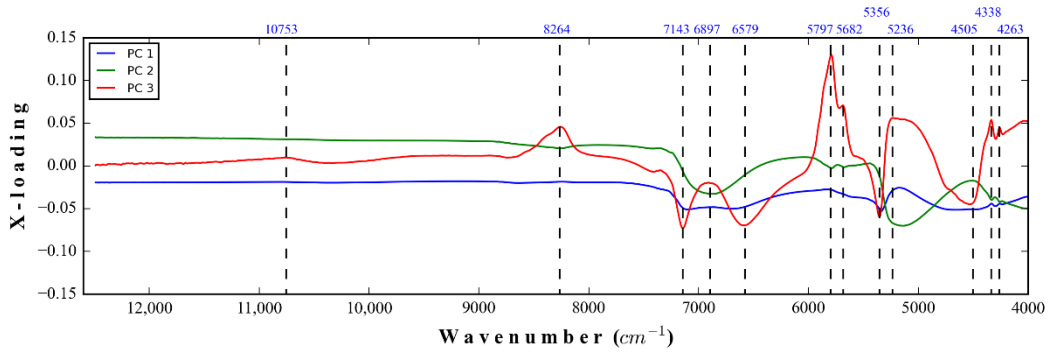
Figure 3.2. NIR spectra of samples (a) FCM (b) ICM (c) A-FCM.

3.6.2 PCA Scores Scatter Plot Analysis

Principal component analysis (PCA) was performed on the raw NIR spectra of the samples to determine their classification change characteristics preliminarily and to see the structure of the population by analyzing the scores obtained from the PCA of the samples. Figure 3.3a displays the three-dimensional score scatter plot obtained through the PCA of the original based data of all the samples. The first three PCs explained 99.43% of the total variance. Different types of samples tended to exhibit trends of separation, and samples of the same type exhibited significant aggregation. These findings indicate the feasibility of NIR spectroscopy for identifying coconut milk samples based on their type. Nevertheless, some samples of A-FCM are slightly scattered for each level of adulteration, and there is a clear difference between the main groups of samples. Thus, PCA could be used to visualize samples directly. Moreover, LDA, SVM, and MLP, which are supervised machine learning algorithms, were used to establish classification models for all the samples.



(a)



(b)

Figure 3.3. Principal component analysis plot (a) 3D scores scatter of PCA, (b) The first three X-loading lines.

The X-loading lines of the first three PCs are displayed in Figure 3.3b. The high positive and negative X-loading weights show the strong effect of the bond vibration on the classification of coconut milk samples based on their type. In our case, the peaks can be seen at 930 nm (10,753 cm^{-1}), 1210 nm (8264 cm^{-1}), 1400 nm (7143 cm^{-1}), 1450 nm (6897 cm^{-1}), 1520 nm (6579 cm^{-1}), 1725 nm (5797 cm^{-1}), 1760 nm (5682 cm^{-1}), 1860 nm (5376 cm^{-1}), 1910 nm (5236 cm^{-1}), 2220 nm (4505 cm^{-1}), 2305 nm (4338 cm^{-1}), and 2346 nm (4263 cm^{-1}). The vibration bands with a high X-loading from the PCA will be corresponding vibration bands of some functional group (structure). More details of all the corresponding functional groups found in this study are concluded in Table 3.4.

Table 3.4. The vibration bands with a high X-loading from the PCA.

Wavenumber (cm^{-1})	Wavenumber of references (cm^{-1})	Band vibration and functional group	Ref.
10753	10753	CH methylene (CH_2)	Workman and Weyer (2007)
8264	8333 – 8258	-CH_2, 2nd overtone of sym. stretching	Conzen (2006)
7143	7189 – 7138	-CH_2, 2×CH-stretching + 3×CH-bending	Conzen (2006)
	7168 – 7018	Free-OH_2, 1st overtone	
6897	6901 – 6849	1st overtone H_2O	Conzen (2006)
	6901 – 6780	-NH_2 primary amines, 1st overtone, ArNH_2	
	6940 – 6849	-$CONH_2$ prim. Amides, 1st overtone of NH_2 anti-sym. Stretching	

6579	6671 – 6270	Bound-OH, 1st overtone, intermolecular hydrogen bond	Conzen (2006)
	6711 – 6468	NH secondary amine, 1st overtone	
	6618 – 6540	-$CONH_2$ prim. amides, 1st overtone, intermolecular hydrogen bond	
5797	5851 – 5780	-CH_3, 1st overtone of anti-sym. Stretching	Conzen (2006)
5682	5701 – 5631	-CH, 1st overtone	Conzen (2006)
5376	5376	C-Cl chlorinated organics (C-Cl group)	Workman and Weyer (2007)
5236	5291 – 5211	-COOH, -COOR, 2nd overtone, 2×C=O-stetching (carbox.acids)	Conzen (2006)
	5241 – 5179	-CONH- second. Amides, 2nd overtone of amide	
4505	4505	Asymmetrical NH stretch + NH_2 rocking	Workman and Weyer (2007)
4338	4340 – 4320	-CH_3, combination of CH-stretching + CH-bending	Conzen (2006)
	4386 – 4292	CH stretching and CH_2 deformation combinations	Workman and Weyer (2007)
2346	2320 – 2331	-CH_3, combination of CH-stretching + CH-bending	Conzen (2006)
	2342 – 2347	-C=C, combination CH_2-stretching + =CH_2-bending	

3.6.3 Classification Models

A total of 108 combinations of results from preprocessing and hyperparameter optimization (18 types of preprocessing and 6 levels of linear discriminant (LDs) components of hyperparameters) from the LDA algorithm was tested to find their best combination in accuracy through 5-folds cross-validation presented in Table 3.5 (they are presented in full in Table S2-1 in supplementary material). For illustration, Table 3.5 shows 12 types of preprocessing, and just only using one hyperparameter of linear discriminant (LD) component produces the same good accuracy, which is 0.99. After that, the accuracy value produced by preprocessing followed by hyperparameter optimization continued to decrease until it was not defined in the preprocessing range scaling type with an LD of 11. This study shows that with the application of the LDA algorithm, there will be more than one candidate model that

can produce the best classification calibration model for differentiating types of coconut milk (FCM, ICM, A-FCM) based on the NIR spectral. Therefore, an automation concept that can combine both at the same time, preprocessing and optimizing hyperparameters, as was done in this study, will be very necessary and will make it more effortless to develop spectroscopy-based calibration models combined with machine learning algorithms in the future.

Table 3.5. The best preprocessing and hyperparameters using LDA algorithm.

No	Preprocessing	Hyperparameter (LD)	Accuracy (5-folds CV)
1	RD	1	0.99 ± 0.02
2	SNV	1	0.99 ± 0.02
3	MSC	1	0.99 ± 0.02
4	MnS	1	0.99 ± 0.02
5	Mds	1	0.99 ± 0.02
6	MxS	1	0.99 ± 0.02
7	SGF (5, 2)	1	0.99 ± 0.02
8	SGF (9, 2)	1	0.99 ± 0.02
9	SGF (5, 3)	1	0.99 ± 0.02
10	MC	1	0.99 ± 0.02
11	SS	1	0.99 ± 0.02
12	RS	1	0.99 ± 0.02

A total of 1296 combinations of results from preprocessing and hyperparameter optimization (18 types of preprocessing and 3, 4, 3, and 2 levels of hyperparameter C, kernel, degree, and gamma, respectively) from the SVM algorithm were tested to find their best combination in accuracy through 5-folds cross-validation presented in supplementary material Table S2-2. It is verified that by using the SVM algorithm to classify coconut milk types into FCM, ICM, and A-FCM based on NIR spectra, there are at least 215 combinations of preprocessing and hyperparameters that show a perfect level of accuracy. In Table 3.6, the combination is dominated by six preprocessing types: SNV, mean scaling, median scaling, max scaling, standard scaler, and range scaling. The first four of the six preprocessing, according to Engel *et al.* (2013) and Bian *et al.* (2020), are to be able to perform scatter correction noise on the spectra. This is because scatter noise is common to

all analytical techniques that involve light, including infrared (IR), near-infrared (NIR), and UV-spectroscopy. Light scattering can originate from physical differences between samples, including particle size and shape, sample packing, and sample surface (Li *et al.*, 2019).

Table 3.6. The best preprocessing and hyperparameters using SVM algorithm.

No	Preprocessing	Hyperparameter (C, Kernel, Degree, Gamma)	Accuracy (5-folds CV)
1-36	SNV	1, linear, 2, scale; 1, rbf, 2, scale; 1, linear, 2, auto; 1, rbf, 2, auto; 1, linear, 3, scale; 1, rbf, 3, scale; 1, linear, 3, auto; 1, rbf, 3, auto; 1, linear, 4, scale; 1, rbf, 4, scale; 1, linear, 4, auto; 1, rbf, 4, auto; 10, linear, 2, scale; 10, rbf, 2, scale; 10, linear, 2, auto; 10, rbf, 2, auto; 10, linear, 3, scale; 10, rbf, 3, scale; 10, linear, 3, auto; 10, rbf, 3, auto; 10, linear, 4, scale; 10, rbf, 4, scale; 10, linear, 4, auto; 10, rbf, 4, auto; 100, linear, 2, scale; 100, rbf, 2, scale; 100, linear, 2, auto; 100, rbf, 2, auto; 100, linear, 3, scale; 100, rbf, 3, scale; 100, linear, 3, auto; 100, rbf, 3, auto; 100, linear, 4, scale; 100, rbf, 4, scale; 100, linear, 4, auto; 100, rbf, 4, auto	1.0 ± 0.0
37-75	MnS	1, linear, 2, scale; 1, poly, 2, scale; 1, rbf, 2, scale; 1, linear, 2, auto; 1, linear, 3, scale; 1, poly, 3, scale; 1, rbf, 3, scale; 1, linear, 3, auto; 1, linear, 4, scale; 1, poly, 4, scale; 1, rbf, 4, scale; 1, linear, 4, auto; 10, linear, 2, scale; 10, poly, 2, scale; 10, rbf, 2, scale; 10, linear, 2, auto; 10, linear, 3, scale; 10, poly, 3, scale; 10, rbf, 3, scale; 10, linear, 3, auto; 10, linear, 4, scale; 10, poly, 4, scale; 10, rbf, 4, scale; 10, linear, 4, auto; 100, linear, 2, scale; 100, poly, 2, scale; 100, rbf, 2, scale; 100, linear, 2, auto; 100, poly, 2, auto; 100, linear, 3, scale; 100, poly, 3, scale; 100, rbf, 3, scale; 100, linear, 3, auto; 100, poly, 3, auto; 100, linear, 4, scale; 100, poly, 4, scale; 100, rbf, 4, scale; 100, linear, 4, auto; 100, poly, 4, auto	1.0 ± 0.0
76-110	MdS	1, linear, 2, scale; 1, poly, 2, scale; 1, rbf, 2, scale; 1, linear, 2, auto; 1, linear, 3, scale; 1, poly, 3, scale; 1, rbf, 3, scale; 1, linear, 3, auto; 1, linear, 4, scale; 1, rbf, 4, scale; 1, linear, 4, auto; 10, linear, 2, scale; 10, poly, 2, scale; 10, rbf, 2, scale; 10, linear, 2, auto; 10, linear, 3, scale; 10, poly, 3, scale; 10, rbf, 3, scale; 10, linear, 3, auto; 10, linear, 4, scale; 10, rbf, 4, scale; 10, linear, 4, auto; 100, linear, 2, scale; 100, poly, 2, scale; 100, rbf, 2, scale; 100, linear, 2, auto; 100, poly, 2, auto; 100, linear, 3, scale; 100, poly, 3, scale; 100, rbf, 3, scale; 100, linear, 3, auto; 100, poly, 3, auto; 100, linear, 4, scale; 100, rbf, 4, scale; 100, linear, 4, auto	1.0 ± 0.0
111-149	MxS	1, linear, 2, scale; 1, poly, 2, scale; 1, rbf, 2, scale; 1, linear, 2, auto; 1, linear, 3, scale; 1, poly, 3, scale; 1, rbf, 3, scale; 1, linear, 3, auto; 1, linear, 4, scale; 1, poly, 4, scale; 1, rbf, 4, scale; 1, linear, 4, auto; 10, linear, 2, scale; 10, poly, 2, scale; 10, rbf, 2, scale; 10, linear, 2, auto; 10, linear, 3, scale; 10, poly, 3, scale; 10, rbf, 3, scale; 10, linear, 3, auto; 10, linear, 4, scale; 10, poly, 4, scale; 10, rbf, 4, scale; 10, linear, 4, auto; 100, linear, 2, scale; 100, poly, 2, scale; 100, rbf, 2, scale; 100, linear, 2, auto; 100, rbf, 2, auto; 100, linear, 3, scale; 100, poly, 3, scale; 100, rbf, 3, scale; 100, linear, 3, auto; 100, rbf, 3, auto; 100, linear, 4, scale; 100, poly, 4, scale; 100, rbf, 4, scale; 100, linear, 4, auto; 100, rbf, 4, auto	1.0 ± 0.0
150-185	SS	1, linear, 2, scale; 1, rbf, 2, scale; 1, linear, 2, auto; 1, rbf, 2, auto; 1, linear, 3, scale; 1, rbf, 3, scale; 1, linear, 3, auto; 1, rbf, 3, auto; 1, linear, 4, scale; 1,	1.0 ± 0.0

		rbf, 4, scale; 1, linear, 4, auto; 1, rbf, 4, auto; 10, linear, 2, scale; 10, rbf, 2, scale; 10, linear, 2, auto; 10, rbf, 2, auto; 10, linear, 3, scale; 10, rbf, 3, scale; 10, linear, 3, auto; 10, rbf, 3, auto; 10, linear, 4, scale; 10, rbf, 4, scale; 10, linear, 4, auto; 10, rbf, 4, auto; 100, linear, 2, scale; 100, rbf, 2, scale; 100, linear, 2, auto; 100, rbf, 2, auto; 100, linear, 3, scale; 100, rbf, 3, scale; 100, linear, 3, auto; 100, rbf, 3, auto; 100, linear, 4, scale; 100, rbf, 4, scale; 100, linear, 4, auto; 100, rbf, 4, auto	
186-215	RS	1, linear, 2, scale; 1, rbf, 2, scale; 1, linear, 2, auto; 1, linear, 3, scale; 1, rbf, 3, scale; 1, linear, 3, auto; 1, linear, 4, scale; 1, rbf, 4, scale; 1, linear, 4, auto; 10, linear, 2, scale; 10, rbf, 2, scale; 10, linear, 2, auto; 10, linear, 3, scale; 10, rbf, 3, scale; 10, linear, 3, auto; 10, linear, 4, scale; 10, rbf, 4, scale; 10, linear, 4, auto; 100, linear, 2, scale; 100, rbf, 2, scale; 100, linear, 2, auto; 100, sigmoid, 2, auto; 100, linear, 3, scale; 100, rbf, 3, scale; 100, linear, 3, auto; 100, sigmoid, 3, auto; 100, linear, 4, scale; 100, rbf, 4, scale; 100, linear, 4, auto; 100, sigmoid, 4, auto	1.0 ± 0.0

A total of 1512 combinations of results from preprocessing and hyperparameter optimization (18 types of preprocessing and 7, 4, and 3 levels/types of hyperparameter hidden layer sizes, activation function, and learning rate initial, respectively) from the MLP algorithm were tested to find their best combination in accuracy through 5-folds cross-validation which is presented in supplementary material Table S2-3. After extracting and showing in Table 3.7, it can be seen that there are at least two possible types of preprocessing with different hyperparameter structures that produce the same accuracy performance in the use of the MLP algorithm. Both types of preprocessing are eligible for the MLP algorithm, according to Engel *et al.* (2013), to carry out scaling and transformations of spectral data processing. In other words, mean centering and range scaling are scaling techniques to normalize all variables into a specific range to match the MLP algorithm's hyperparameters.

Table 3.7. The best preprocessing and hyperparameters using MLP algorithm.

No	Preprocessing	Hyperparameter (Hidden layer sizes, Activation function, Learning rate initial)	Accuracy (5-folds CV)
1	MC	5, ReLU, 0.01	1.0 ± 0.0
2	RS	(3, 5, 7), identity, 0.01	1.0 ± 0.0

All the classification algorithms used in this study achieved a discrimination rate of 100% (Table 3.8). The LDA algorithm achieved high classification accuracy for the

calibration and prediction sets without preprocessing and using only one LD as a predictor. The SVM method exhibited the same classification accuracy (for the calibration and validation) as the MLP algorithm. This result indicated a linear relationship in the classification problem. However, the SVM algorithm confirms that it can adjust its analysis (both linear and nonlinear) so that they perform as well as the LDA algorithm (Li *et al.*, 2021). This can be seen from the optimized hyperparameters in the SVM algorithm, it is recommended to use a linear type of kernel. In line with the SVM algorithm, the MLP algorithm can also adjust its machine to be as good as the LDA algorithm in dealing with this situation by placing ReLU (rectified linear unit) as a kernel hyperparameter. The kernel of ReLU, in general, is a piecewise linear function that will output the input directly if it is positive, otherwise, it will output zero (Katz *et al.*, 2022). Therefore, LDA, SVM, and MLP-based models could be used with NIR spectral data to classify the three groups of coconut milk samples accurately.

Table 3.8. Comparison of confusion matrix among the classification models.

Model	Preprocessing	Hyper-parameter	Spliting	Class (Sample number)	Pr	Rc	Fs	Ac
LDA	RD	LD=1	Calibration	FCM(18)	1.0	1.0	1.0	1.0
				ICM(40)	1.0	1.0	1.0	
				A-FCM(70)	1.0	1.0	1.0	
			Validation	FCM(2)	1.0	1.0	1.0	1.0
				ICM(10)	1.0	1.0	1.0	
				A-FCM(20)	1.0	1.0	1.0	
SVM	SNV	C=1.0, Kernel=linear, Degree=2, Gamma=scale	Calibration	FCM(18)	1.0	1.0	1.0	1.0
				ICM(40)	1.0	1.0	1.0	
				A-FCM(70)	1.0	1.0	1.0	
			Validation	FCM(2)	1.0	1.0	1.0	1.0
				ICM(10)	1.0	1.0	1.0	
				A-FCM(20)	1.0	1.0	1.0	
MLP	MC	Hidden layer sizes=(5), Activation function=ReLU, Learning rate initial=0.01	Calibration	FCM(18)	1.0	1.0	1.0	1.0
				ICM(40)	1.0	1.0	1.0	
				A-FCM(70)	1.0	1.0	1.0	
			Validation	FCM(2)	1.0	1.0	1.0	1.0
				ICM(10)	1.0	1.0	1.0	
				A-FCM(20)	1.0	1.0	1.0	

3.6.4 Regression Models

A total of 72 combinations of results from preprocessing and hyperparameter optimization (18 types of preprocessing and 4 levels of LV hyperparameters) from the PLS algorithm were tested to find their best combination in R^2 through 5-folds cross-validation which is presented in Table S2-4 of the supplementary material. This study shows that by applying the PLS algorithm together with the combination of preprocessing and hyperparameter optimization, there is only one best type of preprocessing followed by the optimal LV hyperparameter component and produces an R_{cv}^2 of 0.92 to predict the adulteration level of fresh coconut milk. After that, the other combinations produced an R_{cv}^2 value that decreased until the lowest was 0.53. This implies that inappropriate preprocessing and hyperparameters are detrimental to the predictive power of a chemometric model. Moreover, the existing preprocessing selection combined with the optimization of the hyperparameters of a machine learning algorithm will produce the possibility of a huge number of model combinations being tried and tested, and the process of exploring manually is a very time-consuming procedure. Even though according to Gerretzen *et al.* (2015), the straightforward way to get the best calibration model is by evaluating all possible strategies and selecting the optimal one.

A total of 3888 combinations of results from preprocessing and hyperparameter optimization (18 type preprocessing and 3, 4, 3, 2, and 3 level hyperparameters C, kernel, degree, gamma, and nu, respectively) from the SVM algorithm were tested to find their best combination in R^2 through 5-folds cross-validation which is presented in supplementary material Table S2-5. Applying the SVM algorithm to predict the adulteration level of fresh coconut milk with preprocessing and hyperparameter optimization produces at least 12 models that are as good as with an R_{cv}^2 of 0.90 (Table 3.9). The best model is dominated by two types of preprocessing, including standard scaler and SNV. The standard scaler preprocessing method from the scikit-learn library works for standardization by removing the mean and scaling to unit variance from the feature. It has been reported for its use in the prediction of the intramuscular fat content of lamb loin (Fowler *et al.*, 2021) and classification of bruise damage to apple fruit (Nturambirwe *et al.*, 2023) based on NIR spectral combined with machine learning. In addition, the

linear kernel type is optimal for the SVM algorithm, which shows that the level of adulteration of fresh coconut milk indicated a linear relationship with its spectroscopy properties.

A total of 1512 combinations of results from preprocessing and hyperparameter optimization (18 types of preprocessing and 7, 4, and 3 levels/types of hyperparameter hidden layer sizes, activation function, and learning rate initial, respectively) from the MLP algorithm were tested to find their best combination in R^2 through 5-folds cross-validation which is presented in supplementary material Table S2-6. It can be seen that the MLP algorithm generates only one best model to predict the level of adulteration of fresh coconut milk with an R_{cv}^2 of 0.86. Preprocessing type of standard scaler which functions to remove the mean and scaling to unit variance and is followed by a 2-layer neural network architecture where each neuron is 3 and 5 with an activation function of identity is the best combination in predicting the level of adulteration of fresh coconut milk. After that, R_{cv}^2 from the other combinations gradually decreased until it was undefined in the preprocessing type of combination Standard scaler with hidden layer sizes, activation functions, and initial learning rates of (100, 100), identity, and 1.0, respectively.

Table 3.9. The best preprocessing and hyperparameters using SVM algorithms to predict the level of fresh coconut milk adulteration.

No	Preprocessing	Hyperparameter (C, Kernel, Degree, Gamma, Nu)	R_{cv}^2 (5-folds CV)
1	SS	100, linear, 2, scale, 0.1	0.90 ± 0.05
2	SS	100, linear, 2, auto, 0.1	0.90 ± 0.05
3	SS	100, linear, 3, scale, 0.1	0.90 ± 0.05
4	SS	100, linear, 3, auto, 0.1	0.90 ± 0.05
5	SS	100, linear, 4, scale, 0.1	0.90 ± 0.05
6	SS	100, linear, 4, auto, 0.1	0.90 ± 0.05
7	SNV	100, linear, 2, scale, 0.1	0.90 ± 0.05
8	SNV	100, linear, 2, auto, 0.1	0.90 ± 0.05
9	SNV	100, linear, 3, scale, 0.1	0.90 ± 0.05
10	SNV	100, linear, 3, auto, 0.1	0.90 ± 0.05
11	SNV	100, linear, 4, scale, 0.1	0.90 ± 0.05
12	SNV	100, linear, 4, auto, 0.1	0.90 ± 0.05

The best performance test results with preprocessing and hyperparameters from the development of a calibration model to predict the level of adulteration of fresh coconut milk (A-FCM) are shown in Table 3.10. When comparing the performance of the machine learning algorithm for these studies, it is evident that no single preprocessing type and hyperparameter is superior for developing calibration models using NIR spectral data. In other words, many combinations of different preprocessing types and hyperparameters are suitable for developing calibration models based on NIR data combined with a machine learning algorithm. However, which preprocessing types and hyperparameters are optimal depends on the data set characteristics and the goal of data analysis. This raises the challenge of selecting preprocessing types and hyperparameters to generate the best calibration models. Obviously, this study provides a different perspective on solving this problem by presenting preprocessing and hyperparameter optimization combinations at once.

Table 3.10. Prediction results of determination of level adulteration of fresh coconut milk.

Model	Preprocessing	Hyper-parameter	Calibration		Prediction			
			R_c^2	RMSEc	R_p^2	RMSEp	Bias	RPD
PLS	SGF (11, 2)	LV=9	0.967	5.34	0.926	8.00	1.56	3.69
SVM	SS	C=100, Kernel=linear, Degree=2, Gamma=scale, Nu=0.1	0.984	3.74	0.931	8.30	2.63	3.80
MLP	SS	Hidden layer sizes=(3, 5), Activation function=identity, Learning rate initial=0.01	0.945	6.85	0.866	10.24	3.19	2.73

Scatter plots of the machine learning algorithm model to predict level adulteration of coconut milk from the validation are shown in Figure 3.4. According to Janse Van Vuuren and Groenewald (2013), RPD is the most meaningful statistic

used to evaluate the performance of calibrations. High values for the RPD indicate efficient NIR predictions and, at the same time, verify that SEp should be much lower than the SD. The PLS and SVM algorithms have an RPD value greater than three, which, according to Williams & Norris (2001) in Janse Van Vuuren and Groenewald (2013), is included in a fair model and can be utilized as screening application tools. Besides that, the model built with the MLP algorithm is included in the poor category with applications for very rough screening.

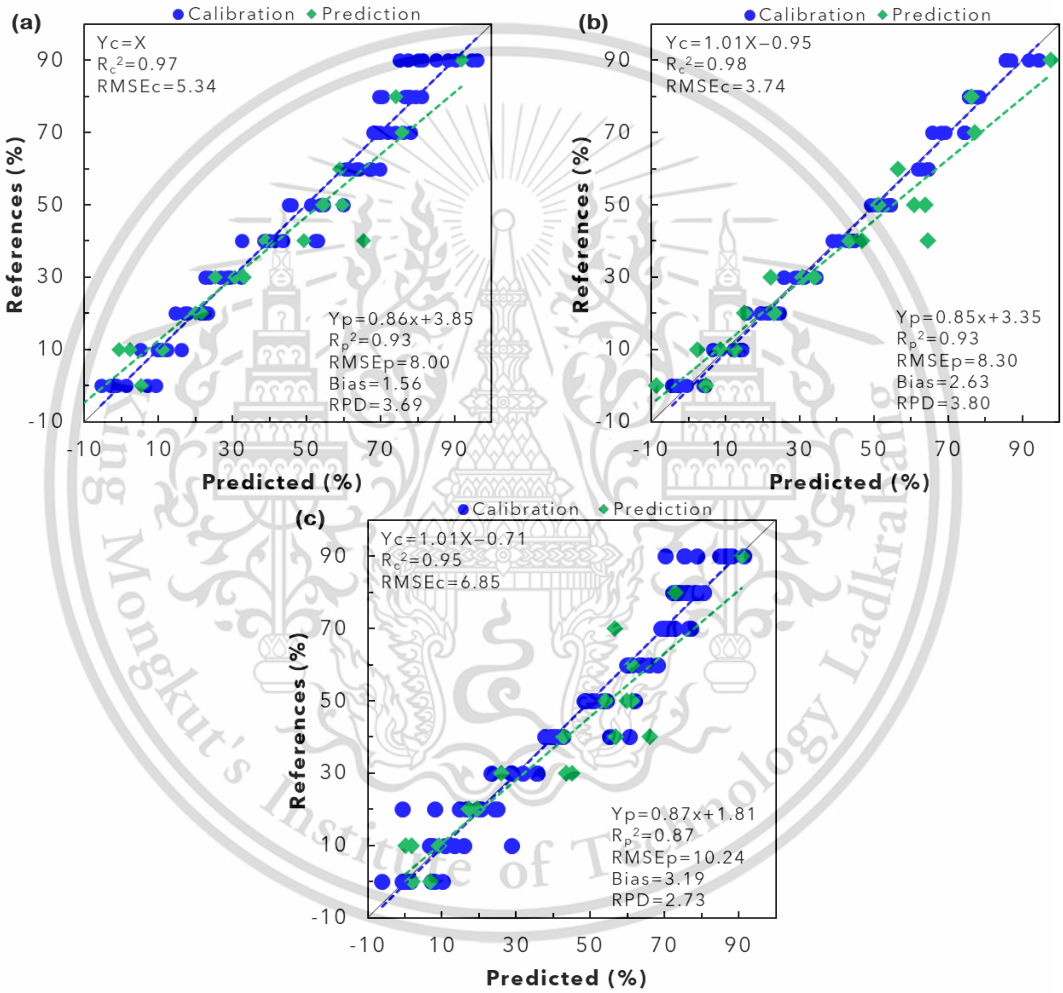


Figure 3.4. The plots of the prediction value versus the reference value of machine learning algorithm (a) PLS, (b) SVM, (c) MLP.

3.7 Conclusions

Near-infrared (NIR) spectroscopy combined with a precise machine learning algorithm can effectively identify the type of coconut milk and quantify the level of

water adulteration in fresh coconut milk; thus, this technique can be used to maintain the safety of coconut milk. One side is that data preprocessing to remove unwanted noise is crucial to achieving the data-analysis goals, including classification and regression. On the other hand, using machine learning algorithms requires hyperparameter tuning to achieve the best model. The combination of both in this study shows how challenging it can be to resolve which parameter can successfully help achieve these goals. However, breaking with the data sciences approach in solving this problem has been proposed and demonstrated in this paper.

The classification problem of coconut milk type (FCM, ICM, A-FCM) based on NIR spectroscopy shown here was solved perfectly (100% accuracy) using machine learning algorithms, including LDA, SVM, and MLP with the best preprocessing and optimum hyperparameters that have been found. Both in calibration and validation, a model generated by a machine learning algorithm can perform accurately. Each machine learning algorithm classification in this study has an optimal number of preprocessing and hyperparameter combinations of 12 for LDA, 215 for SVM, and 2 for MLP, respectively. From all the best compositions, one type of combination of preprocessing and optimal hyperparameters for LDA is chosen without preprocessing, and hyperparameter LD is 1. Furthermore, the SVM algorithm can use the SNV preprocessing method with hyperparameters C, kernel, degree, and gamma are 1, linear, 2, and scale, respectively. Finally, the MLP algorithm can use the mean centering preprocessing method with hidden layer sizes, activation functions, and learning rates are (5), ReLU, and 0.01, respectively.

Model-based on NIR spectroscopy to be able to predict the adulteration level of fresh coconut milk (0–90%) can also be generated using machine learning algorithms (PLS, SVM, MLP) in the performance ranges R^2 and RMSE at their calibrations of 0.95–0.98, and 6.85–3.74%, respectively. Next, R^2, RMSE, and RPD at the validation stage are 0.87–0.93, 10.24–8.00%, and 2.73–3.80, respectively. The most suitable preprocessing type and optimal hyperparameters for each algorithm in classification and regression can be found and optimized simultaneously. The best combination of preprocessing and hyperparameters for the PLS algorithm is using the SGF (11, 2) method, followed by an LV of 9. The SVM algorithm for the regression case in this study produces 12 combinations of preprocessing and hyperparameters

that have the same performance and further in this study use the preprocessing type standard scaler with hyperparameters C, kernel, degree, gamma, and nu are 100, linear, 2 scales, and 0.1, respectively. Finally, for the MLP algorithm, there is only one combination of the best preprocessing and hyperparameters with the type standard scaler and Hyperparameter hidden layer sizes, activation function, and initial learning rate are (3, 5), identity and 0.01, respectively.

The application of NIR spectroscopy technology with appropriate chemometrics can be exploited for non-destructive coconut milk quality monitoring in the future. Furthermore, as a chemometrics tool, the machine learning algorithm could guarantee classification accuracy and regression performance when used carefully. Last but not least, a data-analysis strategy can be used to obtain a robust model that combines the methods appropriate for preprocessing and optimizing hyperparameters simultaneously to achieve any possible data-analysis goals.

3.8 References

Alyaqoubi, S., Abdullah, A., Samudi, M., Abdullah, N., Addai, Z. R., & Musa, K. H. (2015). Study of antioxidant activity and physicochemical properties of coconut milk (Pati santan) in Malaysia. *Journal of Chemical and Pharmaceutical Research, 7*(4), 967-973.

Azlin-hashim, s., Siang, q. l., Yusof, F., Zainol, M. K., & Yusof, H. M. (2019). Chemical composition and potential adulterants in coconut milk sold in Kuala Lumpur. *Malaysian Applied Biology, 48*(3), 27-34.

Bala, M., Sethi, S., Sharma, S., Mridula, D., & Kaur, G. (2022). Prediction of maize flour adulteration in chickpea flour (besan) using near infrared spectroscopy. *Journal of Food Science and Technology, 59*(8), 3130-3138.

Bázár, G., Romvári, R., Szabó, A., Somogyi, T., Éles, V., & Tsenkova, R. (2016). NIR detection of honey adulteration reveals differences in water spectral pattern. *Food Chemistry, 194*, 873-880.

Bian, X., Wang, K., Tan, E., Diwu, P., Zhang, F., & Guo, Y. (2020). A selective ensemble preprocessing strategy for near-infrared spectral quantitative analysis of complex samples. *Chemometrics and Intelligent Laboratory Systems, 197*, 103916.

Cardoso, V. G. K., & Poppi, R. J. (2021). Non-invasive identification of commercial green tea blends using NIR spectroscopy and support vector machine. *Microchemical Journal, 164*, 106052.

Chu, X., Huang, Y., Yun, Y.-H., & Bian, X. (2022). *Chemometric methods in analytical spectroscopy technology*: Springer.

CODEX-STAN-240. (2003). Standard for Aqueous Coconut Products-Coconut Milk and Coconut Cream.: FAO/WHO Food Standards Programme.

Conzen, J.-P. (2006). *Multivariate calibration*.

Cruz-Tirado, J., da Silva Medeiros, M. L., & Barbin, D. F. (2021). On-line monitoring of egg freshness using a portable NIR spectrometer in tandem with machine learning. *Journal of Food Engineering, 306*, 110643.

Engel, J., Gerretzen, J., Szymanska, E., Jansen, J. J., Downey, G., Blanchet, L., & Buydens, L. M. C. (2013). Breaking with trends in pre-processing? *TrAC Trends in Analytical Chemistry, 50,* 96-106.

Faith Ndlovu, P., Samukelo Magwaza, L., Zeray Tesfay, S., & Ramaesele Mphahlele, R. (2022). Destructive and rapid non-invasive methods used to detect adulteration of dried powdered horticultural products: A review. *Food Research International, 157*, 111198.

Fowler, S. M., Wheeler, D., Morris, S., Mortimer, S. I., & Hopkins, D. L. (2021). Partial least squares and machine learning for the prediction of intramuscular fat content of lamb loin. *Meat Science, 177*, 108505.

Gerretzen, J., Szymanska, E., Jansen, J. J., Bart, J., van Manen, H.-J., van den Heuvel, E. R., & Buydens, L. M. C. (2015). Simple and Effective Way for Data Preprocessing Selection Based on Design of Experiments. *Analytical Chemistry, 87*(24), 12096-12103.

Guido, R., Groccia, M. C., & Conforti, D. (2022, 2022//). *Hyper-Parameter Optimization in Support Vector Machine on Unbalanced Datasets Using Genetic Algorithms.* Paper presented at the Optimization in Artificial Intelligence and Data Sciences, Cham.

Janse Van Vuuren, J., & Groenewald, C. (2013). Use of scanning near-infrared spectroscopy as a quality control indicator for bulk blended inorganic fertilizers. *Communications in soil science and plant analysis, 44*(1-4), 120-135.

Katz, G., Barrett, C., Dill, D. L., Julian, K., & Kochenderfer, M. J. (2022). Reluplex: a calculus for reasoning about deep neural networks. *Formal Methods in System Design, 60*(1), 87-116.

Kaufmann, K. C., Sampaio, K. A., García-Martín, J. F., & Barbin, D. F. (2022). Identification of coriander oil adulteration using a portable NIR spectrometer. *Food Control, 132*, 108536.

Kucharska-Ambrozej, K., & Karpinska, J. (2020). The application of spectroscopic techniques in combination with chemometrics for detection adulteration of some herbs and spices. *Microchemical Journal, 153*, 104278.

Lakshanasomya, N., Danudol, A., & Ningnoi, T. (2011). Method performance study for total solids and total fat in coconut milk and products. *Journal of Food Composition and Analysis, 24*(4), 650-655.

Lapcharoensuk, R., Danupattanin, K., Kanjanapornprapa, C., & Inkawee, T. (2020). *Combination of NIR spectroscopy and machine learning for monitoring chili sauce adulterated with ripened papaya.* Paper presented at the E3S Web of Conferences.

Li, L., Jin, S., Wang, Y., Liu, Y., Shen, S., Li, M., Ma, Z., Ning, J., & Zhang, Z. (2021). Potential of smartphone-coupled micro NIR spectroscopy for quality control of green tea. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 247*, 119096.

Li, L., Peng, Y., Li, Y., & Wang, F. (2019). A new scattering correction method of different spectroscopic analysis for assessing complex mixtures. *Analytica Chimica Acta, 1087*, 20-28.

Mabood, F., Hussain, J., Jabeen, F., Abbas, G., Allaham, B., Albroumi, M., Alghawi, S., Alameri, S., Gilani, S. A., & Al-Harrasi, A. (2018). Applications of FT-NIRS combined with PLS multivariate methods for the detection & quantification of saccharin adulteration in commercial fruit juices. *Food Additives & Contaminants: Part A, 35*(6), 1052-1060.

Mabood, F., Jabeen, F., Ahmed, M., Hussain, J., Al Mashaykhi, S. A. A., Al Rubaiey, Z. M. A., Farooq, S., Boqué, R., Ali, L., Hussain, Z., Al-Harrasi, A., Khan, A. L., Naureen, Z., Idrees, M., & Manzoor, S. (2017). Development of new NIR-spectroscopy method combined with multivariate analysis for detection of adulteration in camel milk with goat milk. *Food Chemistry, 221*, 746-750.

Manley, M. (2014). Near-infrared spectroscopy and hyperspectral imaging: non-destructive analysis of biological materials. *Chemical Society Reviews, 43*(24), 8200-8214.

Nturambirwe, J. F. I., Hussein, E. A., Vaccari, M., Thron, C., Perold, W. J., & Opara, U. L. (2023). Feature Reduction for the Classification of Bruise Damage to Apple Fruit Using a Contactless FT-NIR Spectroscopy with Machine Learning. *Foods, 12*(1), 210.

Osborne, B. G., Fearn, T., & Hindle, P. H. (1993). *Practical NIR spectroscopy with applications in food and beverage analysis*: Longman scientific and technical.

Panmanas, S., & Chin Hock, L. (2019). Rapid Evaluation of the Properties of Natural Rubber Latex and Its Products Using Near-Infrared Spectroscopy. In S. Arpit & Z. Elsayed (Eds.), *Organic Polymers* (pp. Ch. 2). Rijeka: IntechOpen.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research, 12*, 2825-2830.

Puttipipatkajorn, A., & Puttipipatkajorn, A. (2020). Development of calibration models for rapid determination of moisture content in rubber sheets using portable near-infrared spectrometers. *Journal of Innovative Optical Health Sciences, 13*(02), 2050009.

Ribeiro, J. S., Salva, T. d. J. G., & Silvarolla, M. B. (2021). Prediction of a wide range of compounds concentration in raw coffee beans using NIRS, PLS and variable selection. *Food Control, 125*, 107967.

Sankaran, S., Mishra, A., Maja, J. M., & Ehsani, R. (2011). Visible-near infrared spectroscopy for detection of Huanglongbing in citrus orchards. *Computers and Electronics in Agriculture, 77*(2), 127-134.

Setiadi, I. C., Hatta, A. M., Koentjoro, S., Stendafity, S., Azizah, N. N., & Wijaya, W. Y. (2022). Adulteration detection in minced beef using low-cost color imaging system coupled with deep neural network. *Frontiers in Sustainable Food Systems, 6*.

Simuang, J., Chiewchan, N., & Tansakul, A. (2004). Effects of fat content and temperature on the apparent viscosity of coconut milk. *Journal of Food Engineering, 64*(2), 193-197.

Sitorus, A., Muslih, M., Cebro, I. S., & Bulan, R. (2021). Dataset of adulteration with water in coconut milk using FTIR spectroscopy. *Data in Brief, 36*, 107058.

Sitorus, A., Pambudi, S., Boodnon, W., & Lapcharoensuk, R. (2023). Near-Infrared Spectroscopy with Machine Learning for Classifying and Quantifying Nutmeg Adulteration. *Analytical Letters*, 1-22.

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. *Current Opinion in Food Science, 10*, 45-51.

Subramanian, A., Alvarez, V. B., Harper, W. J., & Rodriguez-Saona, L. E. (2011). Monitoring amino acids, organic acids, and ripening changes in Cheddar cheese using Fourier-transform infrared spectroscopy. *International Dairy Journal, 21*(6), 434-440.

Valinger, D., Longin, L., Grbeš, F., Benković, M., Jurina, T., Gajdoš Kljusurić, J., & Jurinjak Tušek, A. (2021). Detection of honey adulteration – The potential of UV-VIS and NIR spectroscopy coupled with multivariate analysis. *LWT, 145*, 111316.

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.

Workman, J., & Weyer, L. (2007). *Practical guide to interpretive near-infrared spectroscopy*: CRC press.

Yu, D.-x., Guo, S., Zhang, X., Yan, H., Zhang, Z.-y., Chen, X., Chen, J.-y., Jin, S.-j., Yang, J., & Duan, J.-a. (2022). Rapid detection of adulteration in powder of ginger (Zingiber officinale Roscoe) by FT-NIR spectroscopy combined with chemometrics. *Food Chemistry: X, 15*, 100450.

CHAPTER 4 – CASE STUDY 2
DEVELOPED AN AUTOMATIC TUNING OF COMBINATION PREPROCESSING AND HYPERPARAMETER OF MACHINE LEARNING AND ITS APPLICATION TO NIR SPECTRAL DATA OF COCONUT MILK ADULTERATION[3]

4.1 Highlights

1. FT-NIR and Micro-NIR spectrometer are utilized to identify adulteration in coconut milk.
2. An automatic selection method for NIR preprocessing and hyperparameters was developed.
3. This method suggests that Micro-NIR with KNN classifier achieves up to 0.97 accuracy.
4. This method implies that Micro-NIR with KNN regressor achieves an RPD of more than 14.
5. This method worked for coconut milk NIR data and can applied to future other cases.

4.2 Abstract

This work proposes a novel approach to automatically select preprocessing and hyperparameters of machine learning algorithms based on their best performance in cross-validation for NIR spectroscopy data. The proposed method incorporates a single up to multiple preprocessing steps and tuning hyperparameters simultaneously that contribute to discovering the best model performances for FT-NIR and Micro-NIR spectral data of coconut milk adulteration by distilled water and coconut water in the range 0 to 50%. Computational experiments in Python language are developed and tested with nine single preprocessing types, three types of machine learning for classification (LDA, KNN, MLP), and regression (PLS, KNN, MLP).

[3]This chapter constituted the publication article: Sitorus, A., & Lapcharoensuk, R. (2024). Developed an automatic tuning of combination preprocessing and hyperparameter of machine learning and its application to NIR spectral data of coconut milk adulteration. **Food Chemistry**, Submited on 2 December 2023.

The performance strategy proposed in this study effectively addressed and produced satisfactory outcomes in classification and regression challenges problems from coconut milk adulteration. Finally, these results demonstrate that our proposed approach can more deeply discover the best classification and regression model, particularly for the coconut milk adulteration NIR spectroscopy.

Keywords: advance chemometrics; fraud; machine learning; preprocessing; FT-NIR; Micro-NIR.

4.3 Introduction

Near-infrared spectroscopy (NIRs) technology is a rapid and non-destructive method that can predict specific agricultural products' property and their chemical constituent. NIRs is powerful because its measurement does not require complicated sample preparation and provides faster predictive results. NIRs have a shorter wavelength and can penetrate deeper into the sample compared to Mid-infrared spectroscopy (Workman, 2001); therefore, NIRs have been applied in all sorts of applications in food, particularly for adulteration detection. Research papers in the last 10 years report that NIRs can predict qualitatively and quantitatively the adulteration of food and agro-products, including fruit juice (Calle *et al.*, 2022), coriander oil (Kaufmann *et al.*, 2022), bee honey (Bodor *et al.*, 2023; Raypah *et al.*, 2022), soy sauce (Chen *et al.*, 2023), olive oil (Meng *et al.*, 2023), quinoa flour (Wang *et al.*, 2022), chickpea flour (Bala *et al.*, 2022), coconut powder (Pandiselvam *et al.*, 2022), ginger powder (Yu *et al.*, 2022), pepper powder (Wu *et al.*, 2023), and nutmeg (Sitorus *et al.*, 2023). However, spectroscopy data obtained through NIR often overlap, showing less information about the overall structure.

In developing the model of adulteration in food and agro-products, most chemometrics from the research paper above utilize a combination of NIRs with traditional algorithms like partial least squares regression (PLSR). Only a few studies consider using advanced chemometrics from machine learning to analyze spectral data from NIRs and compare it with the PLSR algorithm. On the other hand, the main challenge to getting a robust model from NIR spectral data is the need for various techniques to handle overlapping classes, diverse samples, high sample sizes, imbalanced samples, and nonlinear relationships in data (Cardoso and Poppi, 2021).

Therefore, chemometric experts persistently seek alternative techniques to address this issue by utilizing machine learning that can work more deeply for NIR datasets. The most popular machine learning algorithms employed recently on NIRs datasets of adulteration in food and agro-products are support vector machine (SVM), k-nearest neighbor (KNN), and artificial neural networks (ANN). Cardoso and Poppi (2021) reported using SVM to predict commercial green tea blends with an accuracy between 82.0 and 93.0%. Another research report by Fowler et al. (2021) in lamb loin adulteration using SVM with R^2 maximum 0.65, Dankowska et al. (2022) in dried herbs using KNN with accuracy 86.6%, Valinger et al. (2021) in honey adulteration using ANN with R^2 above 0.8, and Sitorus et al. (2023) in nutmeg adulteration using ANN with R^2 above 0.9. The description of the above research findings underscores that machine learning algorithms have great potential for analyzing NIR data.

In general, generating a robust NIR classification and regression model is started by selecting appropriate data preprocessing. This is because NIRs are often subject to noise, including background shifting, light scattering, varying noises, and other unexpected factors. Currently, many preprocessing methods have been developed to eliminate the noise from NIR spectra, and how-to selection still has pros and cons. Yet, according to Engel et al. (2013), at least selection methods currently available include trial and error, assessment of preprocessed data quality parameters, and expert visual inspection. The trial and error and assessment of preprocessed data quality parameters are time-consuming to execute manually if considering all the existing preprocessing methods and their combination simultaneously. Besides, the visual inspection method by an expert also required a long time to become an expert in this matter. In order to address this gap, some chemometricians are concentrating on developing an automatic preprocessing selection strategy for NIR data through data sciences.

Several studies have investigated the feasibility of an automatic preprocessing selection strategy for full NIR spectra. Start from the research idea of Xu et al. (2008), who developed a strategy for automatic preprocessing of 20 different preprocessing (single to double) for NIR spectra public datasets from wheat kernels and meat using a PLS regressor. Five years later, Xu et al. (2013) reported using an SVM regressor for 62 varieties of combined preprocessing methods from single to double preprocessing

steps. Continues to Gerretzen et al. (2015), who reported using the design of experiments to produce 16 varieties of combined preprocessing methods from single to quadruple preprocessing steps combined with the PLS-DA classifier. Furthermore, Bian et al. (2020) developed a selective preprocessing strategy from 120 varieties of combined preprocessing methods from single to quadruple preprocessing steps combined with the PLS regressor. It can be seen that almost all study focuses on preprocessing combination with traditional chemometrics algorithms, and just one report extends to the SVM algorithm. Up to the research results from Sitorus and Lapcharoensuk (2023) who recently reported automatic selection of preprocessing combined with hyperparameter tuning to improve the performance of the classifier/regressor they used. Unfortunately, they still haven't developed it to the multiple preprocessing stage. On the other hand, tuning of each hyperparameter is required from the beginning to produce a robust classification and regression model from a machine-learning algorithm. Hyperparameter is a parameter value used to control the learning process, and it has to be tuned to generate good performance in a machine-learning algorithm (Guido et al., 2022).

To the best of our knowledge, there is no research investigation of ensembling strategy between preprocessing with tuning hyperparameters machine learning algorithms systematically until quadruple-step preprocessing on classification and regression model, especially in predicting adulteration in coconut milk. This sample was selected due to the widespread utilization of coconut milk as a primary ingredient in numerous culinary dishes and snack preparations, especially in Asia. According to CODEX-STAN-240 (2003), the liquid fresh coconut milk can be categorized into light coconut milk, coconut milk, coconut cream, and coconut cream concentrate, based on total solids, non-fat solids, total fat content, and moisture content. Because of this categorization, the potential for adulteration of fresh coconut milk is tremendous in order to obtain economic benefits from irresponsible people. So far, at least studies related to the adulteration of coconut milk by water, coconut water, and flour have been reported by several research papers (Azlin-hashim *et al.*, 2019; Simuang *et al.*, 2004; Sitorus *et al.*, 2021)

Therefore, the main objective of this study was to develop the strategy of ensembling between selecting preprocessing and tuning hyperparameters of machine

learning algorithms on NIR spectra data for predicting adulteration in coconut milk cases. The specific objectives of this study were to identify the best strategy developed for cases of qualitative (classification of type adulteration) and quantitative (regression of level adulteration) from two types of NIR instruments, including from FT-NIR and from Micro-NIR.

4.4 Materials and Methods

4.4.1 Sample Preparation

Fresh coconut milk (FCM) refers to the liquid extracted directly from the matured endosperm (kernel) of mature coconut fruit, and it is devoid of additional diluents like water or other ingredients. Coconut water in this study refers to the clear, slightly sweet liquid naturally found inside a mature coconut fruit. Coconut water is distinct from coconut milk, as it is obtained from the inner cavity of the young coconut. FCM and coconut water were obtained from local markets in Lad Krabang, Thailand. FCM was collected without additional water or coconut water. All samples were subsequently preserved in glass bottles at room temperature. Before the experiment, both fresh coconut milk and coconut water samples were filtered through 100-mesh cloth filters. Liquid adulterant materials (destilated water and coconut water) with coconut milk were intentionally-adulterated following percentage levels including 0 (pure coconut milk), 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 30, 40, and 50% (w/w). Henceforth in this article, adulterated fresh coconut milk with distilled water and adulterated fresh coconut milk with coconut water are referred to as ADW and ACW, respectively.

Next, samples were stirred in a glass beaker at a speed of 200 rpm for 1 minute and subsequently allowed to equilibrate to a temperature of 25°C before the scanning. For each level of adulteration, 10 samples were prepared, and thus a total of 150 samples were scanned per type of adulteration material. With two types of liquid adulterant materials, the total data analyzed in this study is 310 NIR spectral. Manley (2014) suggests the total NIR spectra samples for developing classification and regression models must at least be more than 100 spectra to get a reliable model, and this study has exceeded that.

4.4.2 NIR Spectral Data Acquisition

The full-wave NIR spectral was measured with benchtop FT-NIR spectrometer (Bruker Ltd., Germany) on 12,500–4000 cm^{-1} (800–2500 nm). Each sample of coconut milk was taken from a glass beaker using a micropipette of about 1 mL. After that, a aluminum reflector was also put in a test glass vial and placed on the FT-NIR spectrometer. One average spectrum was obtained by an average of 32 scans at resolution of 8 cm^{-1}. Secondly, another NIR spectrum was measured with a portable Micro-NIR Pro-1700-ES (VIAVI) on 908–1676 nm (11,013–5967 cm^{-1}) with a resolution of 6.2 nm. Before the acquisition, the integration time and scan count used in this study were 3000 ms, 1000, respectively. After that dark current scanning was conducted in the air, and reference scanning was performed on the teflon material as a reference. Scan results were recorded in absorption mode (log 1/R) for each sample.

4.4.3 Chemometric and Statistical Analysis

4.4.3.1 Principal Component Analysis

Principal component analysis (PCA) is the commonly used unsupervised qualitative multivariate data analysis method with reducing the dimensionality for many purposes (Brereton, 2022). By plotting the scores of the PCA, samples having similar spectra signatures tend to aggregate together or lie close to one another. In this study, PCA was performed on the covariance matrix to identify by examining the grouping of the three samples (FCM, ADW, and ACW) according to their spectral variations.

4.4.3.2 Ensembling Strategy for Developing Model

Following the flowchart shown in Figure 4.1, the classification as qualitative analysis and prediction of adulteration level as a quantitative analysis were created and tested individually based on two type the absorbance NIR spectra using the abovementioned procedures. Methods for developing classification models using machine learning classifier, including linear discriminant analysis (LDA), K-nearest neighbors (KNN) and multilayer perceptron (MLP) were employed in this study. In addition, machine learning regressor, including partial least squares (PLS), KNN, and

MLP, were utilized to develop a regression model to predict the level of FCM adulteration from distilled water and coconut water.

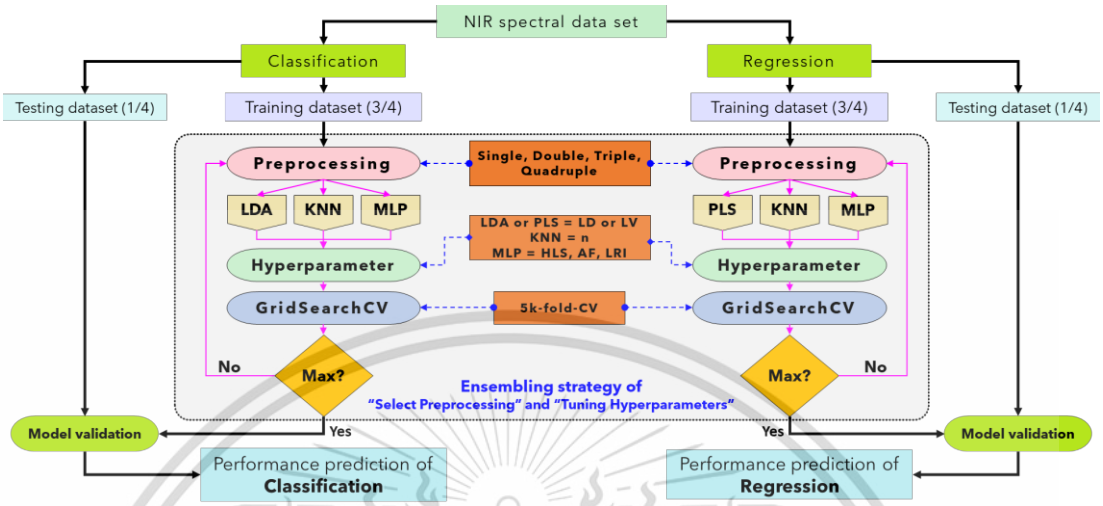


Figure 4.1. Proposed overall methodology for model development.

A comprehensive exploration was conducted to systematically evaluate parameter values to optimize each estimator within the machine learning algorithms. The preprocessing type and hyperparameters of each machine-learning algorithm are optimized together with the values and ranges shown in Table S3-1. Furthermore, as part of the preprocessing selection, a composite of nine type preprocessing methods, including no preprocessing (none), was employed together during the optimization of individual machine learning algorithms via GridSearchCV command. The eight types of preprocessing are baseline second order with 3 points (BSO3), first derivative (FD), second derivative (SD), standard normal variate (SNV), multiplicative signal correction (MSC), mean scaling (MS), Savitzky-Golay filters (SGF) and standard scalers (SS). The combination of 9 types of preprocessing in 4 layers was developed to work from single preprocessing up to quadruple-preprocessing (multiple-preprocessing).

From the total dataset, a random splitting strategy was employed, 3/4 of the experimental samples were used in the training dataset, and 1/4 for the testing dataset or external validation (Sankaran *et al.*, 2011). The statistical parameters of the training and testing dataset are shown in Table S3-2. To internally validate the models, a 5-fold cross-validation (5f-CV) approach was employed on the training dataset. The cross-validation results represent the averaged internal validation

outcomes obtained from 5f-CV. This procedure was carried out utilizing the open-source library from Scikit-learn, Scipy, and NumPy in Python programming language 3.8.8 (Pedregosa *et al.*, 2011).

4.4.3.3 Model Evaluation

Three groups of coconut milk samples will be investigated qualitatively for classification, including FCM, ADW, and ACW. The confusion matrix will be used to evaluate the classification model. In the confusion matrix, each prediction result will be categorized into false positive (FP), false negative (FN), true positive (TP), and false negative (FN). After that, the parameters like precision (Pr), recall (Rc), F1-score (Fs), and accuracy (Ac) will be calculated from the confusion matrix by Equation 4.1 to Equation 4.4.

$$Pr = \frac{TP}{TP+FP} \tag{4.1}$$

$$Rc = \frac{TP}{TP+FN} \tag{4.2}$$

$$Fs = \frac{2 \times Pr \times Rc}{Pr + Rc} \tag{4.3}$$

$$Ac = \frac{TP+TN}{TP+TN+FP+FN} \tag{4.4}$$

Fiveteen impurity levels and one pure fresh coconut milk will be studied quantitatively for predictive levels of its adulteration. The quantitative model was evaluated by coefficient of determination in training and testing (Equation. 4.5), root mean square error (RMSE) in training and testing (Equation 4.6), Bias (Equation 4.7), and the ratio of prediction to deviation (Equation 4.8). The coefficient of determination (R^2) illustrates the ratio of the variance of the response variable obtained and how certain its models can be making predictions by features in the training and testing model. Accuracy is assessed by RMSE in training and testing. The robustness test from the regression model is assessed by the ratio of prediction to deviation (RPD). The ideal model is expected to exhibit a high R^2 value, a low RMSE, a low Bias, and a high RPD (Chu *et al.*, 2022; Conzen, 2006).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(E_i - P_i\right)^2}{\sum_{i=1}^{n}\left(E_i - \bar{E}\right)^2} \tag{4.5}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(E_i - P_i\right)^2}{N}} \tag{4.6}$$

$$\text{Bias} = \frac{\sum_{i=1}^{n}\left(E_i - P_i\right)}{N} \tag{4.7}$$

$$\text{RPD} = \frac{1}{\sqrt{1-R^2}} \tag{4.8}$$

Where E_i is the existing value for point to-*i*, P_i is the prediction value for point to-*i*, N is the number of samples, and $\bar{E}$ is average of existing value.

4.5 Results and Discussions

4.5.1 NIR Spectral Characteristics of Various Samples

Figure 4.2 shows the raw data from FT-NIR and Micro-NIR spectra of FCM, ADW, and ACM. By visual inspection, the spectra are disturbed by the baseline shifting and scatter. Due to the fact that coconut milk primarily consists of water, fat, protein, and carbohydrates with similar chemical structures, the spectral waveform and absorption peak locations appear almost identical, with only minor variations in absorbance. Plotting with degradation color at each level of adulteration shows that the raw NIR spectra both from FT-NIR and Micro-NIR, do not yet have a pattern that can directly predict the adulteration level of FCM in ADW and ACW. Therefore, further processing is needed using chemometrics from this NIR dataset to obtain information on the classification of FCM, ADW, and ACW, as well as the level of adulteration of each type of adulteration.

4.5.2 Spectra Visualization by PCA

PCA was performed on the raw NIR spectra of the samples to determine their classification change characteristics preliminarily. Figure S3-1 displays the 3D score scatter plot obtained through the PCA of the original based data of all the samples and X-loading lines of the first three PCs from 2 spectrometers. Using FT-NIR, the total variance from the first three PCs is 99.30%. Meanwhile, using Micro-NIR, the first three PCs explained 99.37% of the total variance. It can be seen that the initial PCA analysis indicates that the spectral produced by Micro-NIR is better able to explain the total condition of samples with the same number of PCs. However, if you observe the 3D plots in Figure S3-1a and Figure S3-1b, the PCA method can still not truly discriminate between the three sample groups. Only samples of FCM can be identified as being very different from samples that have been filtered with distilled

water or coconut water. While, ADW and ACW samples, are still slightly scattered and overlap each other. This could be because the two main ingredients for alteration are water and the adulteration level range is smaller than 50%. Therefore, PCA could not be used to discriminate samples directly and still requires more advanced chemometric analysis to be able to differentiate them. This challenge is in line with the research results reported by Sitorus and Lapcharoensuk (2023), who reported that pure fresh coconut milk can be differentiated from coconut milk that is adulterated with water in the range of 0 to 100% (with steps of 10%) and instant coconut milk using FT-NIR.

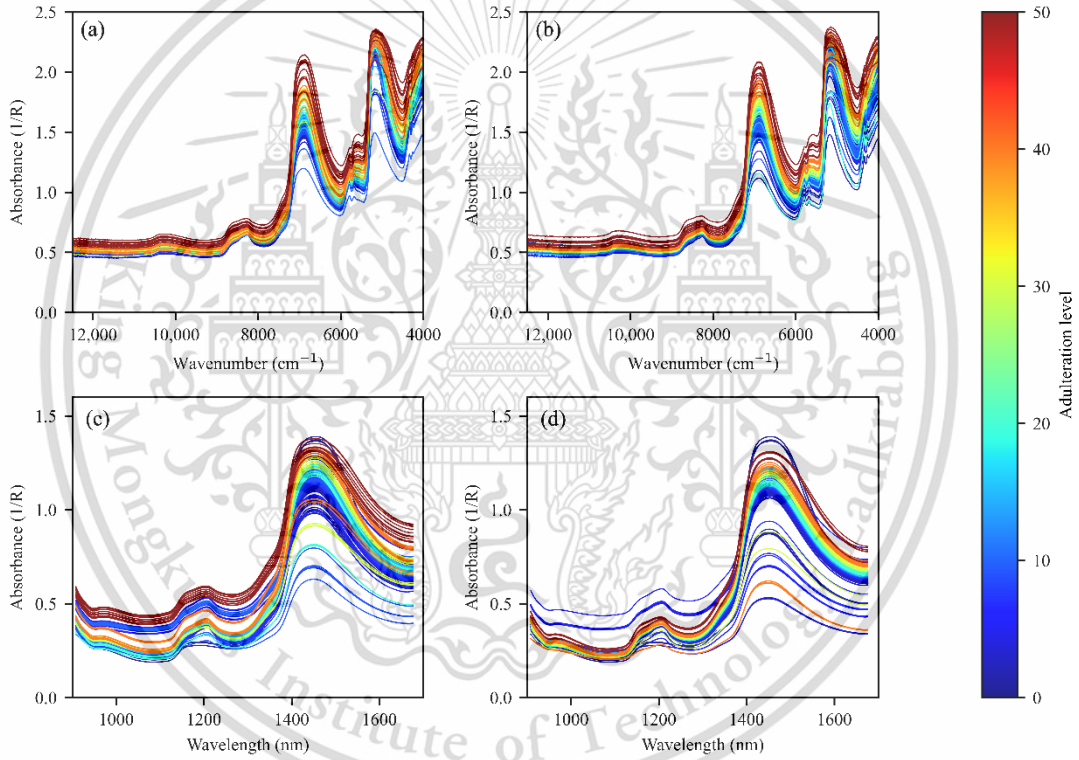


Figure 4.2. NIR spectra of samples from (a) ADW using FT-NIR, (b) ACW using FT-NIR, (c) ADW using Micro-NIR, and (d) ACW using Micro-NIR.

The X-loading lines of the first three PCs for FT-NIR and Micro-NIR are displayed in Figure S3-1c and Figure S3-1d. The positive and negative peaks from X-loading weights show the strong effect of the bond vibration on the classification of samples. The vibration bands from the X-loading of PCA will correspond to vibration bands of some chemical structure X-H. All the corresponding functional groups found in this study are concluded and compared with previous research in Table S3-3 (Conzen,

2006; Osborne *et al.*, 1993; Workman Jr and Weyer, 2007). It can be seen that the originating X-loading PCA peak location comes from OH, CH, and NH because coconut milk is mainly composed of water, fat, protein, and carbohydrates. In general, peaks in X-loading can be found in the previous study, and several peaks have shifted from what they should be. For example, when using Micro-NIR, the 3rd overtone from stretching CH is shifted from 938 nm to 945 nm. Furthermore, 2nd overtone from stretching CH shifted from 1215 nm to 1205 nm for Micro-NIR but not for FT-NIR. The shifting peak on the FT-NIR instrument occurs in the combination of NH stretching and NH_3 bending, where the spectrum is shifted from 4460 cm^{-1} to 4500 cm^{-1}. Additionally, the combination of CH stretching and CH bending was also shifted from 4329 cm^{-1} to 4340 cm^{-1}.

4.5.3 Detection of Adulteration Using FT-NIR

4.5.3.1 Classification of Adulteration Type by FT-NIR

The list of combinations of the three machine learning algorithms (LDA, KNN, MLP) for all combinations of preprocessing and hyperparameters considered in descending order based on 5f-CV accuracy performance is presented in Figure S3-2. For the LDA classifier, there are 39,366 combinations of preprocessing and hyperparameters consisting of 9 types of preprocessing with 4 layers and 6 levels of linear discriminant components (LDs) hyperparameters. Of that number, only 15.37% (6049) combinations had a 5f-CV accuracy value greater than 0. Moreover, the KNN classifier also has a total combination of 39,366 preprocessing and hyperparameters consisting of 9 types of preprocessing with 4 layers and 6 levels of n-neighbors hyperparameters. As many as 90.45% (35,605) combinations using the KNN classifier have a 5f-CV accuracy value greater than 0. Finally, the MLP classifier has a total combination of 209,952 preprocessing and hyperparameters consisting of 9 types of preprocessing with 4 layers and 4, 4, 2 levels/types of hyperparameter hidden layer sizes (HLS), activation function (AF), and learning rate initial (LRI), respectively. The MLP classifier shows that the combination with a 5f-CV accuracy value greater than 0 is 190,133 (90.56%). This total combination is more massive in considering the amount of preprocessing, which simultaneously optimizes the hyperparameters of the machine learning algorithm compared to that carried out in previous studies

(Bian *et al.*, 2020; Xu *et al.*, 2008). This is because this study considers 4 preprocessing layers as well as multiple preprocessing methods before optimizing the hyperparameters of each machine learning algorithm.

It can be seen that the MLP classifier is superior to other machine learning classifiers in testing by the FT-NIR spectra dataset for the classification problem of type adulteration of coconut milk. If we put the order from lowest to highest, its prediction accuracy performance is as follows: KNN < LDA < MLP. The best preprocessing uses the MLP classifier with a combination of MS, MSC, and BSO3 (Figure S3-3d) followed by HLS of 5, AF of identity, and LRI of 0.01 with accuracy 5f-CV is 0.866±0.044. It is clear that although the system developed provides quadruple-preprocessing at once, the strategy proposed in this study is very adaptive to the discovery of the best preprocessing combination. Each preprocessing method has the same opportunity to show its performance when tested using NIR data. Afterward, according to Engel *et al.* (2013), the MS and MSC preprocessing types are preprocessing, which functions to carry out scatter correction, and BSO3 functions to carry out baseline correction of NIR spectra. Next, the best preprocessing uses the LDA classifier with a combination of BSO3, SNV, BSO3, FD (Figure S3-3b) followed by LD component of one with accuracy 5f-CV is 0.763±0.039. Lastly is the KNN classifier with a combination of MSC, MSC, MS, BSO3 (Figure S3-3c) followed by one n-neighbors with an accuracy 5f-CV of 0.75±0.03.

The classification model was validated with the test dataset after finding the best combination of preprocessing and hyperparameters based on 5f-CV accuracy. All the classifier used in this study achieved an overall accuracy rate on validation between 0.69 to 0.92 to classify coconut milk adulteration types (ADW, ACW) with FCM (Table 4.1) based on FT-NIR (12,500–4000 cm^{-1}). The KNN classifier is superior to others at the training stage using FT-NIR. However, at the validation stage, KNN and LDA classifiers were overfitting. Only the MLP classifier can balance its performance with accuracy in validation of 0.92. This is achieved using an architecture of 1 HLZ with 5 neurons followed by an identity-type of AF. An identity of AF is operation activation that is useful to implement linear bottleneck with $f(x)=x$ (Ali *et al.*, 2023). If we look more deeply at the precision, recall, and F1-score parameters, FT-NIR with a combination of preprocessing and hyperparameters can classify FCM and adulterated

coconut milk (ADW, ACW) perfectly. This reconfirms the results of previous research by Sitorus and Lapcharoensuk (2023), which successfully classified 3 groups of coconut milk: fresh coconut milk, instant coconut milk, and adulterated fresh coconut milk with perfect using machine learning algorithms. However, in this study, it was a little challenging to classify between adulterated coconut milk and distilled water or coconut water and pure fresh coconut milk. The challenge arises from the similarity in NIR structure spectra when adulterating coconut milk with water-based adulterants (distilled water and coconut water).

Table 4.1. Performance comparison among the classification models using FT-NIR.

Classifier	The best preprocessing	Hyper-parameter	Spliting	Sample number	PR	RC	Fs	AC
LDA	BSO3+SNV+ BSO3+FD	LD=1	Training	ADW (111)	0.94	0.95	0.94	0.94
				ACW (113)	0.95	0.94	0.94	
				FCM (8)	1.00	1.00	1.00	
			Testing	ADW (39)	0.70	0.79	0.75	0.73
				ACW (37)	0.75	0.65	0.70	
				FCM (2)	1.00	1.00	1.00	
KNN	MSC+MSC+ MS+BSO3	n=1	Training	ADW (111)	1.00	1.00	1.00	1.00
				ACW (113)	1.00	1.00	1.00	
				FCM (8)	1.00	1.00	1.00	
			Testing	ADW (39)	0.68	0.72	0.70	0.69
				ACW (37)	0.69	0.65	0.67	
				FCM (2)	1.00	1.00	1.00	
MLP	MS+MSC +BSO3	HLS=(5) AF=Identity LRI=0.01	Training	ADW (111)	0.87	0.93	0.90	0.90
				ACW (113)	0.92	0.87	0.89	
				FCM (8)	1.00	1.00	1.00	
			Testing	ADW (39)	0.95	0.90	0.92	0.92
				ACW (37)	0.90	0.95	0.92	
				FCM (2)	1.00	1.00	1.00	

4.5.3.2 Prediction of Adulteration Level by FT-NIR

A list of all combinations of preprocessing and hyperparameters using the PLS regressor from the FT-NIR dataset for determining the level of adulteration from ADW and ACW in FCM arranged based on descending of R^2 5f-CV performance is presented in Figure S3-4. By using a full-combination design, a total of 39,366 combinations of preprocessing and hyperparameters were tested to find their best variety. It comes from 9 types of preprocessing on 4 layers and 6 levels of latent variables (LVs) hyperparameters. Only 30,086 of the total combinations, which is 76.43%, showed a combinations have a R^2 on a 5f-CV value greater than 0 for the ADW of the FT-NIR dataset. Meanwhile, for the ACW of the FT-NIR dataset, there are 73.28% (28,847) combinations with a R^2 on a 5f-CV value greater than 0. The combined totals in this study have been much larger than reported by Bian *et al.* (2020), which uses a similar algorithm for 120 types of preprocessing in completing the public NIR spectra dataset from the corn dataset, blood dataset, and edible blend oil dataset. This is because they only consider selective multiple preprocessing combinations. In contrast, this study assesses 4 preprocessing layers simultaneously as multiple-preprocessing

The best preprocessing for detection level ADW in FCM based on FT-NIR using the PLS regressor is with a combination of MS, SNV, MS, BSO3 (Figure S3-5b). Hyperparameter LV is 11. With this combination, it is found that the model performance via R^2 5f-CV is 0.965±0.0172. Meanwhile, to detect the ACW level in FCM based on FT-NIR using the PLS regression, it is possible to combine preprocessing SGF, SGF, SGF, SGF (Figure S3-5b) with hyperparameter LV is 9. With this combination, the model performance via R^2 5f-CV is about 0.965±0.018. The application strategy proposed in this study shows that the best solution for regression cases (both ADW and ACW) using FT-NIR data with PLS as a regressor is to use quadruple-preprocessing. According to Engel *et al.* (2013), the most suitable preprocessing group for the ADW of the FT-NIR dataset is the preprocessing group of scatter correction methods (MS, SNV) and baseline correction methods (BSO3). Meanwhile, for the ACW of the FT-NIR dataset, according to Xu *et al.* (2008), the preprocessing group removes part of the random noise present in the signal and

enhances the signal noise ratio, in this case, represented by SGF, which is the best preprocessing method.

For regressors using KNN, a list of preprocessing and hyperparameter combinations based on the FT-NIR dataset for determining the level of adulteration from ADW and ACW in FCM arranged based on descending R^2 5f-CV performance is presented in Figure S3-6. There are 39,366 combinations of 9 types of preprocessing on 4 layers and 6 levels of n-neighbors hyperparameters. In the case of the ADW of the FT-NIR dataset, as many as 73.33% (28,867) combinations have R^2 on 5f-CV values greater than 0, where 6 of them have equally good performance of 0.931±0.037 with 7 of n-neighbors. The six combinations are combination-1 (MSC, None, MS, BSO3), combination-2 (MSC, MS, BSO3, None), combination-3 (MSC, MS, BSO3, BSO3), combination-4 (MSC, MS, MS, BSO3), combination-5 (None, MSC, MS, BSO3), and combination-6 (MSC, MS, None, BSO3). Meanwhile, for the ACW of the FT-NIR dataset, there are 72.64% (28,596) combinations that have an R^2 on a 5f-CV value greater than 0, where 2 of these combinations have equally good performance of 0.926±0.061 with 3 of n-neighbors. The two combinations are combination-1 (BSO3, MS, MSC, SS) and combination-2 (BSO3, MS, MSC, SNV). This is a signal that multi-preprocessing can also provide more than one best combination option depending on the hyperparameters of the regressor. A similar incident was also reported by Sitorus and Lapcharoensuk (2023) in a quantitative case using single-preprocessing for SVM regressor, where they found at least 12 combinations of preprocessing with SVM hyperparameters to predict the level of fresh coconut milk adulteration by distilled water in the range 0–90%. For further analysis using the KNN regressor, combination 5 (None, MSC, MS, BSO3) was selected for the case of the ADW of the FT-NIR dataset (Figure S3-5c), and combination-1 (BSO3, MS, MSC, SS) for the case of the ACW of the FT-NIR dataset (Figure S3-7c).

The application of the MLP regressor in this study produces a combination of preprocessing and hyperparameter based on a FT-NIR dataset for determining the level of adulteration ADW and ACW, which is arranged based on the descending R^2 of 5f-CV performance presented in Figure S3-8. From that, there are 157,464 combinations originating from 9 types of preprocessing on 4 layers and 4, 3, and 2 levels/types of HLS, AF, and LRI, respectively. Of that total, only 52.83% (83,183) and

52.58% (82,787) combinations have a R^2 on a 5f-CV value greater than 0 for the ADW and ACW of the FT-NIR datasets, respectively. The best preprocessing for the ADW of the FT-NIR dataset is represented by the preprocessing combination of SNV, SS with HLS, AF, and LRI are (3, 5, 7), identity, and 0.01, respectively. Meanwhile, for the ACW of the FT-NIR dataset, the preprocessing combination of MS, SGF, and BSO3 with HLS of (5, 7), AF of identity, and LRI of 0.1 generates the best regression model. The R^2 on 5f-CV for the ADW and ACW are 0.953±0.037 and 0.956±0.018, respectively. For regressors using the MLP, it can be seen that double-preprocessing for ADW and triple-preprocessing for ACW of the FT-NIR datasets work best compared to others. This is the advantage of the strategy developed in this study, where no matter how many multi-preprocessing and hyperparameters are considered, each combination has the same opportunity to be evaluated. This research gap is filled and improved by this study compared with previous research (Sitorus and Lapcharoensuk, 2023; Torniainen *et al.*, 2020).

A comparison of the best results from each regressor used in this study (PLS, KNN, MLP) for the ADW of the FT-NIR dataset problem is presented in Table S3-4. If we put the order from lowest to highest based on RPD performance is as follows: MLP < KNN < PLS. It can be seen that in this study, there is no superior preprocessing method combination for developing regression models using NIR spectral data. In short, combinations of different preprocessing types and hyperparameters for developing regression models depend on the dataset characteristics and the machine learning algorithm chosen as the regressor. Unquestionably, this study provides a new perspective on solving this problem by combining preprocessing and hyperparameter optimization.

Scatter plots of the regression model from PLS, KNN, and MLP regressor to predict the level of adulteration ADW in FCM are shown in Figure 4.3. It can be seen that the measured and predicted values of all machine learning are good predictions. The R^2 value of the regression model for all regressors is above 0.90. However, only the PLS regressor can achieve an R^2 of more than 0.99. Next, in the testing stage, the R^2 value for regressors is above 0.90, yet only the PLS regressor can achieve R^2 up to 0.98. From that, the RPD for predicting the level of adulteration ADW in FCM is more than 4.0 for all regressors. RPD is the most meaningful statistic

used to evaluate the performance of the regression model because it can indicate efficient NIR predictions and, at the same time, verify that standard error predictions should be much lower than the SD. Following the category that was reported by Chu *et al.* (2022) using RPD classification of models for predicting grain chemical composition content, the performance of the MLP regressor for this problem is included in the fair class for use as screening application tool (3.1 < RPD < 4.9). However, the performance of the KNN regressor can be used as a quality control application prediction with a good model accuracy (5.0 < RPD < 6.4). Finally, the PLS regression can be used as a process control with a prediction with an accuracy of the model is very good (6.5 < RPD < 8.0).

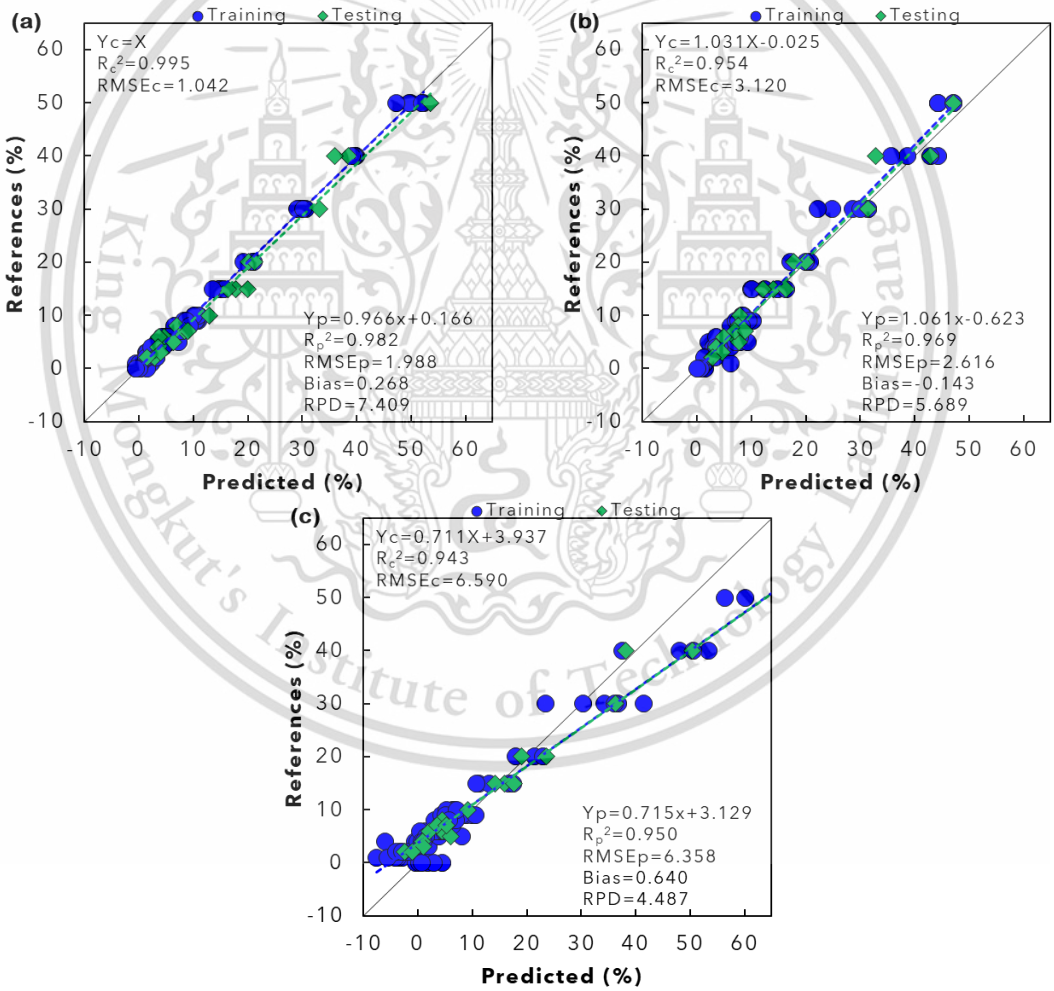


Figure 4.3. The plots of the prediction vs. reference value of ADW in FCM using FT-NIR for for (a) PLS, (b) KNN, and (c) MLP.

A brief of the best regression model for the ACW of the FT-NIR dataset using PLS, KNN, and MLP is presented in Table S3-5. In this case study, we may organize the performance analysis order according to its error as KNN < MLP < PLS. Similar to the previous case, it can be seen that in this study, there is no superior preprocessing method combination for developing regression models using NIR spectral data that is always suitable for all machine learning algorithms. This is because each NIR dataset has its own characteristics, so the machine learning algorithm also requires its own hyperparameter tuning when running. The results of this study on the ADW problem show that the proposed strategy is successful in carrying out a more in-depth search regarding models that have the potential to perform best.

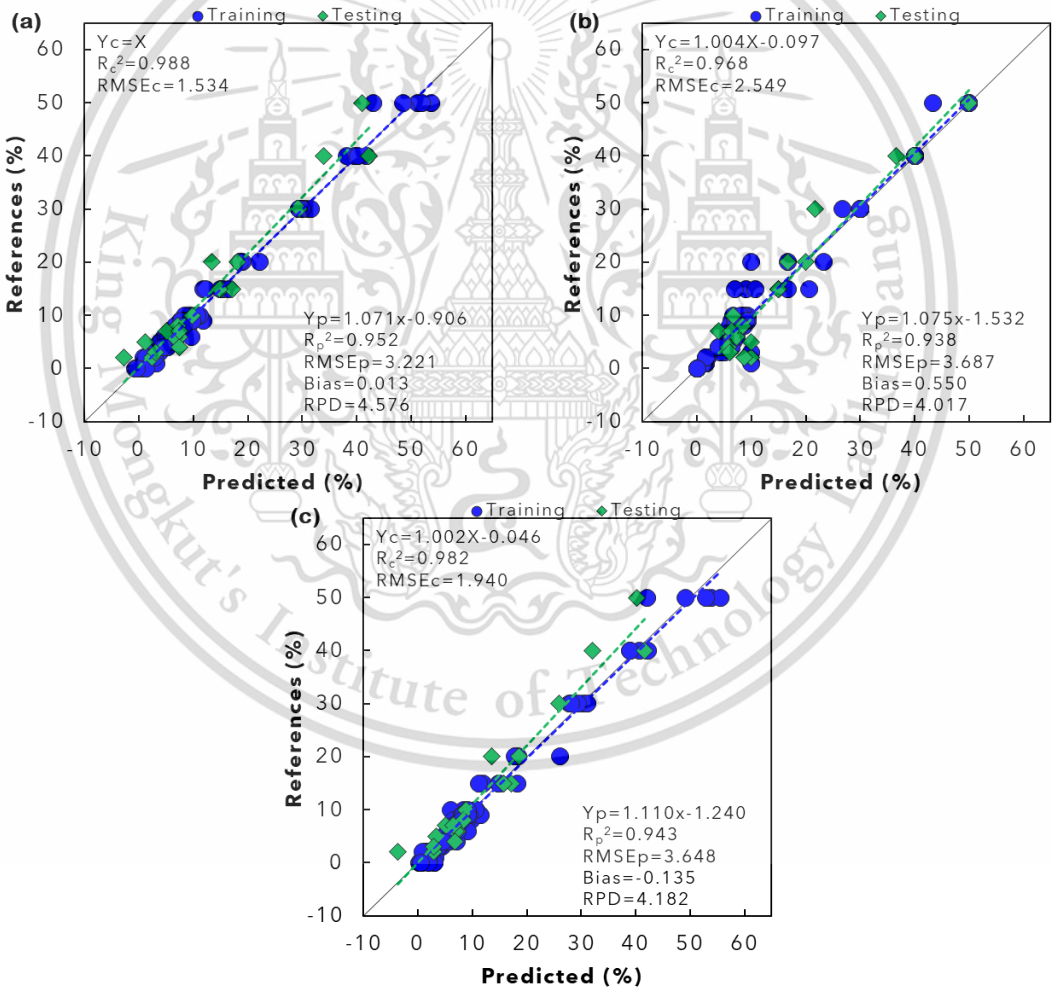


Figure 4.4. The plots of the prediction vs. reference value of ACW in FCM using FT-NIR for (a) PLS, (b) KNN, and (c) MLP.

The scatter plot between references and predicted ACW of the FT-NIR dataset for all machine learning algorithms (PLS, KNN, MLP) is shown in Figure 4.4. The R^2 value on training for all regressors is between 0.968 to 0.988, and the R^2 value of validation is above 0.90. RPD for predicting the level of adulteration ACW for all machine learning models is between 4.017 to 4.576. This shows that the use of FT-NIR to detect the adulteration level of fresh coconut milk from coconut water can only be used as a screening application prediction, with the accuracy of the model being fair (3.1 < RPD < 4.9) (Chu *et al.*, 2022).

4.5.4 Detection of Adulteration Using Micro-NIR

4.5.4.1 Classification of Adulteration Type by Micro-NIR

The list of preprocessing and hyperparameter combinations is arranged descending by 5f-CV accuracy using Micro-NIR for classifying ADW, ACM, and FCM, presented in Figure S3-9. The LDA classifier provides 39,366 combinations of preprocessing and hyperparameters consisting of 9 types of preprocessing with 4 layers and 6 levels of LDs hyperparameters. Of that total, 16.14% (6354) combinations have a 5f-CV accuracy value greater than 0. The KNN classifier has 39,366 preprocessing and hyperparameter combinations consisting of 9 types of preprocessing with 4 layers and 6 levels of n-neighbors hyperparameters. As many as 93.03% (36,624) combinations using the KNN classifier have a 5f-CV accuracy value greater than 0. Finally, the MLP classifier has a total combination of 209,952 preprocessing and hyperparameters consisting of 9 types of preprocessing with 4 layers and 4, 4, 2 levels/types of HLS, AF, and LRI, respectively. The MLP classifier shows that the combination with a 5f-CV accuracy value greater than 0 is 203,328 (96.85%). The combined preprocessing strategy that simultaneously optimizes the hyperparameters of the machine learning algorithm in this study is much greater than that reported by Engel *et al.* (2013), where in their study, they investigated 4914 combinations with the PLS algorithm as the regressor.

In testing this strategy, it can be seen that the KNN and MLP classifiers are superior to LDA, while they (KNN, MLP) are almost equally competitive in internal testing by cross-validation accuracy. In this stage, we can put the order from lowest to highest, its prediction accuracy performance is as follows: LDA < KNN < MLP. The

best preprocessing operates by MLP classifier with a combination of SGF, BSO3, SS, and FD followed by HLS of (5, 7), AF of tanh, and LRI of 0.01 with an accuracy of 5f-CV is 0.97±0.011. In contrast to other classifiers, in this problem, the KNN classifier produces 6 preprocessing combinations that are equally good in 5f-CV accuracy (0.965±0.018). They are combination-1 (SS, FD, SS, SGF), combination-2 (SNV, FD, SNV, SGF), combination-3 (BSO3, MS, FD, SS), combination-4 (BSO3, MS, FD, SNV), combination-5 (SS, FD, SNV, SGF), and combination-6 (SNV, FD, SS, SGF). All Hyperparameter n-neighbors is 1. For further analysis, combination-1 (SS, FD, SS, SGF) was selected using the KNN classifier. Lastly is the LDA classifier with a combination preprocessing of SNV, SGF, MS, and SGF followed by linear discriminant component of 1 with accuracy 5f-CV is 0.923±0.044. The preprocessing of spectra from Micro-NIR is presented in supplementary material Figure S3-10.

All the classification algorithms used in this study achieved an overall accuracy rate on validation between 0.87 to 0.97 (Table 4.2) to classify coconut milk adulteration types (ADW, ACW) that scanned by Micro-NIR (908–1676 nm or 11,013–5967 cm^{-1}). At the training stage, it was seen that the LDA and KNN classifiers showed perfect performance in discriminating coconut milk adulteration types (ADW, ACW) with FCM. Meanwhile, the MLP classifier provides a maximum accuracy performance of 0.98. However, in contrast to the validation stage using the testing dataset, it was found that the LDA classifier provided classification predictions that were much smaller than the training stage (overfitting). Even so, the KNN classifier can still predict better than the LDA and MLP classifiers with an accuracy of 0.97. The KNN classifier achieves this best accuracy using only one n-neighbor component as its hyperparameter. On the other hand, the accuracy performance of the MLP classifier during validation is not much different between FT-NIR and Micro-NIR.

Table 4.2. Performance comparison among the classification models using Micro-NIR.

Classifier	The best preprocessing	Hyper-parameter	Spliting	Sample number	PR	RC	Fs	AC
LDA	SNV+SGF+ MS+SGF	LD=1	Training	ADW (111)	1.00	1.00	1.00	1.00
				ACW (113)	1.00	1.00	1.00	
				FCM (8)	1.00	1.00	1.00	

			Testing	ADW (39)	0.85	0.90	0.88	0.87
				ACW (37)	0.89	0.84	0.86	
				FCM (2)	1.00	1.00	1.00	
KNN	SS+FD+ SS+SGF	n=1	Training	ADW (111)	1.00	1.00	1.00	1.00
				ACW (113)	1.00	1.00	1.00	
				FCM (8)	1.00	1.00	1.00	
			Testing	ADW (39)	0.95	1.00	0.97	0.97
				ACW (37)	1.00	0.95	0.97	
				FCM (2)	1.00	1.00	1.00	
MLP	SGF+BSO3+ SS+FD	HLZs=(5, 7), AF=tanh, LRI=0.01	Training	ADW (111)	0.98	0.98	0.98	0.98
				ACW (113)	0.98	0.98	0.98	
				FCM (8)	1.00	1.00	1.00	
			Testing	ADW (39)	0.88	0.95	0.91	0.91
				ACW (37)	0.94	0.86	0.90	
				FCM (2)	1.00	1.00	1.00	

Similar to the problem using FT-NIR, the best combination of preprocessing and hyperparameters can classify FCM and adulterated coconut milk (ADW, ACW) utilizing Micro-NIR perfectly (if looked at in more depth using the precision, recall, and F1-score parameters). However, the performance of the Micro-NIR spectra for the classification model is better than using FT-NIR. This is presumed because coconut milk samples and adulterant material are in a liquid phase, so lower wavelength NIR energy can give the possibility to sense deep beneath the sample surface compared to higher wavelength NIR (Beć *et al.*, 2021) and capture more information about molecular vibrational excitations that can differentiate them. The best classification model for FT-NIR utilizes the MLP classifier with an accuracy of 0.92, while using Micro-NIR, the optimal model, employs the KNN classifier, achieving an accuracy of 0.97.

4.5.4.2 Prediction of Adulteration level by Micro-NIR

A total of 39,366 combinations of preprocessing and hyperparameter optimization from the PLS regressor were tested to find their best combination in R^2 5f-CV for determination of the level of adulteration from ADW and ACW in FCM employing Micro-NIR are presented in Figure S3-11. The combinations present 9 types of preprocessing on 4 layers and 6 levels of LV hyperparameters. Of that total,

only 88.00% (34,643) combinations have an R^2 on a 5f-CV value greater than 0 for the ADW dataset acquired by Micro-NIR. Meanwhile, for the ACW dataset, there are 82.92% (32,643) combinations that have an R^2 on a 5f-CV value greater than 0. The multiple preprocessing method combined with hyperparameter optimization in this study is more massive than that reported by Xu *et al.* (2008), who used a similar algorithm in completing the public NIR spectra dataset from the wheat kernels dataset and meat dataset. This is because they only compiled multiple preprocessing consisting of only 2 layers from the 8 existing preprocessing methods. Also, they made subjective selections based on their expertise and experience from the number of combinations, so they only left 20 preprocessing methods for reasons of simplicity and feasibility. The research gap is attempted to be resolved by applying data science through an ensembling strategy of preprocessing and hyperparameters for machine learning algorithms, at least in this study, considers PLS, KNN, and MLP regressors.

The best preprocessing for prediction level ADW in FCM scanned by Micro-NIR using the PLS regressor is with a combination of BSO3, BSO3, MS, BSO3 (Figure S3-12b). Hyperparameter LV is 11. With this combination, it is found that the model performance via R^2 5f-CV is 0.976±0.010. Meanwhile, to detect the ACW level in FCM scanned by Micro-NIR using the PLS regression, it combines preprocessing FD, BSO3, SS, FD (Figure S3-13b) with hyperparameter LV of 11. With this combination, model performance via R^2 5f-CV will be obtained about 0.929±0.012. From this, it can be seen that the strategy proposed in this study shows that the best solution for the regression case (both ADW and ACW) using Micro-NIR data with PLS as the regressor is to use quadruple-preprocessing. If observed more deeply, the overall preprocessing appropriate for the PLS regressor works to carry out baseline correction via the BSO3, and FD preprocessing method and perform scaling and transformation via the SS preprocessing method (Bian *et al.*, 2020; Torniainen *et al.*, 2020).

A list of all combinations of preprocessing and hyperparameters using the KNN regressor from the Micro-NIR for determining the level of adulteration from ADW and ACW in FCM, which is arranged based on the descanding of R^2 5f-CV performance is presented in Figure S3-14. Of that, a total of 39,366 combinations of preprocessing and hyperparameters were tested to find their best combination. It comes from 9

types of preprocessing on 4 layers and 6 levels of n-neighbors hyperparameters. In the case of ADW dataset, as many as 92.45% (36,393) combinations have an R^2 on a 5f-CV value greater than 0, where 2 combinations of them have equally good performance of 0.994±0.006 with 1 of n-neighbors. The two combinations are combination-1 (FD, MS, SD, SNV) and combination-2 (FD, MS, SD, SS). Meanwhile, for the ACW dataset, there are 91.30% (35,941) combinations that have an R^2 on a 5f-CV value greater than 0, of which 2 combinations have equally good performance, namely 0.961±0.050 with one of n-neighbors. The two combinations are combination-1 (BSO3, MS, SD, SS) and combination-2 (BSO3, MS, SD, SNV). Similarly to the FT-NIR dataset, using the KNN regressor for the case of Micro-NIR data produces more than one combination of the best preprocessing and hyperparameters. This is because SS and SNV preprocessing, in this case, seems to be able to replace each other to remove or reduce the noise that, according to Engel *et al.* (2013) and Bian *et al.* (2020), SS preprocessing is basically to perform scaling and transformation of spectra data, while SNV preprocessing is tasked with correcting scatter effects from NIR spectra.

For regressors using MLP, a total of 157,464 combinations of preprocessing and hyperparameters using Micro-NIR for determining the level of adulteration from ADW and ACW in FCM, which are arranged based on descanding R^2 5f-CV performance are presented in Figure S3-15. These conditions were found from 9 types of preprocessing on 4 layers and 4, 3, and 2 levels/types of HLZ, AF, and LRI, respectively. Of that total, only 57.90% (93,163) and 57.03% (89,805) of combinations had R^2 on 5f-CV values greater than 0 for the ADW and ACW, respectively. The best preprocessing for ADW dataset is represented by a combination of preprocessing BSO3, SD, SNV, SGF, and HLS, AF, and LRI are (1000), logistic, 0.01, respectively. Meanwhile, the combination of MS, SD, SS, SGF preprocessing with HLS of (100, 100), AF of logistic, LRI of 0.01 are the best for the ACW dataset. The R^2 on 5f-CV for the two Micro-NIR datasets (ADW and ACW) is 0.990±0.003 and 0.973±0.015, respectively. If observed more deeply, the overall preprocessing appropriate for the MLP regressor works to correct the scatter effect of the NIR spectra through the MS and SNV preprocessing methods, carry out baseline correction through the BSO3 and SD preprocessing methods, remove part of random noise present in the signal and

enhance the signal noise ratio via the SGF preprocessing method, and perform scaling and transformation via the SS preprocessing method (Bian *et al.*, 2020; Engel *et al.*, 2013).

A collection of regressor performance (PLS, KNN, MLP) for the ADW problem using Micro-NIR is presented in Table S3-6. Model performance from lowest to highest based on RPD performance is PLS < MLP < KNN. It can be seen that there is no single preprocessing combination method that is superior to all machine learning algorithms (at least of the preprocessing methods considered in this study). This shows that each machine learning algorithm requires its own combination of preprocessing, which must also be aligned with the algorithm's hyperparameters. From that, the preprocessing selection method, as stated by Xu *et al.* (2008), which only considers preprocessing based on the most frequently used and very simple to compute to get simplicity and feasibility, it seems that it has started to be relatively irrelevant, at least until now.

Scatter plots of the regression model to predict the level of adulteration ADW in FCM are shown in Figure 4.5. It can be seen that the predicted values of all regressors are excellent, especially for KNN. The R^2 value on training and testing for all regressors ranges between 0.992–1.00 and 0.988–0.995, respectively. The most superior model uses the KNN regressor with R^2 values on training and testing at 1.00 and 0.995, respectively. Based on RPD criteria that were proposed by Chu *et al.* (2022) using RPD classification of models for predicting grain chemical composition content, the performance of all machine learning algorithms for this problem is included in the excellent class for utilizing any application (RPD > 8.1). If we compare with using FT-NIR in Section 4.4.3 to predict ADW in FCM, Micro-NIR is more recommended.

The results of the performance comparison for the ACW dataset using PLS, KNN, and MLP regressor are presented in Table S3-7. Based on their performance, we can sort the models as PLS < MLP < KNN. It can be seen that in this study, there is no one combination of preprocessing methods that is superior to all regressors. Each NIR dataset has its own characteristics so that it can be developed using machine learning algorithms. Regardless of whether it is done manually with all its limitations as reported in previous research (Bian *et al.*, 2020; Gerretzen *et al.*, 2015; Xu *et al.*,

2008), or in this study, it has been conducted automatically via developing code programming in Python languages.

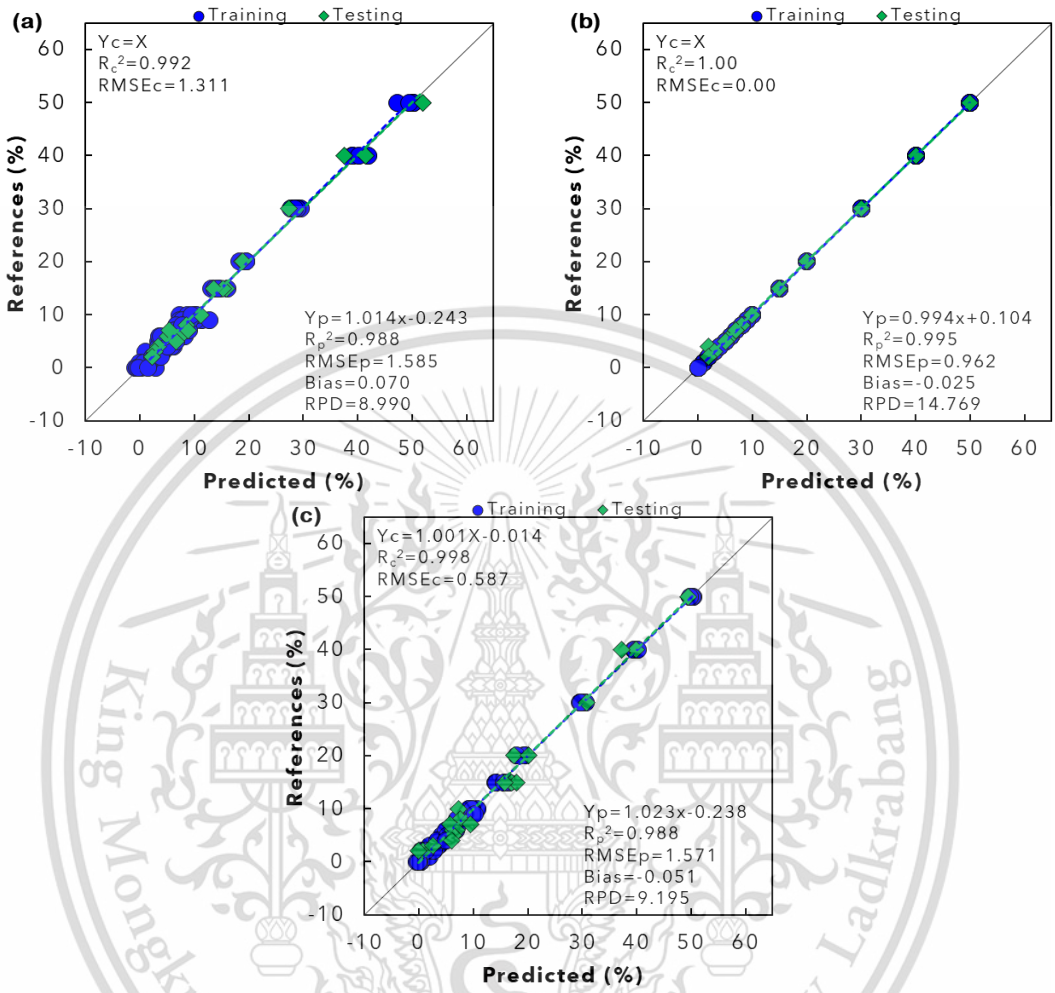


Figure 4.5. The plots of the prediction vs. reference value of ADW in FCM using Micro-NIR for (a) PLS, (b) KNN, (c) MLP.

Scatter plots of all regressors in this study (PLS, KNN, MLP) to predict the level of adulteration ACW in FCM using Micro-NIR are shown in Figure 4.6. Similar to the ADW case, all regressors in ACW dataset have an excellent performance in training, especially for KNN regressors. The R^2 value on training and testing for all regressors ranges between 0.974–1.00 and 0.896–0.996, respectively. Once again, it appears that using Micro-NIR is more appropriate to use the KNN regressor to predict the level of impurity ACW in FCM with R^2 values on training and testing at 1.00 and 0.996, respectively. Following criteria that were reported by Chu *et al.* (2022) using RPD classification of models for predicting grain chemical composition content, the

performance of the PLS regressor for this problem is included in the fair for utilized screening application class (3.1 < RPD < 4.9). However, the performance of the MLP regressor can be used as a process control application prediction, with the accuracy of the regression model being very good (5.0 < RPD < 6.4). Finally, the KNN regressor can be used as any application with a prediction with accuracy of the model is Excellent (RPD > 8.1). Once again, it can be seen that Micro-NIR is more recommended if we compare it with FT-NIR in Section 4.5.3 to predict ACW in FCM.

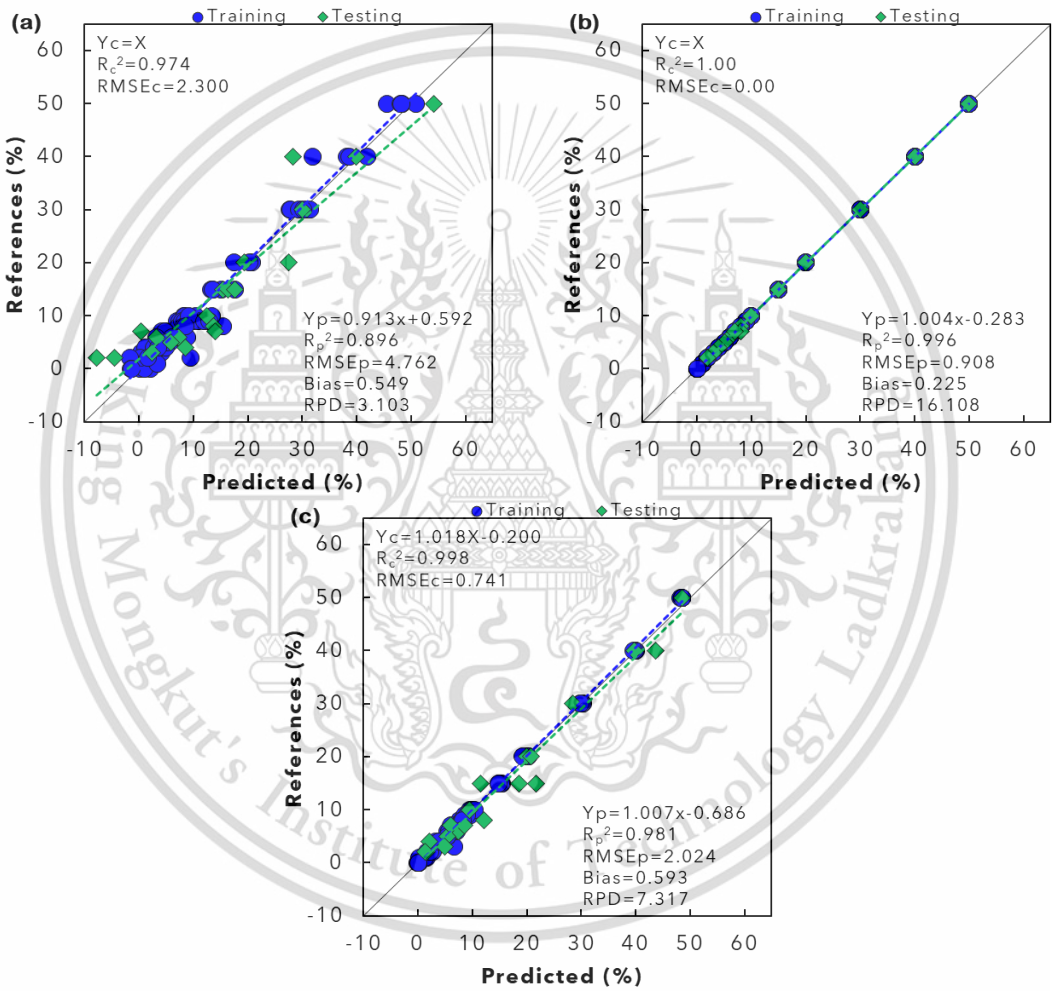


Figure 4.6. The plots of the prediction vs. reference value of ACW in FCM using Micro-NIR for (a) PLS, (b) KNN, (c) MLP.

4.6 Conclusions

New concept to find the appropriate preprocessing and hyperparameter of a machine learning algorithm for coconut milk adulteration NIRs data was developed

and tested. It is set to find preprocessing (single to multiple-preprocessing) in tandem with finding hyperparameters from machine learning algorithms. Systematic ensembling preprocessing and hyperparameter strategy breaks through the shortcomings of preprocessing and hyperparameter tuning classical methods and is more reasonable in treating NIR spectra. It is promising to be applied in NIR spectral preprocessing and tuning hyperparameters directly and jointly to improve model performance.

In the FT-NIR, the MLP classifier excels in classifying ADW, ACW, and FCM with a triple-preprocessing combination of MS, MSC, BSO3, HLS 5, AF identity, and LRI 0.01, achieving 92% accuracy. PLS regressor with 11 LV is best for ADW prediction (RPD of 7.409). For ACW prediction, the PLS regressor with 9 LV and quadruple-preprocessing of SGF achieves an RPD of 4.576. In Micro-NIR, the KNN classifier achieves 98% accuracy for ADW, ACW, and FCM classification with quadruple-preprocessing SS, FD, SS, and SGF. For ADW level prediction in FCM, Micro-NIR with KNN regressor and quadruple-preprocessing FD, MS, SD, SNV, one n-neighbor achieves an RPD of 14.769. For ACW level prediction in FCM, Micro-NIR with KNN regressor and multiple-preprocessing BSO3, MS, SD, SS, one n-neighbor achieves an RPD of 16.108. Finally, this study highlights the value of ensembling preprocessing and hyperparameter strategies in NIR modeling, offering significant benefits to the scientific community saving time and resources.

4.7 References

Ali, H., Muthudoss, P., Ramalingam, M., Kanakaraj, L., Paudel, A., & Ramasamy, G. (2023). Machine Learning–Enabled NIR Spectroscopy. Part 2: Workflow for Selecting a Subset of Samples from Publicly Accessible Data. *AAPS PharmSciTech, 24*(1), 34.

Azlin-hashim, s., Siang, q. l., Yusof, F., Zainol, M. K., & Yusof, H. M. (2019). Chemical composition and potential adulterants in coconut milk sold in Kuala Lumpur. *Malaysian Applied Biology, 48*(3), 27-34.

Bala, M., Sethi, S., Sharma, S., Mridula, D., & Kaur, G. (2022). Prediction of maize flour adulteration in chickpea flour (besan) using near infrared spectroscopy. *Journal of Food Science and Technology, 59*(8), 3130-3138.

Beć, K. B., Grabska, J., & Huck, C. W. (2021). Principles and Applications of Miniaturized Near-Infrared (NIR) Spectrometers. *Chemistry – A European Journal, 27*(5), 1514-1532.

Bian, X., Wang, K., Tan, E., Diwu, P., Zhang, F., & Guo, Y. (2020). A selective ensemble preprocessing strategy for near-infrared spectral quantitative analysis of complex samples. *Chemometrics and Intelligent Laboratory Systems, 197*, 103916.

Bodor, Z., Majadi, M., Benedek, C., Zaukuu, J.-L. Z., Veresné Bálint, M., Csajbókné Csobod, É., & Kovacs, Z. (2023). Detection of Low-Level Adulteration of Hungarian Honey Using near Infrared Spectroscopy. *Chemosensors, 11*(2).

Brereton, R. G. (2022). Numerical introduction to principal components analysis. *Journal of Chemometrics, 36*(8), e3405.

Calle, J. L. P., Barea-Sepúlveda, M., Ruiz-Rodríguez, A., Álvarez, J. Á., Ferreiro-González, M., & Palma, M. (2022). Rapid Detection and Quantification of Adulterants in Fruit Juices Using Machine Learning Tools and Spectroscopy Data. *Sensors, 22*(10).

Cardoso, V. G. K., & Poppi, R. J. (2021). Non-invasive identification of commercial green tea blends using NIR spectroscopy and support vector machine. *Microchemical Journal, 164*, 106052.

Chen, X., Yuan, L., Huang, Y., Chen, J., & Pan, T. (2023). Miniaturized wavelength model optimization for visible–near-infrared spectroscopic discriminant analysis of soy sauce adulteration identification. *Journal of Food Measurement and Characterization*.

Chu, X., Huang, Y., Yun, Y.-H., & Bian, X. (2022). *Chemometric methods in analytical spectroscopy technology*: Springer.

CODEX-STAN-240. (2003). Standard for Aqueous Coconut Products-Coconut Milk and Coconut Cream.: FAO/WHO Food Standards Programme.

Conzen, J. (2006). *Multivariate Calibration: A practical guide for developing methods in the quantitative analytical chemistry*.

Dankowska, A., Majsnerowicz, A., Kowalewski, W., & Włodarska, K. (2022). The Application of Visible and Near-Infrared Spectroscopy Combined with Chemometrics in Classification of Dried Herbs. *Sustainability, 14*(11).

Engel, J., Gerretzen, J., Szymanska, E., Jansen, J. J., Downey, G., Blanchet, L., & Buydens, L. M. C. (2013). Breaking with trends in pre-processing? *TrAC Trends in Analytical Chemistry, 50*, 96-106.

Fowler, S. M., Wheeler, D., Morris, S., Mortimer, S. I., & Hopkins, D. L. (2021). Partial least squares and machine learning for the prediction of intramuscular fat content of lamb loin. *Meat Science, 177*, 108505.

Gerretzen, J., Szymanska, E., Jansen, J. J., Bart, J., van Manen, H.-J., van den Heuvel, E. R., & Buydens, L. M. C. (2015). Simple and Effective Way for Data Preprocessing Selection Based on Design of Experiments. *Analytical Chemistry, 87*(24), 12096-12103.

Guido, R., Groccia, M. C., & Conforti, D. (2022, 2022//). *Hyper-Parameter Optimization in Support Vector Machine on Unbalanced Datasets Using Genetic Algorithms.* Paper presented at the Optimization in Artificial Intelligence and Data Sciences, Cham.

Kaufmann, K. C., Sampaio, K. A., García-Martín, J. F., & Barbin, D. F. (2022). Identification of coriander oil adulteration using a portable NIR spectrometer. *Food Control, 132*, 108536.

Manley, M. (2014). Near-infrared spectroscopy and hyperspectral imaging: non-destructive analysis of biological materials. *Chemical Society Reviews, 43*(24), 8200-8214.

Meng, X., Yin, C., Yuan, L., Zhang, Y., Ju, Y., Xin, K., Chen, W., Lv, K., & Hu, L. (2023). Rapid detection of adulteration of olive oil with soybean oil combined with chemometrics by Fourier transform infrared, visible-near-infrared and excitation-emission matrix fluorescence spectroscopy: A comparative study. *Food Chemistry, 405*, 134828.

Osborne, B. G., Fearn, T., & Hindle, P. H. (1993). *Practical NIR spectroscopy with applications in food and beverage analysis*: Longman scientific and technical.

Pandiselvam, R., Mahanti, N. K., Manikantan, M. R., Kothakota, A., Chakraborty, S. K., Ramesh, S. V., & Beegum, P. P. S. (2022). Rapid detection of adulteration in desiccated coconut powder: vis-NIR spectroscopy and chemometric approach. *Food Control, 133*, 108588.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research, 12*, 2825-2830.

Raypah, M. E., Zhi, L. J., Loon, L. Z., & Omar, A. F. (2022). Near-infrared spectroscopy with chemometrics for identification and quantification of adulteration in high-quality stingless bee honey. *Chemometrics and Intelligent Laboratory Systems, 224*, 104540.

Sankaran, S., Mishra, A., Maja, J. M., & Ehsani, R. (2011). Visible-near infrared spectroscopy for detection of Huanglongbing in citrus orchards. *Computers and Electronics in Agriculture, 77*(2), 127-134.

Simuang, J., Chiewchan, N., & Tansakul, A. (2004). Effects of fat content and temperature on the apparent viscosity of coconut milk. *Journal of Food Engineering, 64*(2), 193-197.

Sitorus, A., & Lapcharoensuk, R. (2023). A rapid method to predict type and adulteration of coconut milk by near-infrared spectroscopy combined with machine learning and chemometric tools. *Microchemical Journal, 195*, 109461.

Sitorus, A., Muslih, M., Cebro, I. S., & Bulan, R. (2021). Dataset of adulteration with water in coconut milk using FTIR spectroscopy. *Data in Brief, 36*, 107058.

Sitorus, A., Pambudi, S., Boodnon, W., & Lapcharoensuk, R. (2023). Near-Infrared Spectroscopy with Machine Learning for Classifying and Quantifying Nutmeg Adulteration. *Analytical Letters*, 1-22.

Torniainen, J., Afara, I. O., Prakash, M., Sarin, J. K., Stenroth, L., & Töyräs, J. (2020). Open-source python module for automated preprocessing of near infrared spectroscopic data. *Analytica Chimica Acta, 1108*, 1-9.

Valinger, D., Longin, L., Grbeš, F., Benković, M., Jurina, T., Gajdoš Kljusurić, J., & Jurinjak Tušek, A. (2021). Detection of honey adulteration – The potential of UV-VIS and NIR spectroscopy coupled with multivariate analysis. *LWT, 145*, 111316.

VIAVI. *Data Sheet VIAVI MicroNIR™ 1700 ES*. Retrieved from: https://www.viavisolutions.com/en-us/literature/micronir-1700es-data-sheets-en.pdf

Wang, Z., Wu, Q., & Kamruzzaman, M. (2022). Portable NIR spectroscopy and PLS based variable selection for adulteration detection in quinoa flour. *Food Control, 138*, 108970.

Workman, J. (2001). *Handbook of organic compounds: NIR, IR, Raman and UV-Vis spectra featuring polymers and surfactants*: Academic Press.

Workman Jr, J., & Weyer, L. (2007). *Practical guide to interpretive near-infrared spectroscopy*: CRC press.

Wu, S., Wang, L., Zhou, G., Liu, C., Ji, Z., Li, Z., & Li, W. (2023). Strategies for the content determination of capsaicin and the identification of adulterated pepper powder using a hand-held near-infrared spectrometer. *Food Research International, 163*, 112192.

Xu, L., Yan, S.-M., Cai, C.-B., Yu, X.-P., Jiang, J.-H., Wu, H.-L., & Yu, R.-Q. (2013). Nonlinear Multivariate Calibration of Shelf Life of Preserved Eggs (Pidan) by Near Infrared Spectroscopy: Stacked Least Squares Support Vector Machine with Ensemble Preprocessing. *Journal of Spectroscopy, 2013*, 797302.

Xu, L., Zhou, Y.-P., Tang, L.-J., Wu, H.-L., Jiang, J.-H., Shen, G.-L., & Yu, R.-Q. (2008). Ensemble preprocessing of near-infrared (NIR) spectra for multivariate calibration. *Analytica Chimica Acta, 616*(2), 138-143.

Yu, D.-x., Guo, S., Zhang, X., Yan, H., Zhang, Z.-y., Chen, X., Chen, J.-y., Jin, S.-j., Yang, J., & Duan, J.-a. (2022). Rapid detection of adulteration in powder of ginger (Zingiber officinale Roscoe) by FT-NIR spectroscopy combined with chemometrics. *Food Chemistry: X, 15*, 100450.

CHAPTER 5 – CASE STUDY 3
EXPLORING DEEP LEARNING TO PREDICT COCONUT MILK ADULTERATION USING FT-NIR AND MICRO-NIR SPECTROSCOPY[4]

5.1 Abstract

Accurately identifying adulterants in agriculture and food products is associated with preventing food safety and commercial fraud activities. However, a rapid, accurate, and robust prediction model for adulteration detection is hard to achieve in practice. Therefore, this study aimed to explore deep-learning algorithms as an approach to accurately identify the level of adulterated coconut milk using two types of NIR spectrophotometer, including benchtop FT-NIR and portable Micro-NIR. Coconut milk adulteration samples came from deliberate adulteration with corn flour and tapioca starch in the 1 to 50% range. A total of four types of deep-learning algorithm architecture that were self-modified to a one-dimensional framework were developed and tested to the NIR dataset, in-cluding simple CNN, S-AlexNET, ResNET, and GoogleNET. The results confirmed the feasibility of deep-learning algorithms for predicting the degree of coconut milk adulteration by corn flour and tapioca starch using NIR spectra with reliable performance (R^2 of 0.886–0.999, RMSE of 0.370–6.108%, and Bias of -0.176–1.481). Furthermore, the ratio of percent deviation (RPD) of all algo-rithms with all types of NIR spectrophotometers indicates an excellent capability for quantitative predictions for any application (RPD > 8.1) except for case predicting tapioca starch, using FT-NIR by ResNET (RPD < 3.0). This study demonstrated the feasibility of using deep-learning algorithms and NIR spectral data as a rapid, accurate, robust, and non-destructive way to evaluate coconut milk adulterants. Last but not least, Micro-NIR is more promising than FT-NIR in predicting coconut milk adulteration from solid adulterants, and it is portable for in situ measurements in the future.

---

[4]This chapter constituted the publication article: Sitorus, A., & Lapcharoensuk, R. (2024). Exploring Deep-learning for Predict Coconut Milk Adultera-tion using FT-NIR and Micro-NIR Spectroscopy. **Sensors**, *24*(7), 2362. https://doi.org/10.3390/s24072362.

Keywords: adulteration; chemometric; coconut milk; deep learning; food; non-destructive

5.2 Introduction

Adulteration in agriculture and food products is an essential safety and control area requiring rapid, accurate, robust, and automated methods for detecting, identifying, and quantifying adulteration, including coconut milk products. Coconut milk is generally extracted from grated coconut meat after pressing or squeezing with or without the addition of water. Coconut milk has been used as a major ingredient in several cuisines, such as curries and desserts (Tansakul and Chaisawang, 2006). There are two common reasons for adulteration in coconut milk products. The first reason is to increase production volume and reduce costs by adding tap water or mature coconut water to coconut milk. The second reason is an attempt to boost the apparent carbohydrate content by adding corn flour.

Accurately identifying adulterants is important for controlling coconut milk product adulteration. The main content of coconut milk (moisture, total fat, carbohydrates, protein, and ash) will be changed when mixed with other materials. As reported by Lakshanasomya *et al.* (2011), laboratory testing can measure the total solids and total fat in coconut milk by drying it in a hot air oven or using a vacuum oven, which takes more than 2 hr to prepare one sample. Although accurate, this method is time-consuming and requires complicated sample pretreatments and well-trained technicians, so it cannot be relied on to carry out rapid monitoring. Near-infrared (NIR) and mid-infrared spectroscopy have gained considerable interest among the approaches to physical properties, particularly for detecting adulteration in many agricultural and food products. Compared with the above methods, NIR and mid-infrared spectroscopy are analytical techniques with the advantages of rapid response in real time, simplicity in testing, and are non-destructive. However, this method requires the development of a calibration model before it can be used to make predictions.

Several efforts to develop calibration models have been created using a chemometrics approach to achieve better performance prediction of coconut milk adulteration based on NIR spectroscopy. For instance, Azlin-Hashim *et al.* (2019)

employed partial least squares (PLS) regression to quantitatively determine the concentration of corn flour in the coconut milk using an FT-IR spectrometer. This study used spectroscopic techniques in mid-infrared zones combined with classical chemometrics. Although advantageous, classical chemometric analysis is frequently criticized for its requirement of expertise and subjectivity in elaborating spectral data, including selecting an excellent preprocessing method based on what worked well on a previous data set and how to highlight important spectral regions (Acquarelli *et al.*, 2017; Nallan Chakravartula *et al.*, 2022). Therefore, Sitorus and Lapcharoensuk (2023) adopted a machine-learning algorithm with automatic preprocessing to predict water in coconut milk using an FT-NIR spectrometer. Al-Awadhi and Deshmukh (2021) utilized linear discriminant analysis (LDA) and K-nearest neighbors (KNN) from machine learning as a classifier to detect water in coconut milk using an FT-IR spectrometer. They succeeded in improving model accuracy but were observed to be complicated structures that were difficult to train and with apparent risks of over-fitting. Moreover, although robust and accurate, these strategies have drawbacks related to data dimensionality and higher entropy apart from efforts and practical feasibility (Acquarelli *et al.*, 2017; Cui and Fearn, 2018; Engel *et al.*, 2013). Furthermore, efforts related to learning representations of the data that identify and highlight the underlying explanatory factors hidden in the data are still challenging in machine-learning applications (Bengio *et al.*, 2013). Consequently, some studies are probing for a shift in the paradigm toward applying deep learning to resolve the issues related to classical and feed-forward neural network approaches.

Deep learning is a branch of machine learning that begins with images as input and learns to identify patterns within their spatial dimensions. Deep learning consists of multiple processing layers to automatically learn complex representations from data without introducing hand-coded rules or human domain knowledge. Among deep-learning algorithms, convolutional neural networks (CNNs) are presently one of the most trending models since they do not require manual feature extraction and have several network architecture types. CNNs are constructed with a series of convolutional layers that act as feature extractors, followed by fully connected layers at the end of the network that serve as predictors. For processing NIR spectral data, it was also recently seen to be useful for one-dimensional (1D) spectroscopy

data, as well as for regression tasks wherein this supervised approach could perform both feature extraction and learning related to features of interest (Cui and Fearn, 2018). Presently, CNN techniques are developing rapidly so many network architecture variants are found for various analysis purposes, such as AlexNet, ResNET, GoogLeNet, etc. (Gron, 2019). Furthermore, in the case of chemometrics data, the use of CNN can enable training on smaller weights, thereby lowering data complexity as opposed to fully connected or feed-forward neural networks. Some of the advantages of CNN are a reduction in neuron interdependence, adaptability to datasets beyond the training, and reduced risk of over-fitting, which is a common criticism in feed-forward networks. Moreover, some researchers (Cui and Fearn, 2018; Nallan Chakravartula *et al.*, 2022) note that CNN can eventually simplify preprocessing and model development, thereby reducing the complexity of model development and improving accurate and robust model predictions. Recent papers on the utilization of CNN for NIR spectroscopy, especially in agricultural and food products, have reported adulteration in coffee products (Nallan Chakravartula *et al.*, 2022), adulteration in infant formula products (Liu *et al.*, 2021), adulteration in dairy products (Said *et al.*, 2022), and adulteration in minced beef products (Weng *et al.*, 2020). This shows the increasing number of studies using the CNN algorithm in NIR-based adulteration detection for agricultural and food products.

To the best of our knowledge, even though coconut milk adulteration was investigated with another adulterant material and another type of spectroscopy (Azlin-Hashim *et al.*, 2019; Sitorus and Lapcharoensuk, 2023), no study explored deep learning as advanced computational algorithms for quantifying adulterants by two types of NIR spectroscopy, including benchtop FT-NIR and portable Micro-NIR and by two types of solid adulterants, including corn flour and tapioca starch. Therefore, the objective of this study was to bridge the gap between advanced perceptual sensors from NIR spectroscopy and data science by developing and testing the performance of four types of regressor architecture CNN of deep-learning to detect coconut milk adulteration from corn flour and tapioca starch.

5.3 Materials and Methods

5.3.1 Sample Collection

This study comprised two parts. The first part was to detect the level of coconut milk adulteration by corn flour and tapioca starch using a benchtop FT-NIR spectrometer. The second part was to detect coconut milk adulteration levels by corn flour and tapioca starch utilizing a portable Micro-NIR spectrometer. For this purpose, the spectra of two potential adulterants (corn flour and tapioca starch) and coconut milk were collected. The adulterants were purchased from a grocery market around Lat Krabang, Thailand. Adulterant samples were stored at room temperature under clean and dry conditions. Coconut milk is the liquid extracted from mature coconut fruit's endosperm, with no added diluents like water or other materials. Coconut milk was obtained from traditional markets (Lad Krabang, Thailand) on 6 January 2024 and processed on the same day.

The coconut milk samples were adulterated by mixing the identified adulterant in the range of 1–50% (w/w) at different adulteration levels (approximately 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 15%, 20%, 30%, 40%, and 50%). The reasons for selecting adulteration levels in this study were as follows. Selecting different mixing levels can help represent various levels of mixing severity. Covering low mixing levels (1%) to high levels (50%) can provide a more complete view of the range of possible mixing situations that will be encountered in practice. Relevance to practical applications where mixing levels with minimum conditions for agriculture and food products are more likely to occur frequently in practical situations in the field so that they have a more significant impact on this study. Also, selecting various mixing levels can help test the model's calibration ability to detect mixing at various severity levels. This can help identify the extent to which the model is reliable in identifying mixing at different levels.

After the coconut milk samples had been intentionally adulterated to the level of conditions specified above, all samples were subsequently preserved in glass bottles at room temperature after being mixed in a glass beaker at a speed of 200 rpm for 1 min and allowed to equilibrate to ±25°C before scanning. Ten samples were prepared at each level of adulteration and each of adulterant. Therefore, a total of 150 samples were prepared per type of adulterant. The total data analyzed

in this study were higher than in the suggestion by Manley (2014) that developing regression models, which must have at least more than 100 spectra to obtain a reliable model, has exceeded that.

5.3.2 NIR Spectroscopy Data Acquisition

NIR spectra were measured using a benchtop FT-NIR spectrometer (Bruker Ltd., Ettlingen Germany) and a portable Micro-NIR spectrometer (MicroNIR OnSite-W, VIAVI Solutions Inc., Chandler, United States). Spectra in the 12,500–4000 cm^{-1} (800–2500 nm) region were recorded using the benchtop FT-NIR (Figure 5.1a) with an average spectrum of 32 scans into one spectrum at a resolution of 8 cm^{-1}. Secondly, portable Micro-NIR (Figure 5.1b) scanned in the spectra range from 908 to 1676 nm (11,013–5967 cm^{-1}). The spectral resolution was set to 6.2 nm. The integration time and scan count were 10 ms and 100, respectively. FT-NIR and Micro-NIR were used together to obtain more complete and comprehensive information about NIR spectra of coconut milk samples adulterated with two potential adulterants (corn flour and tapioca starch).

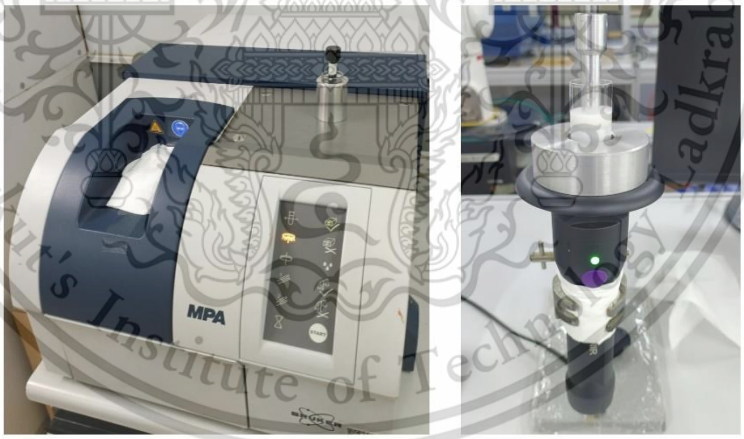


Figure 5.1. Detection conditions for scanning NIR data by (a) FT-NIR and (b) Micro-NIR.

A sample of coconut milk (1 mL) that was intentionally adulterated was taken to test a glass vial with a diameter of 20 mm and a height of 43 mm. After that, an aluminum reflector (with a path length of 0.35 mm) was also put in a test glass vial and placed on top of the benchtop FT-NIR and portable Micro-NIR spectrometer.

Scanning was performed triplicated for each sample, accumulating 3 scanning ×10 samples ×15 level adulteration. All of the measurements were taken at room temperature (±25°C). Scanning was performed in absorption mode (log 1/R).

5.3.3 Data Handling for Modelling

The whole dataset in this study is 1800 NIR spectra consisting of 900 spectra scanned by benchtop FT-NIR and 900 spectra scanned by portable Micro-NIR. Each NIR spectrum acquired from these two instruments consists of 450 spectra from adul-teration coconut milk by corn flour and 450 spectra from adulteration coconut milk by tapioca starch. From the NIR spectra data, each adulteration was split into training and testing subsets. The training data were used to develop a deep-learning model. Then, the models were applied to the test dataset to assess the predictive abilities of the models for predicting the adulteration level of coconut milk by corn flour and tapioca starch. Table 5.1 summarizes adulteration coconut milk data spectra, including NIR spectra data collected from benchtop FT-NIR and portable Micro-NIR. The number of training (70%) and testing (30%) data used in this study was 315:135 (separated using the random splitting method with a random state of 42).

Table 5.1. Summary statistics of data for developing a deep-learning model.

Adulteration material	Instruments	*m*	Training				Testing			
			Min-Max	Mean	SD	*n*	Min-Max	Mean	SD	*n*
Corn flour	FT-NIR	1102	1 – 50	14.00	14.329	315	1 – 50	14.00	14.359	135
	Micro-NIR	125	1 – 50	14.00	14.329	315	1 – 50	14.00	14.359	135
Tapioca starch	FT-NIR	1102	1 – 50	14.00	14.329	315	1 – 50	14.00	14.359	135
	Micro-NIR	125	1 – 50	14.00	14.329	315	1 – 50	14.00	14.359	135

m, number of features; *n*, number of samples; SD, standard deviation

5.3.4 Deep-learning Model Development

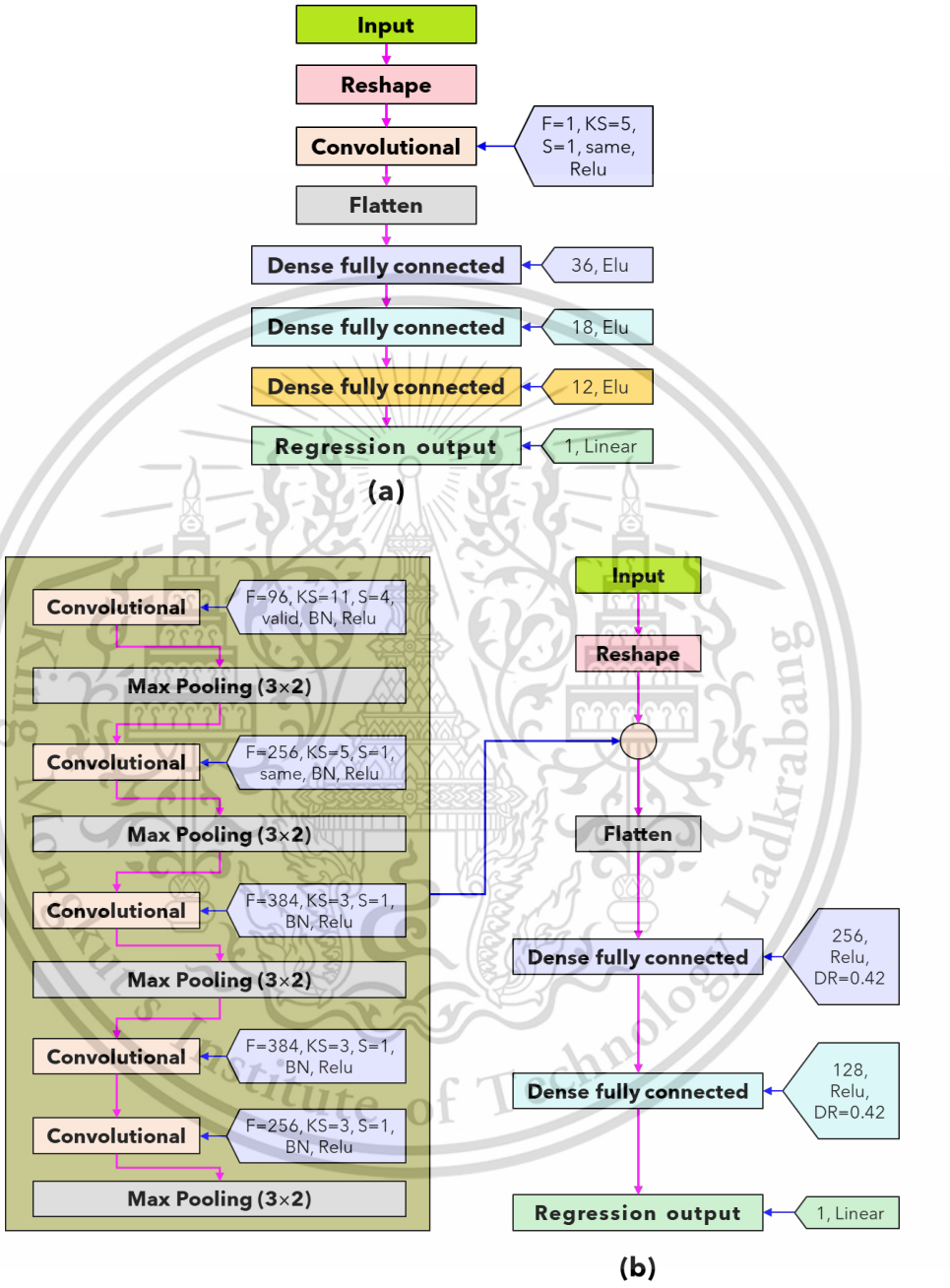
For analyzing the NIR datasets with deep-learning models, a unique structure is required to provide suitable training and improve the feature extraction process. The proposed deep-learning regressor model for predicting coconut milk adulteration in this study uses four types of network architecture that are modified to a one-dimensional framework was developed, including simple convolutional neural network (Simple CNN), A-AlexNet, ResNET, and GoogleNET, which are presented in Figure 5.2. The reason for using simple CNN was based on its effectiveness and

efficiency in processing data, which is particularly suitable for relatively simple datasets (Acquarelli *et al.*, 2017). S-AlexNet was selected as an adaptation of the successful AlexNet architecture, offering a lighter architecture suitable for smaller datasets (Passos and Mishra, 2023). ResNet was chosen for its ability to address the vanishing gradient problem and enable the training of deeper models, making it suitable for complex datasets requiring deep feature representations (Yang *et al.*, 2021). GoogleNet was selected for its innovative architecture, particularly the efficient use of inception modules for feature extraction, making it ideal for complex datasets requiring multi-scale feature representations (Jin *et al.*, 2022). In addition, the original spectral data were preprocessed using standard normal variate (SNV) to obtain input features spectrum to have zero mean and standard deviation of one. The SNV preprocessing effectively reduced specific noise appearing in the spectral data due to the effect of ambient light, the spectrometer used, and the type of lamp (Benmouna *et al.*, 2022).

All types of network architecture model deep-learning training procedures in this study were performed using Adam optimizer. In the simple CNN network architecture (Figure 5.2a), the number of batches, epochs per running, validation split, and learning rate are 128, 1000, 10%, and 5×10^{-3}, respectively. In the S-AlexNET network architecture (Figure 5.2b), the number of batches, epoch per running, validation split, and learning rate are 16, 300, 10%, and 10^{-5}, respectively. In the ResNET network architecture (Figure 5.2c), the number of batches, epochs per running, validation split, and learning rate are 160, 1000, 10%, and 10^{-5}, respectively. In the GoogleNET network architecture (Figure 5.2d), the number of batches, epochs per running, validation split, and learning rate are 160, 1000, 10%, and 10^{-5}, respectively. The total running time is 11 times with early stopping patience of 200 epochs for simple CNN, ResNET, and GoogleNET and 300 epochs for S-AlexNET. All network architectures use random split validation at the training stage with 10% from the training dataset (10%×315=32) with a random state 42. After that, the best deep-learning regression was evaluated with a testing dataset of as much as 135.

In this study, all of the deep-learning algorithms were programmed on the JupyterLab interface using the open source platform Python version 6.5.4, Keras library version 2.13.1 (Gulli and Pal, 2017) with TensorFlow version 2.13.0 backend

(Abadi *et al.*, 2016). The CPU is Intel (R) core (TM) i9-13900H CPU @ 2.60 Ghz, and the graphics card is NVIDIA Geforce RTX 4060.



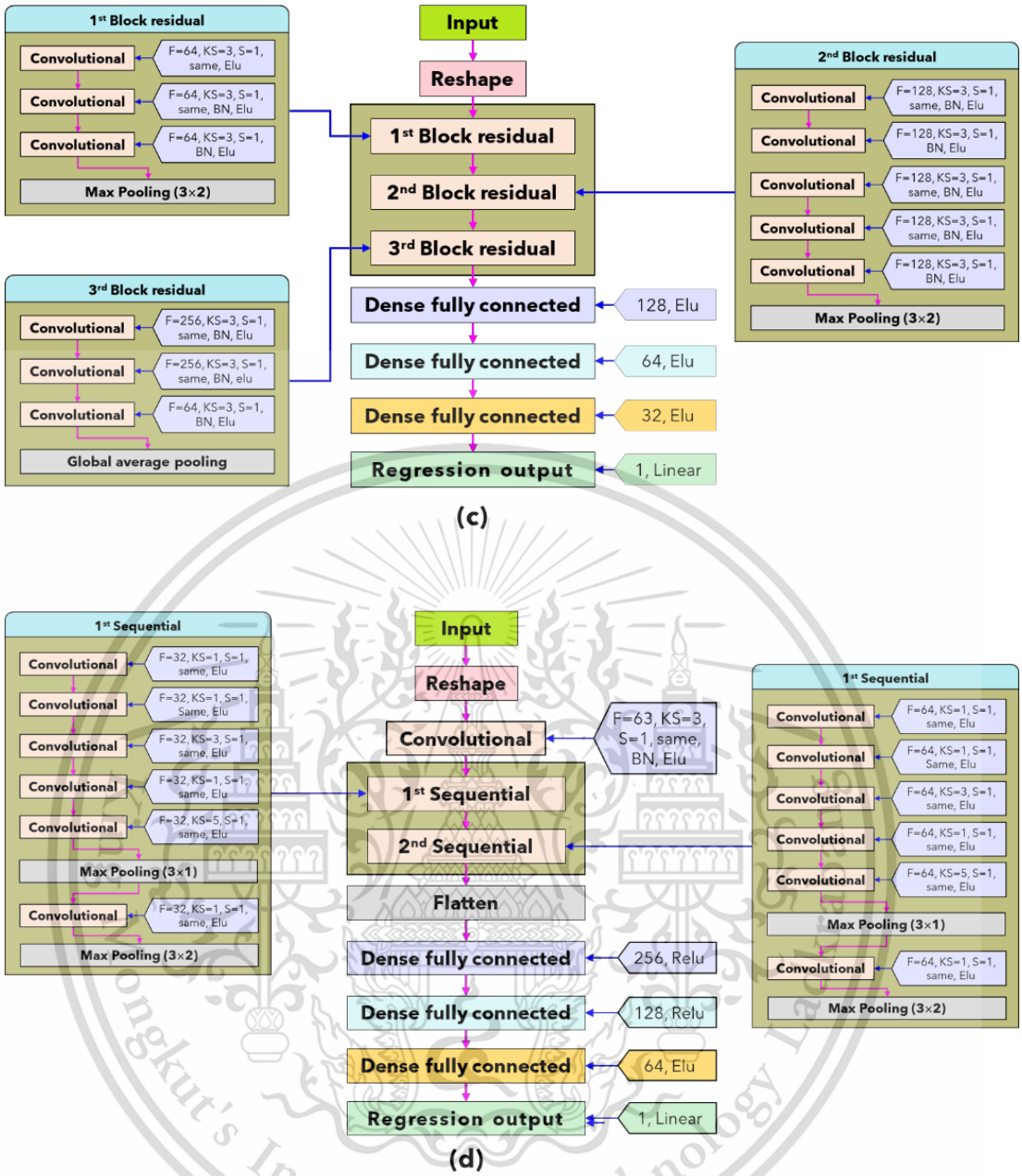


Figure 5.2. Architecture and parameters of the proposed algorithms in this study. (a) Simple CNN regressor; (b) S-AlexNet regressor; (c) ResNET regressor; and (d) GoogleNET regressor.

5.3.5 Performance Model Evaluation

The performance of models was assessed by the coefficient of determination, root-mean-square error, Bias, and the ratio of percent deviation (RPD), which were calculated by Equations (5.1)–(5.4). The higher coefficient of determination values and lower root-mean-square error and Bias indicate an accurate model. In food

adulteration, spectral modeling follows predicting grain chemical composition content; RPD < 3.0 shows a poor and unreliable model; 3.1 < RPD < 4.9 indicates a fair model (just for screening); 5.0 < RPD < 6.4 shows a good model for quality control; 6.5 < RPD < 8.0 indicates very good model for process control, and RPD > 8.1 indicates an excellent capability of the model for quantitative predictions for any application (Chu *et al.*, 2022).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(E_i - P_i\right)^2}{\sum_{i=1}^{n}\left(E_i - \bar{E}\right)^2} \quad (5.1)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(E_i - P_i\right)^2}{N}} \quad (5.2)$$

$$\text{Bias} = \frac{\sum_{i=1}^{n}\left(E_i - P_i\right)}{N} \quad (5.3)$$

$$\text{RPD} = \frac{1}{\sqrt{1 - R^2}} \quad (5.4)$$

Where, R^2 is the coefficient of determination; RMSE is root-mean-square error; RPD is a ratio of percent deviation; E_i is the existing value for point to-*i*; P_i is the prediction value for point to-*i*; N is the number of samples, and $\bar{E}$ is average of existing value.

To interpret the obtained models, this study proposes a method to visualize the regression coefficients of neural networks numerically. This method is modified from Cui and Fearn (2018) and can be used for linear predictors. As a black box that maps the input spectrum, the predictor was treated to a single prediction value using Equations (5.5). The single total weight is calculated using Equation (5.6) (finite difference approximation) from the main Equation (5.7).

$$w = \frac{\sum_{i=1}^{N} w_i}{N} \quad (5.5)$$

$$w_i = \frac{f\left(x_1 + \varepsilon, \ldots x_i + \varepsilon, \ldots, x_n + \varepsilon\right) - f(x_1, \ldots x_i, \ldots, x_n)}{\varepsilon} \quad (5.6)$$

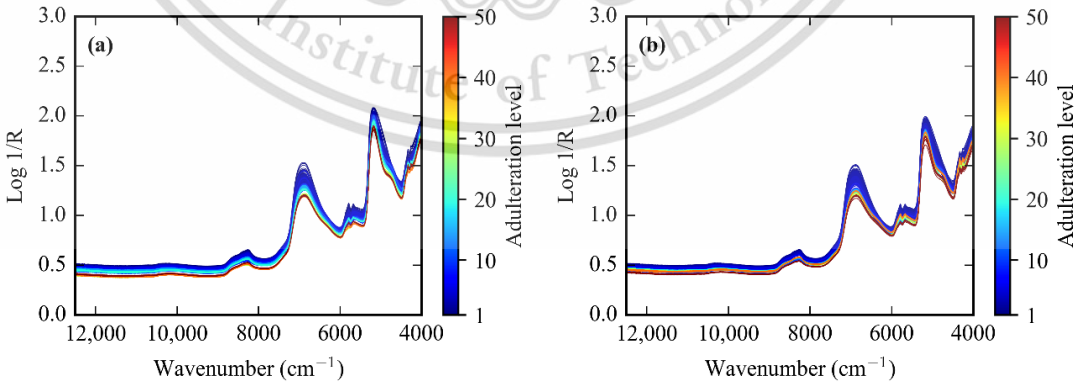
$$f(x) = \left[\sum_{i=1}^{N}\left(w_i x_i\right)\right] + b \quad (5.7)$$

Where, w is the average of weight, w_i is weight to-*i*, N is the number of samples, x_1 to x_n is the absorbance of feature NIR spectra, b is the intercept, and ε is perturbation coefficient (10^{-6}).

5.4 Results

5.4.1 NIR Spectra Features

NIR spectra to detect coconut milk adulteration in this study were acquired using two spectrophotometers, benchtop FT-NIR (450 spectra) and portable Micro-NIR (450 spectra). The original spectrum from the benchtop FT-NIR is presented in Figure 5.3a,b, and the original spectrum from the portable Micro-NIR is shown in Figure 5.3c,d. Figure 5.3a,c shows the FT-NIR and Micro-NIR spectra of coconut milk adulteration by corn flour from 1–50%. Meanwhile, the spectrum of FT-NIR and Micro-NIR of coconut milk adulteration by tapioca starch from 1–50% is shown in Figure 5.3b,d. The NIR spectrum indicates the presence of organic materials resulting from the interaction of molecular bonds of XH with the incident radiation from coconut milk, corn flour, and tapioca starch. The absorption peak positions appear almost indistinguishable, with only slight differences in absorption between adulteration levels. These bonds are subject to vibrational energy changes, including stretching and bending. The presence of strong water absorbance bands from FT-NIR was observed at around 5176 cm^{-1} (1932 nm) and 6889 cm^{-1} (1452 nm) because of the OH combination and its first overtone for both adulterant corn flour and tapioca starch. When utilizing Micro-NIR, strong water absorbance bands were identified around 1447 nm (6911 cm^{-1}) as an OH first overtone and 1212 nm (8251 cm^{-1}) as the second overtone of CH stretching for both the corn flour and tapioca stretching.



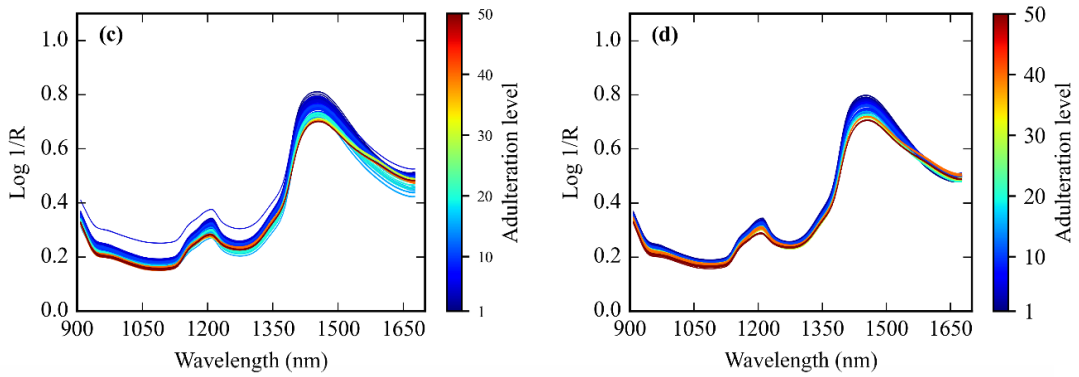


Figure 5.3. The original NIR spectroscopy data. Spectra by benchtop FT-NIR from coconut milk adulteration by (a) corn flour and (b) tapioca starch. Spectra by portable Micro-NIR for coconut milk adulteration by (c) corn flour and (d) tapioca starch.

5.4.2 Calibration Models Development Base on FT-NIR

5.4.2.1 Adulteration by Corn Flour

The results of the prediction of level adulteration corn flour in coconut milk utilization benchtop FT-NIR on training and testing data sets are presented in Table 5.2. The coefficient of determination (R^2) for all architecture network regressors was decreased from training to testing, which was inversely proportional to Bias and RMSE performance, which have increased from training to testing. As can be seen, all types of architecture networks have excellent performance model capability (RPD > 8.1) that can be expected for predictions of future samples. As for comparing the four architecture networks from the deep-learning regressor, the GoogleNET regressor (best RPD=20.866) possesses much higher detection accuracy than the other regressor. In other words, this study may organize the performance of regressor analysis order according to its RPD as ResNET < Simple CNN < S-AlexNET < GoogleNET. However, the GoogleNET regressor requires a higher epoch than the others.

The regression scatter plots of all architecture networks of deep-learning regression to predict level adulteration corn flour in coconut milk utilization FT-NIR are illustrated in Figure 5.4. The regression coefficient (slope) for all architecture network regressors decreased from training to testing. The intercept coefficient for

simple CNN and GoogleNET is positive, while S-AlexNET and ResNET are negative for training. However, all regressors' intercept coefficients are negative, except simple CNN in testing.

Table 5.2. Regression model performance to predict corn flour in coconut milk utilizing FT-NIR.

Regressor	Epoch	Training			Testing			
		R^2	RMSE	Bias	R^2	RMSE	Bias	RPD
Simple CNN	8035	0.999	0.370	−0.120	0.993	1.204	−0.012	11.884
S-AlexNET	3300	0.999	0.520	0.076	0.997	0.858	0.176	17.213
ResNET	5929	0.996	0.958	0.027	0.992	1.256	0.101	11.429
GoogleNET	10202	0.998	0.601	−0.037	0.998	0.686	0.012	20.866

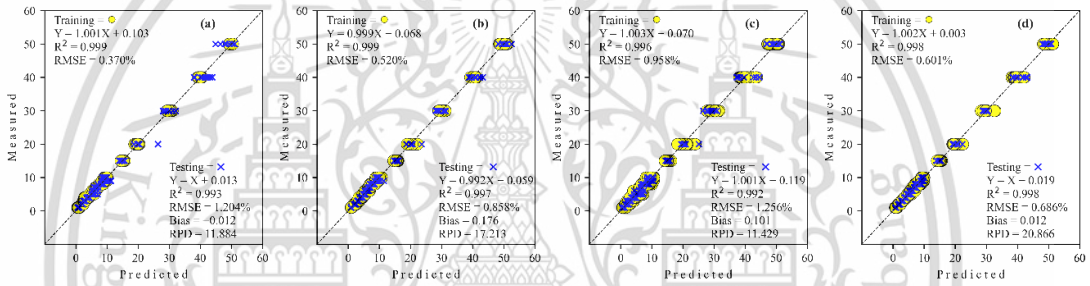


Figure 5.4. Regression plots obtained by deep learning to detect adulteration of coconut milk by corn flour using FT-NIR. (a) Simple CNN; (b) S-AlexNET; (c) ResNET; and (d) GoogleNET.

The range of the simple CNN coefficient from coconut milk adulteration by corn flour using FT-NIR is between 69.996 and −83.475 (Figure 5.5a). The wavenumbers that have more than score threshold 50% in the range of simple CNN coefficient of architecture are 7236 cm^{-1} (1382 nm), 7228 cm^{-1} (1384 nm), 7190 cm^{-1} (1391 nm), 7182 cm^{-1} (1392 nm), 7167 cm^{-1} (1395 nm), 5346 cm^{-1} (1871 nm), and 5338 cm^{-1} (1873 nm). They have as many as seven wavenumbers of feature importance.

Next, the range of the S-AlexNET coefficient is between 40.231 and −37.471 (Figure 5.5b). The wavenumbers that have more than score threshold 50% in the range of S-AlexNET coefficient of architecture are 7421 cm^{-1} (1348 nm), 7360 cm^{-1} (1359 nm), 7306 cm^{-1} (1369 nm), 7282 cm^{-1} (1373 nm), 7259 cm^{-1} (1378 nm), 7236

cm^{-1} (1382 nm), 7190 cm^{-1} (1391 nm), 7182 cm^{-1} (1392 nm), 7120 cm^{-1} (1404 nm), 6542 cm^{-1} (1529 nm), 6519 cm^{-1} (1534 nm), 5099 cm^{-1} (1961 nm), 4883 cm^{-1} (2048 nm), 4852 cm^{-1} (2061 nm), 4706 cm^{-1} (2125 nm), 4698 cm^{-1} (2129 nm), 4636 cm^{-1} (2157 nm), 4544 cm^{-1} (2201 nm), 4482 cm^{-1} (2231 nm), 4413 cm^{-1} (2266 nm), 4397 cm^{-1} (2274 nm), 4336 cm^{-1} (2306 nm), and 4328 cm^{-1} (2311 nm). They contain a total of 23 wavenumbers that are of important features.

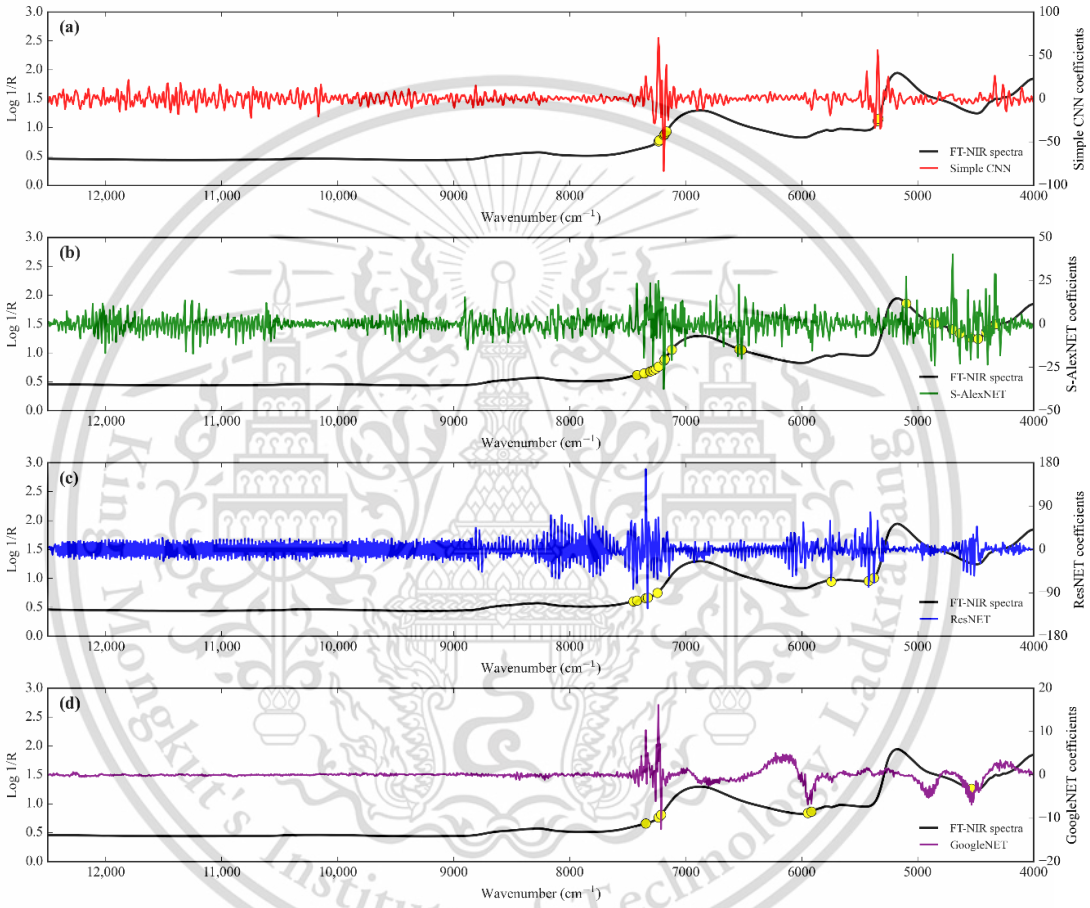


Figure 5.5. Comparison of the regression coefficients of the four deep-learning calibration approaches of adulteration coconut milk by corn flour using FT-NIR. (a) Simple CNN; (b) S-AlexNET; (c) ResNET; and (d) GoogleNET. ○ Feature importance.

Also, the range of the ResNET coefficient is between 166.725 and −122.456 (Figure 5.5c). The wavenumbers that have more than score threshold 50% in the range of ResNET coefficient of architecture are 7452 cm^{-1} (1342 nm), 7421 cm^{-1} (1348 nm), 7344 cm^{-1} (1362 nm), 7329 cm^{-1} (1364 nm), 7244 cm^{-1} (1380 nm), 5747 cm^{-1}

(1740 nm), 5423 cm^{-1} (1844 nm), and 5377 cm^{-1} (1860 nm). They consist of a total of eight spectral important features.

Finally, the range of the GoogleNET coefficient is between 16.124 and −12.524 (Figure 5.5d). The wavenumbers that have more than score threshold 50% in the range of GoogleNET coefficient of architecture are 7344 cm^{-1} (1362 nm), 7236 cm^{-1} (1382 nm), 7213 cm^{-1} (1386 nm), 5948 cm^{-1} (1681 nm), 5917 cm^{-1} (1690 nm), and 4536 cm^{-1} (2205 nm). There are a total of six wavenumbers of important features.

5.4.2.2 Adulteration by Tapioca Starch

The results of the prediction of level adulteration tapioca starch in coconut milk utilization benchtop FT-NIR on training and testing data sets are presented in Table 5.3. It is clear that the GoogleNET regressor requires a more increased epoch than the others. The coefficient of determination (R^2) for all architecture network regressors was decreased from training to testing, which is conversely proportional to Bias and RMSE performances, which have increased from training to testing. If focused on RPD, all types of architecture networks have excellent performance model capability (RPD > 8.1), except ResNET (RPD < 3), which shows poor and unreliable performance. As for comparing the four architecture networks from the deep-learning regressor, the GoogleNET regressor (best RPD=21.421) possesses a much higher prediction than the other regressor. If organized in the best possible way (based on RPD), the performance of deep-learning architecture network regressors of this study can be arranged into GoogleNET > S-AlexNET > Simple CNN > ResNET.

Table 5.3. Regression model performance to predict tapioca starch in coconut milk utilizing FT-NIR.

Regressor	Epoch	Training			Testing			
		R^2	RMSE	Bias	R^2	RMSE	Bias	RPD
Simple CNN	5633	0.995	0.977	0.039	0.995	1.034	0.034	14.067
S-AlexNET	3000	0.998	0.711	−0.299	0.996	0.951	−0.202	15.631
ResNET	2603	0.892	5.850	1.017	0.886	6.108	1.481	2.958
GoogleNET	10202	0.999	0.482	−0.035	0.998	0.670	0.054	21.421

The regression scatter plots of all architecture networks of deep-learning regression to predict level adulteration tapioca starch in coconut milk utilization FT-NIR are shown in Figure 5.6. The regression coefficient (slope) for all architecture network regressors decreases from training to testing, except S-AlexNET, which decreases very little (more than three decimal places), so the effect is minimal. Also, the intercept coefficient for all regressors is positive, while Simple CNN is negative for training. Inversely proportional to testing, all regressors are negative except S-AlexNET.

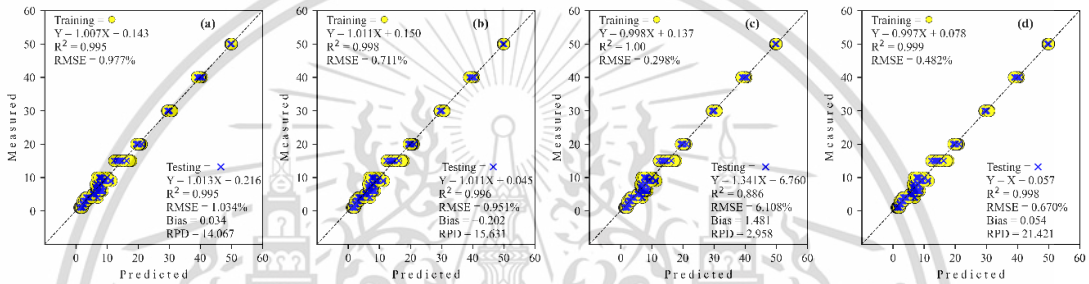


Figure 5.6. Regression plots obtained by deep learning to detect adulteration of coconut milk by tapioca starch using FT-NIR. (a) Simple CNN, (b) S-AlexNET, (c) ResNET and (d) GoogleNET.

The simple CNN coefficient range for detecting tapioca starch in coconut milk using FT-NIR is 1.8127 to −1.892. Wavenumbers with more than score threshold 50% of the simple CNN coefficient of architecture as many as 51 important features include 7213–7182 cm^{-1} (1386–1392 nm), 6256–5963 cm^{-1} (1598–1677 nm), and 5832–5778 cm^{-1} (1715–1731 nm) (Figure 5.7a).

Next, the S-AlexNET coefficient range is from 21.049 to −24.092 (Figure 5.7b). Wavenumbers with more than score threshold 50% of the S-AlexNET coefficient of architecture include 11,926 cm^{-1} (839 nm), 11,865 cm^{-1} (843 nm), 11,556 cm^{-1} (865 nm), 8208 cm^{-1} (1218 nm), 7190 cm^{-1} (1391 nm), 6542 cm^{-1} (1529 nm), 6380 cm^{-1} (1567 nm), 6148 cm^{-1} (1627 nm), 5007–4690 cm^{-1} (1997–2132 nm), 4667 cm^{-1} (2143 nm), 4636 cm^{-1} (2157 nm), 4413 cm^{-1} (2266 nm), 4328 cm^{-1} (2311 nm), 4289 cm^{-1} (2332 nm), and 4266 cm^{-1} (2344 nm) (a total of 51 important wavenumbers).

Also, the range of the ResNET coefficient is from 20.277 to −20.429. Wavenumbers with more than score threshold of 50% of the ResNET coefficient of architecture as many as 32 importance spectral include 7090 cm^{-1} (1410 nm), 7074 cm^{-1} (1414 nm), 7012 cm^{-1} (1426 nm), 6982–6750 cm^{-1} (1432–1481 nm), 6434–6403 cm^{-1} (1554–1562 nm), 5693–5408 cm^{-1} (1757–1849 nm), and 4621–4498 cm^{-1} (2164–2223 nm) (Figure 5.7c).

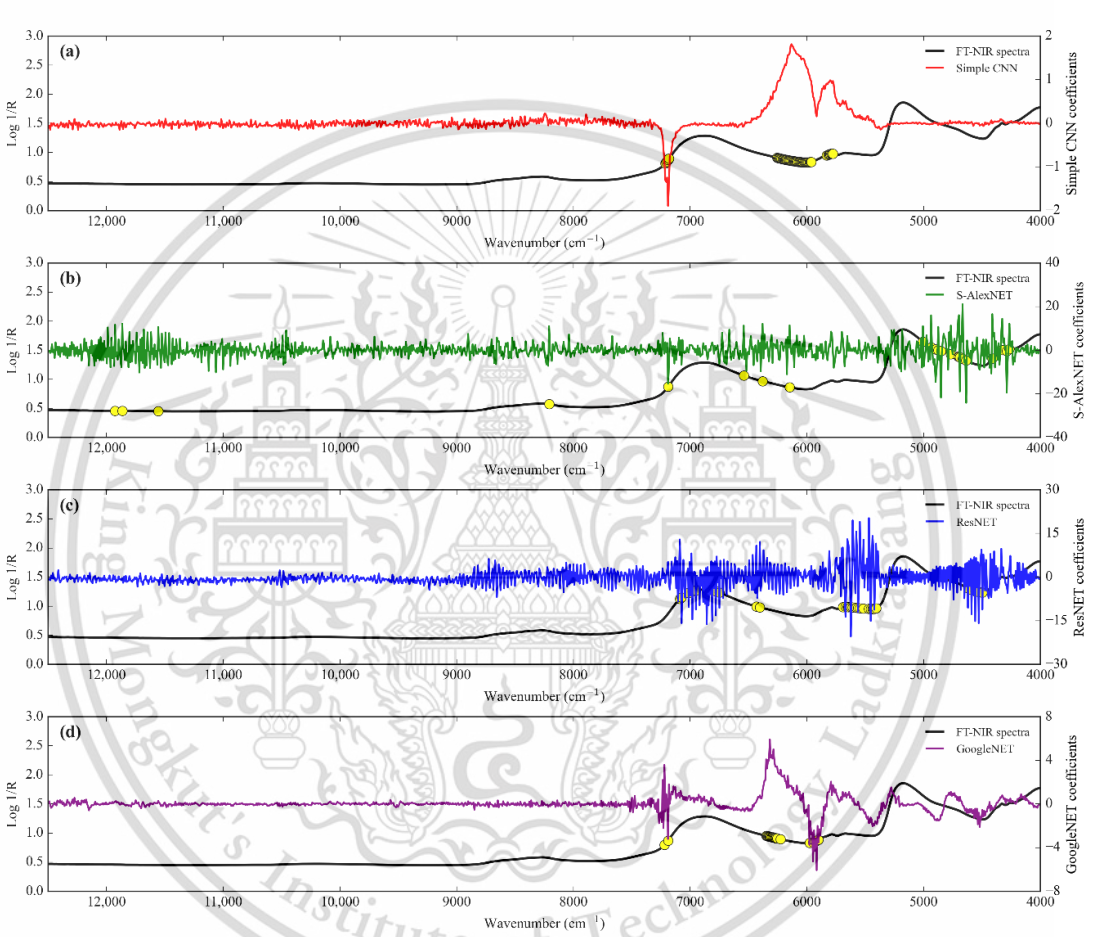


Figure 5.7. Comparison of the regression coefficients of the four deep-learning calibration approaches to adulteration of coconut milk by tapioca starch using FT-NIR. (a) Simple CNN; (b) S-AlexNET; (c) ResNET; and (d) GoogleNET. ○ Feature importance.

Finally, the GoogleNET coefficient range is from 5.915 to −6.067 (Figure 5.7d). Wavenumbers with more than a score threshold of 50% of the GoogleNET coefficient of architecture include 7221–7190 cm^{-1} (1385–1391 nm), 6349–6226 cm^{-1} (1575–

1606 nm), and 5979–5902 cm^{-1} (1673–1694 nm) (a total of 22 importance wavenumbers).

5.4.3 Calibration Models Development Base on Micro-NIR

5.4.3.1 Adulteration by Corn Flour

The performances of the different network architectures to predict the level of adulteration of coconut milk by corn flour using Micro-NIR during training and testing are summarized in Table 5.4. The GoogleNET regressor needs a more significant number of epochs compared to the other regressors. The ResNET regressor is the one that provided the best results (same R^2 with GoogleNET regressor but with the lowest RMSE) during training. The GoogleNET regressor possesses a much higher coefficient of determination (R^2) and lowest RMSE than the other regressor in the test set. This is confirmed by the RPD parameter for the GoogleNET regressor, which has excellent performance model capability (RPD=31.094). In simpler terms, this study may also arrange the performance of regressor order according to its RPD as Simple CNN < ResNET < S-AlexNET < GoogleNET.

Table 5.4. Regression model performance to predict corn flour in coconut milk utilizing Micro-NIR.

Regressor	Epoch	Training			Testing			
		R^2	RMSE	Bias	R^2	RMSE	Bias	RPD
Simple CNN	6596	0.998	0.706	−0.084	0.998	0.597	−0.023	23.981
S-AlexNET	3300	0.998	0.603	−0.183	0.999	0.532	−0.123	28.599
ResNET	6091	0.999	0.363	−0.129	0.998	0.575	−0.065	25.210
GoogleNET	10128	0.999	0.414	−0.053	0.999	0.463	−0.029	31.094

All predictive performances from deep-learning regressors in scatter plots to predict the level adulteration of corn flour in coconut milk operating Micro-NIR are shown in Figure 5.8. The regression coefficient (slope) for S-AlexNET and ResNET regressors decreased from training to testing. Meanwhile, Simple CNN tends to be stable, and GoogleNET overlooks the increase. The intercept coefficient for all regressors, both in training and testing, is positive except for the GoogleNET regressor in the testing stages.

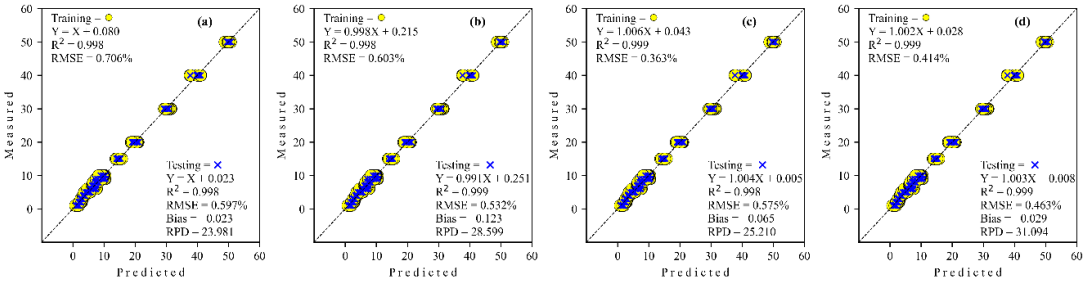


Figure 5.8. Regression plots obtained by deep learning to detect adulteration of coconut milk by corn flour using Micro-NIR. (a) Simple CNN; (b) S-AlexNET; (c) ResNET; and (d) GoogleNET.

The range of the simple CNN coefficient from coconut milk adulteration by corn flour using Micro-NIR is between 24.486 and –29.476 (Figure 5.9a). The wavelengths that have more than score threshold 50% in the range of simple CNN coefficient of architecture (seven wavelengths) are 921 nm (10858 cm^{-1}), 1206–1212 nm (8292–8251 cm^{-1}), 1224 nm (8170 cm^{-1}), 1385 nm (7220 cm^{-1}), 1391 nm (7189 cm^{-1}), and 1410 nm (7092 cm^{-1}).

Next, the range of the S-AlexNET coefficient is between 61.807 and –70.932 (Figure 5.9b). The wavelengths that have more than score threshold 50% in the range of S-AlexNET coefficient of architecture (15 wavelengths) are 1119 nm (8937 cm^{-1}), 1175 nm (8511 cm^{-1}), 1199 nm (8340 cm^{-1}), 1205–1212 nm (8299–8251 cm^{-1}), 1230–1236 nm (8130–8091 cm^{-1}), 1249–1255 nm (8006–7968 cm^{-1}), 1280 nm (7813 cm^{-1}), 1317 nm (7593 cm^{-1}), 1404 nm (7123 cm^{-1}), 1416 nm (7062 cm^{-1}), 1428 nm (7003 cm^{-1}), and 1515 nm (6601 cm^{-1}).

Likewise, the range of the ResNET coefficient is between 140.050 and –122.771 (Figure 5.9c). The wavelengths that have more than score threshold 50% in the range of ResNET coefficient of architecture are 1106 nm (9042 cm^{-1}), 1150 nm (8696 cm^{-1}), 1181 nm (8467 cm^{-1}), 1212 nm (8251 cm^{-1}), 1224 nm (8170 cm^{-1}), 1274 nm (7849 cm^{-1}), 1342–1360 nm (7452–7353 cm^{-1}), 1385–1391 nm (7220–7189 cm^{-1}), 1515 nm (6601 cm^{-1}), 1540 nm (6494 cm^{-1}), 1559 nm (6414 cm^{-1}), 1590 nm (6289 cm^{-1}), 1602 nm (6242 cm^{-1}), 1614 nm (6196 cm^{-1}), 1627 nm (6146 cm^{-1}), 1639 nm (6101 cm^{-1}), 1645 nm (6079 cm^{-1}), and 1664 nm (6010 cm^{-1}). They contain a total of 22 wavelengths that are of important features.

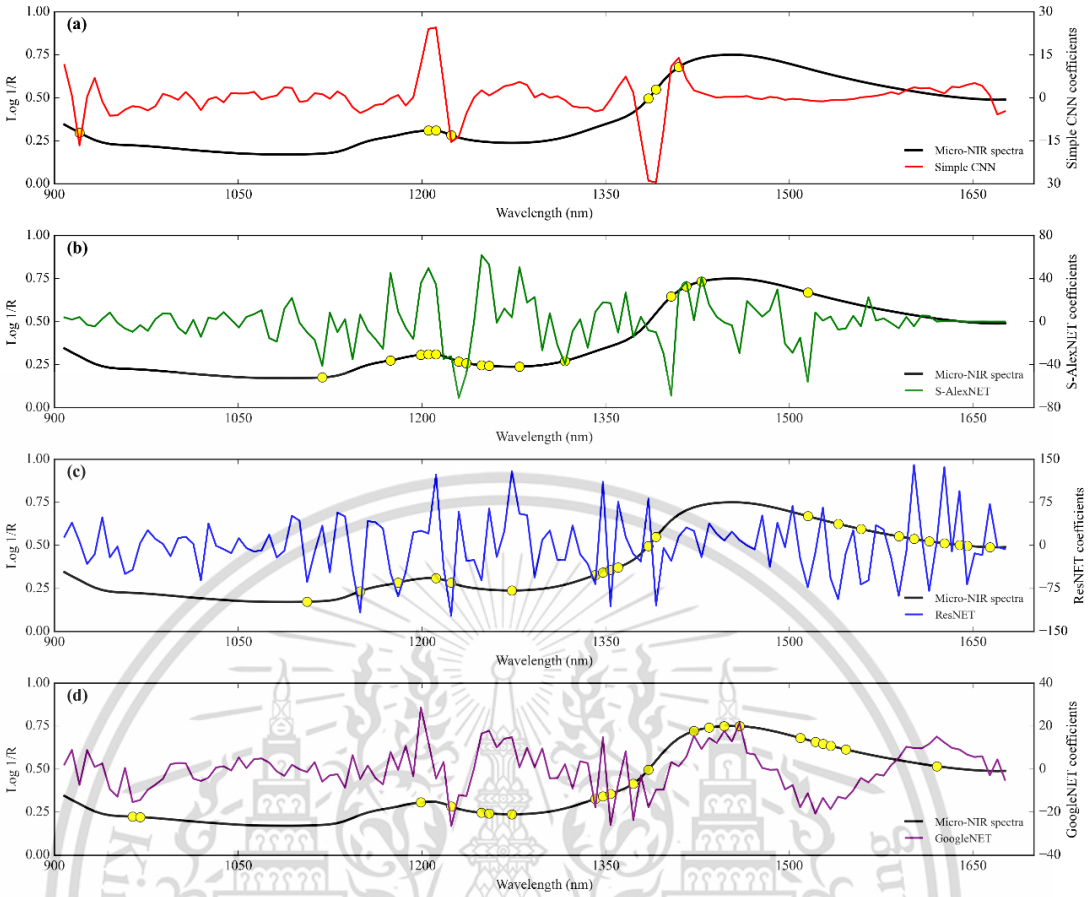


Figure 5.9. Comparison of the regression coefficients of the four deep-learning calibration approaches of coconut milk adulteration by corn flour using Micro-NIR. (a) Simple CNN; (b) S-AlexNET; (c) ResNET; and (d) GoogleNET. ○ Feature importance.

Lastly, the range of the GoogleNET coefficient is between 28.666 and −26.421 (Figure 5.9d). The wavelengths that have more than score threshold 50% in the range of GoogleNET coefficient of architecture are 964–970 nm (10,373–10,309 cm^{-1}), 1199 nm (8340 cm^{-1}), 1224 nm (8170 cm^{-1}), 1249–1255 nm (8006–7968 cm^{-1}), 1274 nm (7849 cm^{-1}), 1342–1348 nm (7452–7418 cm^{-1}), 1354 nm (7386 cm^{-1}), 1373 nm (7283 cm^{-1}), 1385 nm (7220 cm^{-1}), 1422 nm (7032 cm^{-1}), 1435 nm (6969 cm^{-1}), 1447 nm (6911 cm^{-1}), 1459 nm (6854 cm^{-1}), 1509 nm (6627 cm^{-1}), 1521–1528 nm (6575–6545 cm^{-1}), 1534 nm (6519 cm^{-1}), 1546 nm (6468 cm^{-1}), and 1621 nm (6169 cm^{-1}). They have as many as 22 wavelengths that feature importance.

5.4.3.2 Adulteration by Tapioca Starch

Results of the prediction for the adulteration degree of corn flour in coconut milk using Micro-NIR, as seen in the training and testing data sets, are displayed in Table 5.5. The coefficient of determination (R^2) for all architecture network regressors is a teeny change from training to testing. The RMSE performance decreased for simple CNN and S-AlexNET while increasing for ResNET and GoogleNET from training to testing. However, Bias performance for simple CNN and ResNET decreased during training to testing, and S-AlexNET and GoogleNET increased during training to testing. Regarding RPD, all architecture networks demonstrate excellent performance model capability (RPD > 8.1), with the best coming from ResNET (RPD=39.349). To elaborate, this study may organize the best performance of regressor analysis order according to its RPD as ResNET > S-AlexNET > GoogleNET > Simple CNN with GoogleNET regressor requires a higher epoch than others.

Table 5.5. Regression model performance to predict tapioca starch in coconut milk using Micro-NIR.

Regressor	Epoch	Training			Testing			
		R^2	RMSE	Bias	R^2	RMSE	Bias	RPD
Simple CNN	8872	0.998	0.637	−0.105	0.998	0.611	−0.044	23.521
S-AlexNET	2700	0.999	0.428	−0.029	0.999	0.419	−0.065	34.880
ResNET	7814	1.000	0.298	−0.111	0.999	0.370	−0.068	39.349
GoogleNET	9840	0.999	0.431	0.041	0.999	0.461	0.035	31.095

The regression scatter plots for predicting the adulteration level of tapioca starch in coconut milk using Micro-NIR with various deep-learning architecture networks are depicted in Figure 5.10. The regression coefficient (slope) for all architecture network regressors increases during training to testing except GoogleNET, which decreases very little (more than three decimal places), so the effect is minimal. Also, all regressors' intercept coefficients are positive during training and testing, while GoogleNET is negative for both training and testing.

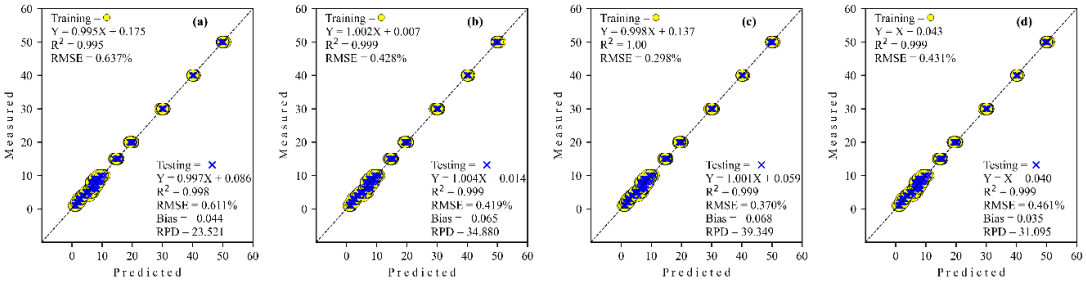


Figure 5.10. Regression plots obtained to detect adulteration of coconut milk by tapioca starch using Micro-NIR. (a) Simple CNN; (b) S-AlexNET; (c) ResNET; and (d) GoogleNET.

The simple CNN coefficient range for detecting tapioca starch in coconut milk using Micro-NIR is from 25.180 to –32.352. Wavelengths with more than score threshold 50% of the simple CNN coefficient of architecture with as many as 13 importance features include 908 nm (11013 cm^{-1}), 1385 nm (7220 cm^{-1}), 1391 nm (7189 cm^{-1}), 1404 nm (7123 cm^{-1}), 1410 nm (7092 cm^{-1}), 1614–1651 nm (6196–6057 cm^{-1}), and 1670 nm (5988 cm^{-1}) (Figure 5.11a).

Next, the S-AlexNET coefficient range is from 80.792 to –114.579 (Figure 5.11b). Wavelengths with more than score threshold 50% of the S-AlexNET coefficient of architecture include 1218 nm (8210 cm^{-1}), 1249 nm (8006 cm^{-1}), 1255 nm (7968 cm^{-1}), 1267 nm (7837 cm^{-1}), 1354 nm (7386 cm^{-1}), 1367 nm (7315 cm^{-1}), 1404 nm (7123 cm^{-1}), 1410 nm (7092 cm^{-1}), 1416 nm (7062 cm^{-1}), 1428 nm (7003 cm^{-1}), 1435 nm (6969 cm^{-1}), and 1453 nm (6882 cm^{-1}) (a total of 12 importance wavelengths).

Likewise, the range of the ResNET coefficient is from 178.959 to –219.921. Wavelengths with more than a score threshold of 50% of the ResNET coefficient of architecture as many as five important spectra include 1187 nm (8425 cm^{-1}), 1193 nm (8382 cm^{-1}), 1360 nm (7353 cm^{-1}), 1552 nm (6443 cm^{-1}), and 1559 nm (6414 cm^{-1}) (Figure 5.11c).

Lastly, the GoogleNET coefficient range is from 28.693 to –36.384 (Figure 5.11d). Wavelengths with more than score threshold of 50% of the GoogleNET coefficient of architecture include 914 nm (10,941 cm^{-1}), 921 nm (10,858 cm^{-1}), 939 nm (10,650 cm^{-1}), 952 nm (10504 cm^{-1}), 964 nm (10,373 cm^{-1}), 976 nm (10,246 cm^{-1}), 1224 nm (8170 cm^{-1}), 1255 nm (7968 cm^{-1}), 1261 nm (7930 cm^{-1}), 1274 nm (7849 cm^{-1}), 1367 nm (7315 cm^{-1}), 1373 nm (7283 cm^{-1}), 1379 nm (7252 cm^{-1}), 1422–1459 nm (7032–

6854 cm^{-1}), 1509 nm (6627 cm^{-1}), 1521–1552 nm (6575–6443 cm^{-1}), and 1602–1633 nm (6242–6124 cm^{-1}) (a total of 33 importance wavelengths).

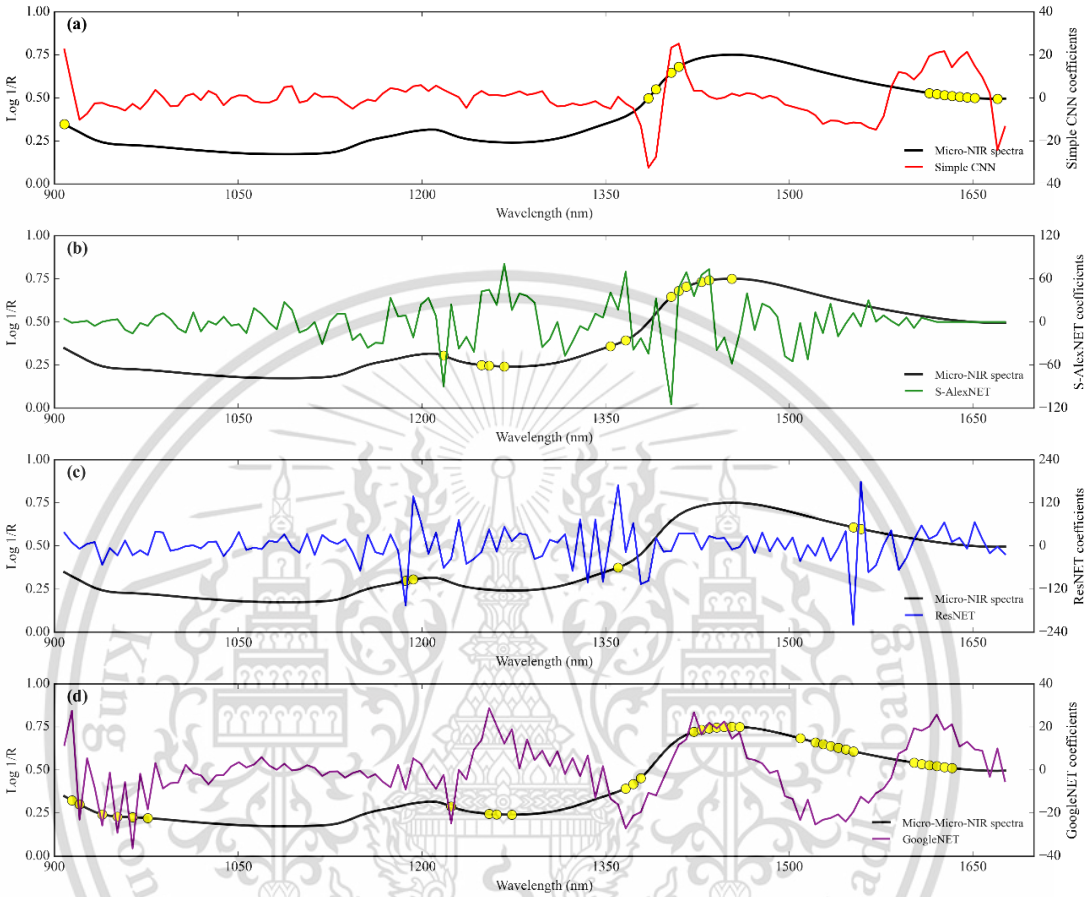


Figure 5.11. Comparison of the regression coefficients of the four deep-learning calibration approaches of adulteration coconut milk by tapioca starch using Micro-NIR. (a) Simple CNN; (b) S-AlexNET; (c) ResNET; and (d) GoogleNET. ● Feature importance.

5.5 Discussions

This study explored the feasibility of using deep learning to create a rapid, accurate, and robust prediction model to predict adulteration levels of coconut milk by corn flour and tapioca starch using FT-NIR and Micro-NIR spectroscopy. Presently, the non-destructive testing of adulteration in agriculture and food products by NIR spectrometer based on laboratory conditions has been widely introduced. However, the procedures and development of the calibration model under which it can be

applied are still limited. The use of deep learning can be a good solution to such problems. Additionally, compared with the results of this study's use of deep learning, for example, coffee adulteration prediction (Nallan Chakravartula *et al.*, 2022), adulteration in infant formula (Liu *et al.*, 2021), cow milk fat content adulteration by water (Said *et al.*, 2022), and minced beef adulteration (Weng *et al.*, 2020), we also obtained equally superior prediction results. Furthermore, the operation of the NIR spectrophotometer is simpler and easier to promote.

This article presents a novel deep-learning regression method for quantitative NIR spectrum analysis. This method utilizes four network architectures: Simple convolutional neural network (Simple CNN); S-AlexNET; ResNET; and GoogleNET. However, as is known, a robust NIR model for adulteration detection is hard to achieve due to multiple variation factors, such as different brands and batches of product, the simultaneous existence of several adulterants, temperature, humidity, and spectral drift of light sources, making it hard to obtain stable applications in practice. Therefore, more advanced modeling investigations should be carried out and prepared to evaluate and improve the robustness of the proposed method for the future. However, the limitations of the proposed method should also be further considered and improved. For example, the deep-learning method is much more time-consuming in training than the traditional method and the regular machine-learning algorithm. Therefore, some adulteration studies in food and agriculture products are based on NIR spectroscopy run deep-learning algorithms on graphics processing units (GPU), such as the assurance of tea quality by Yang *et al.* (2021) and detection of adulteration of minced beef by Weng *et al.* (2020). However, thanks to the fast development of deep-learning hardware, for instance, graphics processing unit (GPU), associative processing unit (APU), tensor processing unit (TPU), and quantum processing unit (QPU), the testing time for the proposed network is acceptable.

As can be observed from Figure 5.3, the spectral profiles of the degree of coconut milk adulterants by corn flour and tapioca starch were similar and characterized by few substantial differences in peak positions and curve trends. In general, for all the sample adulterants, a few characteristic overlapping peaks contributed by the presence of the main content of coconut milk and adulterant

material, including fat, protein, moisture, ash, and carbohydrates. Samples with more adulterant material caused the peak absorbance level to decrease, both for adulteration by corn flour and tapioca starch. This corresponds to the difference in moisture content between coconut milk and adulterant, which causes the free moisture content in the coconut milk to be absorbed by the admixture agent to reach an equilibrium point. As a result, coconut milk that has been adulterated more with a solid adulterant material has a lower absorption spectral ability. This is in line with the report by Malvandi *et al.* (2022), who stated in their study that peak values and their corresponding wavelengths in the NIR region changed as the moisture content altered. Büning-Pfaue (2003) emphasized that the strength and weakness of this absorption band come from the strong effect of hydrogen bonds on organic monomers, ions, and polymers in the sample. The presence of content in adulteration coconut milk samples was observed at the main peaks of the following wavebands, both FT-NIR and Micro-NIR: 1210 nm (8262 cm^{-1}) related predominantly to CH bond stretching with the second overtone; 1453 nm (6881 cm^{-1}) to the first overtone of OH stretching bonds attributed to starch and water; 1728 nm (5786 cm^{-1}) and 1764 nm (5670 cm^{-1}) to the resonance bands of CH bond stretching with the first overtone; and 1929 nm (5184 cm^{-1}) to the CH bonds stretching with the second overtone (Conzen, 2006; Osborne *et al.*, 1993; Workman Jr and Weyer, 2007).

The performance criteria for the prediction model using deep learning in this study were evaluated based on predicting grain chemical composition content. A study by Chu *et al.* (2022) examined the regression model's capacity to classify RPD in the following manner: less than 3 as a poor or unreliable model, 3.1–4.9 as a fair model, 5.0–6.4 as a good model, 6.5–8.0 as a very good model and more than 8.1 as an excellent model. When comparing the RPD results for the prediction degree of adulteration of coconut milk with corn flour using all the network architectures, it was observed that Micro-NIR was superior to using FT-NIR (for all network architectures). At the same time, all the network architectures were considered excellent models. For tapioca starch in coconut milk case, Micro-NIR performed better than FT-NIR based on RPD among the spectrophotometer to predict the degree of adulteration (Table 5.6). Subsequently, only ResNET has lower RPD and weak performance (FT-NIR data set) among the network architectures. Regarding the

comparison between FT-NIR and Micro-NIR, the models developed for the prediction of coconut milk by adulterant material corn flour and tapioca starch seem to give comparable results using the FT-NIR. The performance of FT-NIR slightly reduced RPD, perhaps due to some factors, including a lack of explanatory variables and collinearity, but fortunately, the RPD obtained is still higher than eight (Basri *et al.*, 2018; Lan *et al.*, 2021).

Table 5.6. Summary of the performance of FT-NIR and Micro-NIR.

Adulteration Material	Instruments	The Best Regressor	RPD
Corn flour	FT-NIR	GoogleNET	20.866
	Micro-NIR	GoogleNET	31.094
Tapioca starch	FT-NIR	GoogleNET	21.421
	Micro-NIR	ResNET	39.349

In chemometrics, a limited number of samples with high-dimensional data of features pose common problems like data overfitting and multicollinearity and do not show the main features that are more dominant in the data. Selection of the most important features can lead to the dominant variables in a high-dimensional dataset. In case studies on NIR spectra, this can be represented in many methods, one being by expressing slope coefficients or regression coefficients. According to a study from Palermo *et al.* (2009), regression coefficients can be used to select appropriate predictors according to the magnitude of their absolute values. Even according to a study by Wold *et al.* (2001), in classical chemometric analysis using partial least squares (PLS), small regression coefficients can be ignored as an unimportant term to find the most prominent features and correlate them with the chemical assignment of some structure and bond vibration in the NIR spectroscopy. Additionally, compared with the results of previous research using this approach in analysis in NIR spectroscopy, for example, extra virgin olive oil adulteration prediction by PLS regressor (Jiang and Chen, 2019), adulteration in quinoa flour by PLS regressor (Wang *et al.*, 2022), aged-rice adulteration by competitive adaptive reweighted sampling (CARS) combined with PLS regressor (Li *et al.*, 2023), and adulterants of notoginseng powder by CARS-PLS regressor (Chen *et al.*, 2019). However, in applying

advanced chemometrics using machine learning and deep learning, it is still a challenge to demonstrate coefficients that can represent important features.

The regression coefficients from deep-learning algorithms used in this study can be represented using weight coefficients. Even though it is not strictly identical to the regression coefficients in classical chemometric analysis using PLS, at least the weight coefficients of each deep-learning network architecture can indicate variables for each response that are more important than others. Regression coefficients for the case of deep-learning regressors were first introduced by Cui and Fearn (2018) and tested on three NIR datasets, including the wheat flour dataset, wheat flour dataset, and protein content dataset. In their study, they randomly draw a few spectra from the dataset and plot the corresponding regression coefficients. This is understandable because deep learning is a non-linear approach, so each spectrum will have its own regression coefficient value, different from the PLS regression coefficient, which has the same value for all sample spectra. However, in this study, because the aim of showing the coefficients of weight of each deep-learning network architecture is to find the dominant features in high-dimensional data, we use all training data spectra. Next, we average the weight coefficients of all the training data spectra, which are called regression coefficients for each deep-learning network architecture. In this study, we apply a threshold score of 50% of the maximum and minimum peaks in the regression coefficient, as shown in Figure 5.5, Figure 5.7, Figure 5.9, and Figure 5.11. This approach is similar to the system applied in the variable importance in projection (VIP) approach from PLS regression, which applies a threshold score rule that can be data specific, ranging between 0.83 and 1.21 (Chong and Jun, 2005; Wang *et al.*, 2022).

In the case of scanning corn flour in coconut milk (Figure 5.5 and Figure 5.9), we can see regression coefficients related to the presence of the structural groups CH and OH. In general, the regression coefficients in this study are in the range of 1200–1500 nm (8333–6667 cm^{-1}), which is related to the main wavelength of corn flour found by Jiang and Lu (2018). In the case of scanning with FT-NIR, we can see regression coefficients that overlap with all deep-learning network architectures, at least across nine NIR bands. This starts from wave 7421 cm^{-1} (1348 nm), which is related to the fourth overtone of CH_2 (Workman Jr and Weyer, 2007). Next, waves at

7306–7329 cm^{-1} (1369–1364 nm) and 7344 cm^{-1} (1362 nm) are a combination of CH stretching and CH deformation from CH_3 (Osborne *et al.*, 1993). Waves at 7213–7228 cm^{-1} (1386–1384 nm) and 7236–7244 cm^{-1} (1382–1380 nm) correspond to OH stretching from H_2O (Conzen, 2006). Furthermore, waves at 7167–7190 cm^{-1} (1395–1391 nm) and 7182 cm^{-1} (1392 nm) are related to a combination of CH stretching and CH deformation from CH_2 (Osborne *et al.*, 1993). Lastly, the wave at 4536–4544 cm^{-1} (2205–2201 nm) is related to CH stretching and C=O stretching from CHO (Osborne *et al.*, 1993). However, the regression coefficients that overlap with all deep-learning network architectures that scan using Micro-NIR are 12 NIR bands. The wave was detected starting from 1205–1206 nm (8299–8292 cm^{-1}), the fourth overtone of aromatic CH, to 1212 nm (8251 cm^{-1}) and 1224 nm (8170 cm^{-1}), the second overtone of CH_2 and CH (Osborne *et al.*, 1993; Workman Jr and Weyer, 2007). In addition, waves in the range of 1249–1274 nm (8006–7849 cm^{-1}), 1342–1348 nm (7452–7418 cm^{-1}), and 1354–1404 nm (7386–7123 cm^{-1}) are the fourth overtone beta-diketone, the fourth overtone CH_2, and the third overtone aldehydes, respectively (Workman Jr and Weyer, 2007). Furthermore, waves at 1391 nm (7189 cm^{-1}), 1404–1410 nm (7123–7092 cm^{-1}), and 1416–1422 nm (7062–7032 cm^{-1}) are the representations of combination CH stretching with CH deformation, the first overtone of OH stretching, and a combination CH stretching with CH deformation and the first overtone OH stretching, respectively (Osborne *et al.*, 1993). Finally, the waves at 1515 nm (6601 cm^{-1}), 1540 nm (6494 cm^{-1}), and 1614–1621 nm (6196–6169 cm^{-1}) are related to the first overtone of CH, the first overtone of OH (starch), and the first overtone of $=CH_2$, respectively (Osborne *et al.*, 1993; Workman Jr and Weyer, 2007).

When examining the tapioca starch in coconut milk (Figure 5.7 and Figure 5.11), we see regression coefficients associated with the existence of the structural groups CH, CC, CNO, and OH. The structural groups detected in this sample were relatively slightly different from the adulteration of coconut milk with corn flour. This is due to the composition of the adulterant material, which is also different. The study by Williams (2009) was confirmed by Phetpan and Sirisomboon (2015), who stated that the peak in the 1400 nm (7143 cm^{-1}) region was associated with the glucose molecules in the tapioca starch constituents. The regression coefficients that cover

the overlap for all deep learning network architectures when using FT-NIR spectroscopy are in the five NIR spectral bands. Starting from waves 7213–7221 cm^{-1} (1386–1385 nm), 7182–7190 cm^{-1} (1392–1391 nm), 6380–6403 cm^{-1} (1567–1562 nm), 5963–5979 cm^{-1} (1677–1673 nm), and 4621–4636 cm^{-1} (2164–2157 nm), which correspond to the third overtone carbonyl stretching, CH_2 combination stretching and deformation, the second overtone CC stretching, the first overtone aromatic CH stretching, and the second overtone CNO, respectively (Osborne *et al.*, 1993; Workman Jr and Weyer, 2007). Furthermore, the regression coefficients that cover the overlap for all deep-learning network architectures when using Micro-NIR spectroscopy are in the 11 NIR spectral bands. Starting from waves 1255–1276 nm (7968–7837 cm^{-1}), 1360–1367 nm (7353–7315 cm^{-1}), 1379–1385 nm (7252–7220 cm^{-1}), 1404 nm (7123 cm^{-1}), and 1410 nm (7092 cm^{-1}), which correspond to the third overtone CC stretching, combination stretching, and deformation from CH_3, the third overtone carbonyl stretching, the third overtone carbonates, and the first overtone of OH stretching, respectively (Osborne *et al.*, 1993; Workman Jr and Weyer, 2007). Next, waves at 1416 nm (7062 cm^{-1}), 1428 nm (7003 cm^{-1}), 1552–1553 nm (6443–6882 cm^{-1}), 1614–1621 nm (6196–6169 cm^{-1}), and 1627–1633 nm (6146–6124 cm^{-1}) are related to combination stretching and deformation of CH_2, the first overtone of NH stretching, the third overtone carbonyl stretch, the first overtone of OH stretching, the first overtone of $=CH_2$ stretching, and the first overtone of CH stretching, respectively (Osborne *et al.*, 1993; Workman Jr and Weyer, 2007).

This study analyzed important wavelengths by deep learning and found that not all important wavelengths will be the same for all deep-learning network architectures. In addition, even though the FT-NIR wavelength range covers the wavelength range in Micro-NIR, the important wavelength will not be precisely the same for both. However, in the case of FT-NIR and Micro-NIR instruments, it was still found that some important waves overlapped between them. In the case of corn flour, waves 7421 cm^{-1} (1348 nm) and 7167–7190 cm^{-1} (1395–1391 nm) were found in FT-NIR, and 1342–1348 nm (7452–7418 cm^{-1}) and 1391 nm (7189 cm^{-1}) were found in Micro-NIR. In the case of tapioca starch, the waves at 7213–7221 cm^{-1} (1386–1385 nm) were found in FT-NIR and 1379–1385 nm (7252–7220 cm^{-1}) in Micro-NIR. This may be caused by the nature of each regressor, which in the convolutional

layer stage can transform the spectra to fit in the following regression scheme. In other words, the regressor from deep learning has carried out automatic preprocessing, as reported by Cui and Fearn (2018). This causes the final shape of each spectrum before the "flatten" to the "dense fully connected" stage to differ according to the output variable. This difference will eventually result in differences in important wavelengths for each deep-learning network architecture. Even though they are different, several wavelengths from all deep-learning network architectures are still the same.

5.6 Conclusions

Deep learning as a novel approach to predict the level of adulteration of coconut milk was successfully developed and tested based on spectra from benchtop FT-NIR and portable Micro-NIR. Models based on FT-NIR spectroscopy to be able to predict the adulteration level of corn flour in coconut milk (1–50%) can be generated using architecture network regressor from deep learning (Simple CNN, S-AlexNET, ResNET, GoogleNET) in the performance ranges of R^2, RMSE, and Bias at their training from 0.996 to 0.999, from 0.370 to 0.958%, and from −0.027 to 0.120, respectively. Next, R^2, RMSE, Bias, and RPD at the testing stage are from 0.992 to 0.998, 0.686 to 1.256%, from −0.012 to 0.176, and from 11.429 to 20.866, respectively. Even though it is still as good, the performance based on the FT-NIR prediction model is still lower than that of Micro-NIR with the same regressor network architecture from deep learning. Performance ranges R^2, RMSE, and Bias at their training using Micro-NIR are from 0.998 to 0.999, from 0.363 to 0.706%, and from −0.053 to −0.183, respectively. At the testing stage, R^2, RMSE, Bias, and RPD are from 0.998 to 0.999, from 0.463 to 0.597%, from −0.023 to 0.123, and from 23.981 to 31.094, respectively.

Relatively similar to the case of the model to predict tapioca starch adulteration in coconut milk, the performance based on the Micro-NIR dataset is better than using FT-NIR with the same regressor network architecture from deep learning. Performance ranges at their training (R^2, RMSE, Bias) and testing (R^2, RMSE, Bias, RPD) using Micro-NIR are from 0.998 to 1.000, from 0.298 to 0.637%, from −0.029 to −0.111, from 0.998 to 0.999, from 0.370 to 0.611%, from −0.035 to −0.068, and

from 23.521 to 39.349, respectively. Meanwhile, performance ranges at their training (R^2, RMSE, Bias) and testing (R^2, RMSE, Bias, RPD) using FT-NIR are from 0.892 to 0.999, from 0.482 to 5.850%, from −0.035 to 1.017, from 0.886 to 0.998, from 0.670 to 6.108%, from −0.202 to 1.481, and from 2.958 to 21.421, respectively.

In closing, the prediction results demonstrated that the proposed architecture from the deep-learning method yielded superior regression performance for the FT-NIR and Micro-NIR to predict the level of adulterants (corn flour and tapioca starch) in coconut milk. While finding that the optimal deep-learning architecture is complex and computationally expensive, implementation and training are straightforward once found. Furthermore, developing deep-learning architectures and applying them are two different study matters that should not be confused. This study also indicated that deep learning for NIR spectroscopy data is less dependent on preprocessing than the classical chemometrics method and still can achieve excellent performance.

5.7 References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., & Isard, M. (2016). *TensorFlow: a system for Large-Scale machine learning*. Paper presented at the 12th USENIX symposium on operating systems design and implementation (OSDI 16).

Acquarelli, J., van Laarhoven, T., Gerretzen, J., Tran, T. N., Buydens, L. M. C., & Marchiori, E. (2017). Convolutional neural networks for vibrational spectroscopic data analysis. *Analytica Chimica Acta, 954*, 22-31.

Al-Awadhi, M. A., & Deshmukh, R. R. (2021, 4-6 Dec. 2021). *Detection of Adulteration in Coconut Milk using Infrared Spectroscopy and Machine Learning*. Paper presented at the 2021 International Conference of Modern Trends in Information and Communication Technology Industry (MTICTI).

Azlin-Hashim, S., Siang, Q. L., Yusof, F., Zainol, M. K., & Mohd Yusof, H. (2019). Chemical composition and potential adulterants in coconut milk sold in Kuala Lumpur. *Malaysian Applied Biology, 48*(3), 27-34.

Basri, K. N., Laili, A. R., Tuhaime, N. A., Hussain, M. N., Bakar, J., Sharif, Z., Abdul Khir, M. F., & Zoolfakar, A. S. (2018). FT-NIR, MicroNIR and LED-MicroNIR for

detection of adulteration in palm oil via PLS and LDA. *Analytical Methods, 10*(34), 4143-4151.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(8), 1798-1828.

Benmouna, B., García-Mateos, G., Sabzi, S., Fernandez-Beltran, R., Parras-Burgos, D., & Molina-Martínez, J. M. (2022). Convolutional Neural Networks for Estimating the Ripening State of Fuji Apples Using Visible and Near-Infrared Spectroscopy. *Food and Bioprocess Technology, 15*(10), 2226-2236.

Büning-Pfaue, H. (2003). *Analysis of water in food by near infrared spectroscopy.* Paper presented at the Food Chemistry.

Chen, H., Tan, C., Lin, Z., & Li, H. (2019). Quantifying several adulterants of notoginseng powder by near-infrared spectroscopy and multivariate calibration. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 211*, 280-286.

Chong, I.-G., & Jun, C.-H. (2005). Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems, 78*(1), 103-112.

Chu, X., Huang, Y., Yun, Y.-H., & Bian, X. (2022). *Chemometric methods in analytical spectroscopy technology*: Springer.

Conzen, J. (2006). *Multivariate Calibration: A practical guide for developing methods in the quantitative analytical chemistry*.

Cui, C., & Fearn, T. (2018). Modern practical convolutional neural networks for multivariate regression: Applications to NIR calibration. *Chemometrics and Intelligent Laboratory Systems, 182*, 9-20.

Engel, J., Gerretzen, J., Szymanska, E., Jansen, J. J., Downey, G., Blanchet, L., & Buydens, L. M. C. (2013). Breaking with trends in pre-processing? *TrAC Trends in Analytical Chemistry, 50*, 96-106.

Gron, A. (2019). Hands-On Machine Learning with scikit learn keras&tensorflow . 2019: Sebastopol O'Reilly Media Google Scholar Google Scholar Digital Library ….

Gulli, A., & Pal, S. (2017). *Deep learning with Keras*: Packt Publishing Ltd.

Jiang, H., & Chen, Q. (2019). Determination of Adulteration Content in Extra Virgin Olive Oil Using FT-NIR Spectroscopy Combined with the BOSS–PLS Algorithm. *Molecules, 24*(11). Retrieved from doi:10.3390/molecules24112134

Jiang, H., & Lu, J. (2018). Using an optimal CC-PLSR-RBFNN model and NIR spectroscopy for the starch content determination in corn. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 196*, 131-140.

Jin, B., Zhang, C., Jia, L., Tang, Q., Gao, L., Zhao, G., & Qi, H. (2022). Identification of Rice Seed Varieties Based on Near-Infrared Hyperspectral Imaging Technology Combined with Deep Learning. *ACS Omega, 7*(6), 4735-4749.

Lakshanasomya, N., Danudol, A., & Ningnoi, T. (2011). Method performance study for total solids and total fat in coconut milk and products. *Journal of Food Composition and Analysis, 24*(4), 650-655.

Lan, Z., Zhang, Y., Zhang, Y., Liu, F., Ji, D., Cao, H., Wang, S., Lu, T., & Meng, J. (2021). Rapid evaluation on pharmacodynamics of Curcumae Rhizoma based on Micro-NIR and benchtop-NIR. *Journal of Pharmaceutical and Biomedical Analysis, 200*, 114074.

Li, Z., Song, J., Ma, Y., Yu, Y., He, X., Guo, Y., Dou, J., & Dong, H. (2023). Identification of aged-rice adulteration based on near-infrared spectroscopy combined with partial least squares regression and characteristic wavelength variables. *Food Chemistry: X, 17*, 100539.

Liu, Y., Zhou, S., Han, W., Li, C., Liu, W., Qiu, Z., & Chen, H. (2021). Detection of Adulteration in Infant Formula Based on Ensemble Convolutional Neural Network and Near-Infrared Spectroscopy. *Foods, 10*(4). Retrieved from doi:10.3390/foods10040785

Malvandi, A., Feng, H., & Kamruzzaman, M. (2022). Application of NIR spectroscopy and multivariate analysis for Non-destructive evaluation of apple moisture content during ultrasonic drying. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 269*, 120733.

Manley, M. (2014). Near-infrared spectroscopy and hyperspectral imaging: non-destructive analysis of biological materials. *Chemical Society Reviews, 43*(24), 8200-8214.

Nallan Chakravartula, S. S., Moscetti, R., Bedini, G., Nardella, M., & Massantini, R. (2022). Use of convolutional neural network (CNN) combined with FT-NIR

spectroscopy to predict food adulteration: A case study on coffee. *Food Control, 135*, 108816.

Osborne, B. G., Fearn, T., & Hindle, P. H. (1993). *Practical NIR spectroscopy with applications in food and beverage analysis*: Longman scientific and technical.

Palermo, G., Piraino, P., & Zucht, H.-D. (2009). Performance of PLS regression coefficients in selecting variables for each response of a multivariate PLS for omics-type data. *Advances and Applications in Bioinformatics and Chemistry, 2*(null), 57-70.

Passos, D., & Mishra, P. (2023). Deep Tutti Frutti: Exploring CNN architectures for dry matter prediction in fruit from multi-fruit near-infrared spectra. *Chemometrics and Intelligent Laboratory Systems, 243*, 105023.

Phetpan, K., & Sirisomboon, P. (2015). Evaluation of the moisture content of tapioca starch using near-infrared spectroscopy. *Journal of Innovative Optical Health Sciences, 8*(02), 1550014.

Said, M., Wahba, A., & Khalil, D. (2022). Semi-supervised deep learning framework for milk analysis using NIR spectrometers. *Chemometrics and Intelligent Laboratory Systems, 228*, 104619.

Sitorus, A., & Lapcharoensuk, R. (2023). A rapid method to predict type and adulteration of coconut milk by near-infrared spectroscopy combined with machine learning and chemometric tools. *Microchemical Journal, 195*, 109461.

Tansakul, A., & Chaisawang, P. (2006). Thermophysical properties of coconut milk. *Journal of Food Engineering, 73*(3), 276-280.

Wang, Z., Wu, Q., & Kamruzzaman, M. (2022). Portable NIR spectroscopy and PLS based variable selection for adulteration detection in quinoa flour. *Food Control, 138*, 108970.

Weng, S., Guo, B., Tang, P., Yin, X., Pan, F., Zhao, J., Huang, L., & Zhang, D. (2020). Rapid detection of adulteration of minced beef using Vis/NIR reflectance spectroscopy with multivariate methods. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 230*, 118005.

Williams, P. (2009). Influence of water on prediction of composition and quality factors: The aquaphotomics of low moisture agricultural materials. *Journal of Near Infrared Spectroscopy, 17*(6), 315-328.

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems, 58*(2), 109-130.

Workman Jr, J., & Weyer, L. (2007). *Practical guide to interpretive near-infrared spectroscopy*: CRC press.

Yang, J., Wang, J., Lu, G., Fei, S., Yan, T., Zhang, C., Lu, X., Yu, Z., Li, W., & Tang, X. (2021). TeaNet: Deep learning on Near-Infrared Spectroscopy (NIR) data for the assurance of tea quality. *Computers and Electronics in Agriculture, 190*, 106431.

CHAPTER 6 – CASE STUDY 4
DISCRIMINATION MODEL OF GEOGRAPHICAL AREA FROM COCONUT MILK BY NIR SPECTROSCOPY: EXPLORATION IN TANDEM WITH CLASSICAL CHEMOMETRICS, MACHINE LEARNING, AND DEEP LEARNING[5]

6.1 Highlights

1. FT-NIR and Micro-NIR are utilized to identify the geographical areas of coconut milk.
2. Classical to modern chemometric classifiers were tested for discriminating geographical areas of coconut milk.
3. Methods for discovering features important from all classifiers were also explored.
4. Classifiers from modern chemometrics with proper data preprocessing achieved good performance classification.
5. NIRs coupled with classical to modern chemometric classifiers provide a promising classification tool for discriminating geographical areas of coconut for coconut milk production.

6.2 Abstract

This work proposes exploring the discrimination model by NIR spectroscopy (FT-NIR and Micro-NIR) for geographical source areas of coconut milk in tandem with the classical to modern chemometrics classifier. The discrimination model was developed using qualitative chemometrics techniques from classic (PCA, PLS-DA, LDA) to modern, including classifiers from machine learning (SVM, KNN, ANN) and deep learning (S-CNN, S-AlexNET, ResNET). Three sources as geographical areas of coconut milk originally from Thailand were used, including the south region (Chumphon Province), middle region (Samut Songkhram Province), and east region

[5]This chapter constituted the publication article: Sitorus, A., & Lapcharoensuk, R. (2024). Discrimination model of geographical area from coconut milk by NIR spectroscopy: Exploration in tandem with classical chemometrics, machine learning, and deep learning. **Microchemical Journal**, Submited on 31 March 2024.

(Chonburi Province). Our findings showed that a classifier from the machine learning (SVM) and deep learning (ResNET) groups could yield the optimal performance for discriminating the geographical source area of coconut milk, with an accuracy of 99.1% for the training and 100% for the testing using FT-NIR. Furthermore, when using Micro-NIR, the classifier from group classical (LDA), machine learning (SVM, KNN), and deep learning (ResNET) delivered the highest accuracy of 99.5% for the training and 100% for the testing. The performance discrimination models above were excellent when classified based on the kappa coefficient. This study concluded that both FT-NIR and Micro-NIR supported by classical to modern chemometric classifiers could be used to evaluate the geographical area source from coconut milk. Also, the method in this study includes a strategy for discovering feature important NIR spectra for interpretability purposes, thereby facilitating the qualitative interpretation of results for all types of classifiers.

Keywords: advance chemometrics; classical chemometrics; coconut milk; FT-NIR; Micro-NIR.

6.3 Introduction

Thailand has a significant role in the international coconut milk industry, both as a producer and consumer of the commodity. Coconut milk is a vital ingredient in Thai cuisine, and it is used in various meals such as curries, soups, desserts, and drinks. The worldwide coconut milk market is projected to witness significant growth between 2022 and 2030 due to the increasing demand for plant-based food and the growing popularity of Thai cuisine (Suksangpanomrung *et al.*, 2024). Therefore, the development of agricultural land that cultivates coconuts as coconut milk continues to be intensively carried out in various provinces in Thailand.

Coconut milk is derived from ripe coconuts (brown-colored husks) and shredded flesh lalu diperas. According CODEX-STAN-240 (2003) from Codex Alimentarius Commission Internastion Food Standard classifies liquid fresh coconut milk into light coconut milk (6.6–12.6% total solid, 5% fat, 93.4% moisture), coconut milk (12.7–25.3% total solid, 10% fat, 87.3.4% moisture), coconut cream (25.4–37.3% total solid, 20% fat, 74.6% moisture), and coconut cream concentrate (>37.4% total solid, 29% fat, 62.6% moisture). These categories guarantee that commercially

manufactured coconut milk meets precise criteria, highlighting the significance of strict quality control throughout the entire manufacturing process, from the first receipt of raw coconut materials to the finished product.

Food traceability is one of the main concerns of regulatory agencies and customers worldwide, and it may be considered a crucial characteristic of future food items. The potential for raw coconut materials to provide different coconut milk contents even from the same variety can be caused by many factors, such as planting environment, treatment, and cultivation methods (Mat *et al.*, 2022; Tulashie *et al.*, 2022). This is possible due to the spread of coconut palm plantations in Thailand in several provinces, which the community's culture will influence when cultivating them. This makes products with high-quality coconut milk vulnerable to counterfeiting with different quality products. Consequently, the risk of coconut milk fraud exists, where coconut milk is misrepresented as originating from more desirable regions renowned for superior nutritional properties. Identifying geographical areas of coconut milk is a topic of concern, as it has other effects on food, especially quality and authenticity. This will protect against potential fraud from the beginning. Hence, developing methods to identify geographical areas of agricultural products like coconut milk is essential to ensure their superior quality.

Near-infrared (NIR) spectroscopy emerges as a promising tool for non-destructiveness and rapid analysis of agriculture and food. NIR spectroscopy can be efficiently applied to quantitative and qualitative analyses. Several recent five-year studies report using NIRs for qualitative analysis in classifying the geographical origin of agricultural and food products, including durum wheat (De Girolamo *et al.*, 2019), sea cucumbers (Sun *et al.*, 2021), almonds (Arndt *et al.*, 2021), rice (Srinuttrakul *et al.*, 2021; Wu *et al.*, 2023), and grain maize (Schütz *et al.*, 2022). This shows that NIR spectroscopy effectively assesses the geographical origin of agricultural products. Therefore, NIR spectroscopy has the potential to predict the geographical area of coconut milk intended for coconut milk production.

Nonetheless, the agricultural products sample conditions are often challenging and have high levels of non-uniformity, significantly deteriorating the measurement data quality, particularly coconut milk. Thus, NIR should be supported by a chemometric approach, which can no longer be classical but must also be prepared

with advanced chemometrics. This is because classical chemometrics is less adaptable to capturing cases of nonlinearity in NIR data (Rocha *et al.*, 2020; Zareef *et al.*, 2020). Meanwhile, the current cases based on NIR spectroscopy involve a lot of nonlinearity in their dataset. Based on this, several chemometricians attempted to develop nonlinear classification models based on NIR spectral data.

To date, several chemometric techniques have been applied to qualitative studies as classifiers in geographical origin for agriculture products, including classical chemometrics, including principal component analysis (PCA) (Wu *et al.*, 2023), partial least squares discriminant analysis (PLS-DA) (Wu *et al.*, 2023), and linear discriminant analysis (LDA) (De Girolamo *et al.*, 2019); machine learning, including random forest, gradient boosting decision tree, light gradient boosting machine (Sun *et al.*, 2021), and support vector machines (SVM) (Schütz *et al.*, 2022). Moreover, deep learning, as a part of advanced machine learning approaches based on one dimensional (1D) NIR spectroscopy, has also not been widely explored in the case of geographical origin classification for agricultural products. So far, the most well-known report is the application of deep learning with convolutional neural network (CNN) classifier for the classification of geographical origin based on hyperspectral imaging for various products such as traditional medicine Radix Glycyrrhizae (Yan *et al.*, 2020), Panax notoginseng (Dong *et al.*, 2020), and Tetrastigma hemsleyanum (Zhou *et al.*, 2020). Although many algorithms have been developed from classical to modern, which are robust and reliable for several cases, the main drawback is that there is no one superior algorithm for all classification cases. In particular, a series of steps to develop complex and rigorous model discrimination is often required, thereby limiting the wider applicability of these methods.

During the model development, spectral preprocessing is an essential first step that plays a critical role in the algorithms' performance. Preprocessing techniques such as multiplicative scattering correction and derivative methods can address background noise, baseline drift, and light scattering issues. Several studies have demonstrated that different spectral preprocessing and modeling approaches yield distinct classification results. Therefore, the pretreatment and chemometric methods must be optimized for the spectra of interest to prevent interference from invalid information and extract as much practical information as possible. This is especially

necessary for classical chemometrics and machine learning classifiers, so several studies have developed automatic preprocessing for these algorithms. Nevertheless, according to Cui and Fearn (2018), spectral preprocessing may not be required anymore because deep learning can deal with it. However, some studies still use preprocessing when using deep learning algorithms (Acquarelli *et al.*, 2017; Zhang *et al.*, 2019). At least, Standard normal variate preprocessing is still needed in the beginning to standardize the data with the same mean and standard deviation for all sample spectra.

To the best of our knowledge, there is no NIR spectroscopy research explorer to trace the coconut milk samples' geographical area. Therefore, the objective of this study was to explore the potential of classifier algorithms for NIR spectroscopy (FT-NIR and Micro-NIR) data from classical to modern chemometrics, focusing specifically on geographical area classification from coconut milk in Thailand. This experiment compares the performance of classic chemometrics, including principal component analysis (PCA), partial least squares-discriminant analysis (PLS-DA), linear discriminant analysis (LDA), machine learning including support vector machine (SVM), k-nearest neighbor (KNN), artificial neural networks (ANN), and deep learning, including simple convolutional neural networks (S-CNN), S-AlexNET, residual networks (ResNET) using confusion metrics and support with the kappa coefficient. This valuable information can be used to ensure that chemometric support from classical to modern approaches is currently available, with many options and promising performance for various purposes, especially in qualitative cases. Furthermore, the case study presented in this paper, related to the geographical area classification of coconuts used in coconut milk production, can be traced quickly and non-destructively using NIR spectroscopy supported by a robust chemometric approach. This enables the identification and decision to accept or reject harvested coconut batches that do not meet the required quality standards to be completed more quickly.

6.4 Materials and Methods

6.4.1 Source of Coconut Sample Collection

Forty-five samples of coconut fruit ready to extract into coconut milk from the January 2024 harvest season, representative of different geographical areas of

Thailand (south, middle, and east), were collected from local coconut farms located in three provinces, including Chumphon (CHP), Samut Songkhram (SSK), and Chonburi (CHB). Figure 6.1 provides details about the source of the sample's geographical location. The samples were transported to the Department of Agricultural Engineering, School of Engineering, King Mongkut's Institute of Technology, Ladkrabang, Thailand. After the samples were collected, a small and medium enterprise (SME) unit in the field of coconut milk processing carried out the extraction process at the local market in Lad Krabang, Thailand. This method was selected so that the extraction process was carried out naturally, following the standards of the coconut milk processing industry, without adding water. After that, the coconut milk samples were taken to the laboratory and preserved in glass bottles at room temperature (±25°C).

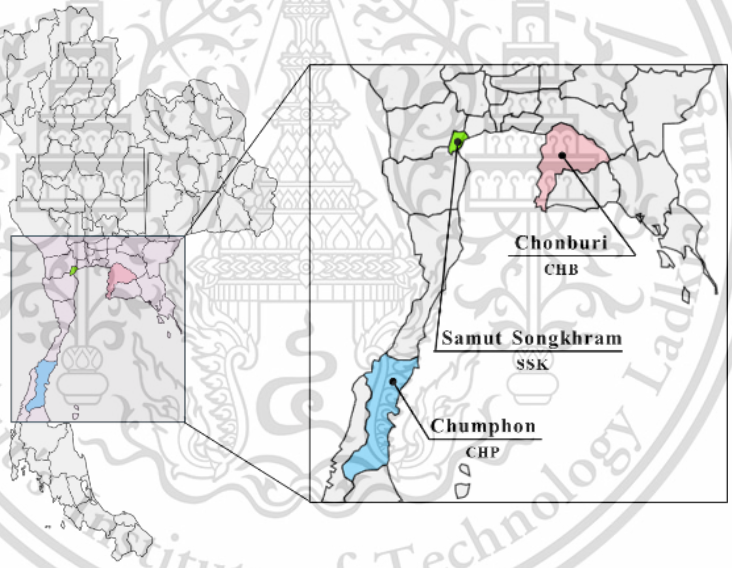


Figure 6.1. Geographical of planted area of collected coconut sample.

Before scanning, coconut milk samples were filtered through 100-mesh cloth filters. For each geographical area, 90 samples of coconut milk were prepared, and a total of 270 samples of coconut milk were scanned. With this condition, the total data that will be analyzed in this study is more than suggested by Manley (2014), who mentions that total NIR spectra samples for developing discrimination models must at least be more than 100 spectra to get a reliable model, and this study has exceeded that. The number of training (4/5) and testing (1/5) data used in this study

was 216:54 (separated using the Kennard stone splitting method with a random state of 62).

6.4.2 NIR Spectral Data Acquisition

The full-wave NIR spectral was measured with benchtop FT-NIR spectrometer (Bruker Ltd., Ettlingen Germany) on 12,500–4000 cm^{-1} (800–2500 nm). Each sample of coconut milk was taken from a glass beaker using a micropipette of about 1 mL. After that, a aluminum reflector was also put in a test glass vial and placed on the FT-NIR spectrometer. One average spectrum was obtained by an average of 32 scans at resolution of 8 cm^{-1}. Secondly, another NIR spectrum was measured with a portable Micro-NIR (MicroNIR OnSite-W, VIAVI Solutions Inc., Chandler, United States) on 908–1676 nm (11,013–5967 cm^{-1}) with a resolution of 6.2 nm. The integration time and scan count used in this study were 10 ms, 100, respectively. After that dark current scanning was conducted in the air, and reference scanning was performed on the teflon material as a reference. Scan results were recorded in absorption mode (log 1/R) for each sample.

6.4.3 Chemometric and Data Analysis

6.4.3.1 Preprocessing Method

The preprocessing employed in this study includes standard normal variate (SNV), multiplicative scattering correction (MSC), baseline second-order (BSO), first-order derivative (FOD), and second-order derivative (SOD) from Savitzky-Golay. Mathematical preprocessing was necessary to reduce systematic noise, such as baseline variation and light scattering, and to enhance the contribution of the chemical composition.

Standard normal variable (SNV) preprocessing is mainly used to eliminate the influence of solid particle size, surface scattering, and optical path change on the diffuse reflectance spectrum. SNV preprocessing standardizes each spectrum using data from the self-spectrum. Therefore, each spectrum's average value and standard deviation obtained from preprocessing using SNV are zero and one, respectively. SNV preprocessing retains the original spectrum shape and creates an artificial absorbance scale with negative values (Chu *et al.*, 2022).

Multiplicative scattering correction (MSC) works by correcting scattering, maintaining the original spectral shape and the same spectral scale. The idea behind MSC is that the two effects, amplification (multiplicative, scattering) and offset (additive, chemical), should be removed from the data table to avoid dominating the information (signal) in the data. Two correction coefficients, intercept and slope, are calculated from a reference from the average spectrum in the data set and used in these computations to determine MSC preprocessing (Xu *et al.*, 2008).

Baseline second order (BSO) or detrending is a correction that is applied to remove nonlinear trends in spectroscopic data by calculating a baseline function as a least squares fit of a polynomial to each individual spectrum (Barnes *et al.*, 1989). BSO corrections are applied to an individual spectrum, which differs from many other data treatments operating at a specific wavelength across the entire spectral set. The order of the polynomial used in the BSO correction determines the removed baseline effects. In this study, a third-order polynomial was used to remove baseline offset, slope, and curvature. Furthermore, the BSO preprocessing does not change the shape of the data, which commonly occurs with derivative-based corrections. BSO is applied to correct baseline effects in spectra to remove nonchemical effects and create robust calibration models. BSO may also help resolve overlapped bands, which can provide a better understanding of the data, emphasizing small spectral variations that are not evident in the raw data.

The first-order derivative (FOD) and second-order derivative (SOD) from Savitzky-Golay are very effective methods for removing such baseline offsets. This method is the most popular derivatives method for preprocessing near-infrared spectral data. The derivatives by Savitzky-Golay are based on a localized linear regression of several neighboring points to determine the best-fit polynomial. This polynomial can be mathematically differentiated and evaluated at the spectra values coincident with wavelength collection points. A convolution with a set of derived coefficients performs a mathematical equivalent of the regression and differentiation procedure (Delwiche and Reeves, 2010). SOD is an effective method for removing the baseline offset and slope from a spectrum. The SOD can help resolve nearby peaks and sharpen spectral features. Peaks in raw spectra change sign and turn to negative peaks on either side in the SOD.

6.4.3.2 Classical Chemometric Classifier

Methods for developing discrimination models using classical chemometrics classifiers, including principal component analysis (PCA), partial least squares-discriminant analysis (PLS-DA), and linear discriminant analysis (LDA), were employed in this study. Each classifier has main hyperparameters that must be optimized, including the principal component (PC) for the PCA classifier, the latent variable (LV) for PLS-DA, and the linear discriminant (LD) component for LDA. Optimization of the hyperparameters from all classical chemometric classifiers utilizing the GridSearchCV command with a tuning range is presented in Table 6.1 until each model yielded the lowest error average of 5-fold cross-validation of the training set. A brief description of each classical chemometrics classifier is defined as follows.

Principal component analysis (PCA) is the commonly used unsupervised qualitative multivariate data analysis method for reducing dimensionality for many purposes (Brereton, 2022). By plotting the PCA scores, samples with similar spectra signatures tend to aggregate together or lie close to one another. In this study, PCA was performed on the covariance matrix to identify by examining the grouping of the three samples (CHP, SSK, CHB) according to their spectral variations.

Table 6.1. Hyperparameters tuning for classical chemometrics and machine learning classifier.

Classifier	Hyperparameter	Tuning values
PCA	PC components (PC)	1–20
PLS-DA	LV components (LV)	1–20
LDA	LD components (LD)	1–2
SVM	Penalty (C)	1, 10, 100
	Kernel (K)	Linear, Poly, RBF, Sigmoid
	Degree (D)	2, 3, 4
	Gamma (G)	scale, auto
KNN	nn-components	1 – 20
ANN	Hidden layer size (H)	(8), (16), (32), (64), (128), (256), (512), (1024), (8, 8), (16, 16), (32, 32), (64, 64), (128, 128), (256, 256), (512, 512), (1024, 1024), (8, 8, 8), (16, 16, 16), (32, 32, 32), (64, 64, 64), (128, 128, 128), (256, 256, 256), (512, 512, 512), (1024, 1024, 1024)

Partial least square discriminant analysis (PLS-DA) is a supervised classification analysis method whose calculations are similar to the partial least square regression (PLSR) method. The basic that distinguishes it from PLSR is related to constructing binary classification for the response. The first step in PLS-DA modeling is recoding the categorical (ordinal or nominal) variables into continuous or numerical variables. After that, the dependent variable or response is converted into a binary classification consisting of only two integers to indicate "out-members" and "in-members", respectively (Brereton and Lloyd, 2014).

Linear discriminant analysis (LDA) is a supervised classification analysis method commonly using dimensionality reduction techniques to solve more than two-class classification problems. The criterion to determine these vectors is the maximization of the ratio of between-class variability and within-class variability in the training set (Silva *et al.*, 2013). In LDA, the eigenvector is generated from the matrix, which is constructed by maximizing the discriminant of the separation matrix between data groups.

6.4.3.3 Machine Learning Classifier

For the machine learning algorithm, this study used classifiers from support vector machine (SVM), k-nearest neighbor (KNN), and artificial neural networks (ANN). The SVM classifier has main parameters that must be optimized, including penalty factor (C), kernel (K), degree (D), and gamma (G). Meanwhile, in KNN and ANN, the classifier has main parameters that must be optimized, namely nearest-neighbors (nn) and hidden layer size (H), respectively. Hyperparameters optimization of the machine learning classifier using the GridSearchCV command with tuning range are presented in Table 6.1 until each model yielded the lowest error average of 5-fold cross-validation of the training set. A short description of each machine learning classifier is defined as follows.

Support vector machines (SVM) are supervised machine learning models credited to possess powerful regression tools that have found applications in numerous prediction problems in various fields. In SVM, the structural risk and statistical learning theory minimization principles are employed to map the initial training samples into a higher dimensional feature space through nonlinear kernel functions, and the optimal solution is obtained by converting the problem to linear

from nonlinear. SVM firstly maps the data into a high-dimensional feature space to transform a linearly separable problem and then classifies these data by hyperplane. The optimal hyperplane separating the data can be obtained to solve the following optimization problem using minimizing Equation (6.1) and subject to Equation (6.2). The number of coefficient weights (*w*) depends on the number of predictors available. After getting the coefficient weight (w_i) from the data set, the hyperplane can be formulated, and the support vector (SV), including SV_1 and SV_2, using Equation (6.3) (Brereton and Lloyd, 2010).

$$\text{Func. obj.} \rightarrow \text{F(obj)} = \min\left\{\frac{1}{2}\sum_{i}^{n}\left(w_i^{\,2}\right)\right\} \tag{6.1}$$

$$\text{Func. bound.} \rightarrow y_i\left[\left(w_i \bullet x_i\right) + b\right] \geq 1 \tag{6.2}$$

$$f(x)\begin{cases} SV_1 \rightarrow -1 = \left[b + \sum_{i}^{n}\left(w_i \bullet x_i\right)\right] \\ \text{hyperplane} \rightarrow 0 = \left[b + \sum_{i}^{n}\left(w_i \bullet x_i\right)\right] \\ SV_2 \rightarrow +1 = \left[b + \sum_{i}^{n}\left(w_i \bullet x_i\right)\right] \end{cases} \tag{6.3}$$

K-nearest-neighbors (KNN) is one of the simplest methods in data mining classification techniques. The idea of the KNN algorithm is if most of the K most similar samples in the feature space belong to a certain category, then this sample also falls into this category. In classification decision-making, this method only determines the class of the sample to be classified according to the class of one or several recent samples. KNN algorithm is run by determining the distance between the new data observation from data set prediction (Xp) to all data points in the training data set (Xt) using the Euclidean distance (*D*) (Equation (6.4)). Next, sort the distance data (*n*) in ascending order. Identify the *k* closet neighbors optimally. Determine the class of the new observation based on the group majority of the KNN (Ni *et al.*, 2009).

$$D\left(\text{Xt}, \text{Xp}\right) = \sqrt{\left(xt_1 - xp_1\right)^2 + \left(xt_2 - xp_2\right)^2} \tag{6.4}$$

An artificial neural network (ANN) is a computational model that imitates the structure and function of a human neural network for estimating or approximating functions. ANN learns arbitrary mappings from input data to outputs to generate computational models and identify highly nonlinear and complex connections with

high prediction accuracy. The network layers of the artificial neural network are divided into the input layer, hidden layer, and output layer. Determining the number of neurons in the hidden layer generally relies on empirical methods, and the appropriate number of layers is determined by testing the network. To describe the neural net operation, a set of neurons from NIR spectra are utilized as input, and the classification of the area of origin of coconut milk as a response. Equation (6.5) calculates net input by connecting all future input to the output over a weighted interconnection network. The function of net input is given in Equation (6.6). This function of net input uses the sigmoid activation function, which is popularly used in the ANN method (Zareef *et al.*, 2020). In addition, the weight (*w*) and bias (*b*) value as a network link can be used temporarily. After that, to get the proper weight (*w*) and bias (*b*) values, they must be updated until the difference between prediction and references (sum square error (SSE)) is as small as possible using Equation (6.7). This is known as backpropagation.

$$a_j = \sum_j \left(w_{i,j} \times x_i \right) + b \tag{6.5}$$

$$y_j = F(\mathrm{a}_j) = \frac{1}{1 + e^{-a_j}} \tag{6.6}$$

$$\mathrm{SSE} = \sum \left(y - y_p \right)^2 \tag{6.7}$$

6.4.3.4 Deep Learning Classifier

This study used three types of deep learning classifier architecture: simple convolutional neural networks (S-CNN), S-AlexNET, and residual networks (ResNET). The main architecture of deep learning is input, feature extraction joint with learning, and output. The difference between the three deep learning classifiers lies in addressing joint feature extraction with learning simultaneously to provide the most accurate classification prediction output (Géron, 2019). The first layer from all classifiers is the input of the convolutional, where this 1D convolution layer functions in extracting features in NIR spectroscopy. This layer has three input dimensions; the first dimension (none) is a sample, many rows of data, or the amount of input data used in a batch (batch size), which has not been determined during the model compilation process. The second input dimension represents the feature used in prediction. This dimension feeds input features as neurons used in the prediction are 1102 neurons for data from FT-NIR and 125 neurons for data from Micro-NIR. In the

end, each deep learning classifier has densely connected layers (4 for the S-CNN, 3 for ResNET, and 3 for the S-AlexNET architecture). The final density is a prediction target value consisting of 3 values with a special activation function for classification cases.

All types of network architecture model deep-learning training procedures in this study were performed using Adam optimizer, number of batches of 108, epochs per running of 1000 in 10 times with restore best weights as an early stopping epoch, and learning rate of 4.22×10^{-3}, respectively. The loss function in this study uses sparse categorical cross-entropy to convert the last dense result using SOFTMAX activation function. Sparse categorical cross-entropy is a loss function commonly used for target labels and integers representing classes. Unlike categorical cross-entropy, it eliminates the need to convert labels into one-hot encoding. This makes it suitable for cases where the number of classes is high, and converting labels to one-hot encoding would be memory-intensive. Sparse categorical cross-entropy calculates the cross-entropy loss between the predicted probability distribution and the true integer labels, helping to train convolutional neural networks to classify input data into the output classes. A brief illustration of each deep learning classifier is described as follows.

The simple convolutional neural networks (S-CNN) architecture used in this study is presented in Table 6.2. This architecture only uses a one-dimensional convolutional layer before being flattened and fed to the neural network (Cui and Fearn, 2018). The filter, kernel size, and stride used are 1, 5, and 1, respectively. The activation function used is ELU with SAME padding. For FT-NIR, the total of parameters, trainable and non-trainable, are 73,257 neurons, 73,257 neurons, and 0 neurons, respectively. The total of parameters, trainable and non-trainable, are 10,729 neurons, 10,729 neurons, and 0 neurons, respectively, for Micro-NIR.

Table 6.2. Parameter setting of S-CNN classifier.

Layer	Output shape (FT-NIR/Micro-NIR)	Parameter (FT-NIR/Micro-NIR)
reshape	(None, 1102/125, 1)	0
conv1d	(None, 1102/125, 1)	6
flatten	(None, 1102/125)	0
dense	(None, 64)	70,592/8064

dense_1	(None, 32)	2080
dense_2	(None, 16)	528
dense_3	(None, 3)	51

The S-AlexNET architecture employed in this study is shown in Table 6.3. The S-AlexNET classifier consists of several layers (Hosseinpour-Zarnaq *et al.*, 2023). The input dimensions are set according to the NIR spectra input dimensions. The L2 regularizer parameter is determined to control the model's complexity. Weight initialization is done with the HE-NORMAL initializer. The model architecture consists of a reshape layer as input and several Conv1D layers, followed by batch normalization, LEAKY-RELU activation function, and max pooling 1D. The dropout method is used to reduce overfitting. Weights for each layer are set using the same initialization and regularization to ensure consistency and control over the model's complexity. This model has a deep architecture with many convolutional and dense layers to obtain strong feature representations from complex spectral data. The kernel size, stride and padding used are 3, 1, and SAME, respectively. Finally, for FT-NIR, the total number of parameters, trainable and non-trainable, is 19,539,203 neurons, 19,535,875 neurons, and 3328 neurons, respectively. For Micro-NIR, the total number of parameters, trainable and non-trainable, is 3,548,419 neurons, 3,545,091 neurons, and 3328 neurons, respectively.

Table 6.3. Parameter setting of S-AlexNET classifier.

Layer	Output shape (FT-NIR/Micro-NIR)	Parameter (FT-NIR/Micro-NIR)
reshape	(None, 1102/125, 1)	0
conv1d	(None, 1100/123, 128)	512
batch_normalization	(None, 1100/123, 128)	512
activation	(None, 1100/123, 128)	0
max_pooling1d	(None, 549/61, 128)	0
conv1d_1	(None, 549/61, 256)	98,560
batch_normalization_1	(None, 549/61, 256)	1024
activation_1	(None, 549/61, 256)	0
max_pooling1d_1	(None, 274/30, 256)	0
conv1d_2	(None, 274/30, 512)	393,728
batch_normalization_2	(None, 274/30, 512)	2048
activation_2	(None, 274/30, 512)	0
conv1d_3	(None, 274/30, 512)	786,944

batch_normalization_3	(None, 274/30, 512)	2048
activation_3	(None, 274/30, 512)	0
conv1d_4	(None, 274/30, 256)	393,472
batch_normalization_4	(None, 274/30, 256)	1024
activation_4	(None, 274/30, 256)	0
max_pooling1d_2	(None, 274/30, 256)	0
flatten	(None, 69632/7168)	0
dense	(None, 256)	1,7826,048/1,835,264
dropout	(None, 256)	0
dense_1	(None, 128)	32896
dropout_2	(None, 128)	0
dense_2	(None, 3)	387

The residual networks (ResNET) architecture utilized in this study is shown Table 6.4. The ResNET architecture classifier consists of one-dimensional convolutional layers followed by residual blocks, each comprising two convolutional layers with a shortcut connection for improved learning. After the residual blocks, dimensionality reduction is performed using max pooling one-dimensional layers, followed by additional residual blocks. Finally, global dimensionality reduction is done using global average pooling one-dimensional layers and connected to several Dense layers for the final classification (Zou *et al.*, 2021). The model utilizes the RELU activation function, HE_NORMAL weight initialization, L2 regularization with a factor of 0.001, and SOFTMAX activation for three output classes. The kernel size, stride and padding used are 3, 1, and SAME, respectively. From these parameters, for FT-NIR, the total parameters, trainable and non-trainable, are 582,211 neurons, 579,907 neurons, and 2304 neurons, respectively. The total of parameters, trainable and non-trainable, are 582,211 neurons, 579,907 neurons, and 2304 neurons, respectively, for Micro-NIR.

Table 6.4. Parameter setting of ResNET classifier.

Layer	Output shape (FT-NIR/Micro-NIR)	Parameter (FT-NIR/Micro-NIR)
reshape_input	(None, 1102/125)	0
reshape_input	(None, 1102/125, 1)	0
conv1d	(None, 1102/125, 64)	256
conv1d_1	(None, 1102/125, 64)	12,352
batch_normalization	(None, 1102/125, 64)	256
activation	(None, 1102/125, 64)	0

conv1d_2	(None, 1102/125, 64)	12,352
batch_normalization_1	(None, 1102/125, 64)	256
add	(None, 1102/125, 64)	0
activation_1	(None, 1102/125, 64)	0
max_pooling1d	(None, 1100/123, 64)	0
conv1d_3	(None, 1100/123, 128)	24,704
batch_normalization_2	(None, 1100/123, 128)	512
activation_2	(None, 1100/123, 128)	0
conv1d_4	(None, 1100/123, 128)	49,280
batch_normalization_3	(None, 1100/123, 128)	512
conv1d_5	(None, 1100/123, 128)	8320
add_1	(None, 1100/123, 128)	0
activation_3	(None, 1100/123, 128)	0
conv1d_6	(None, 1100/123, 128)	49,280
batch_normalization_4	(None, 1100/123, 128)	512
activation_4	(None, 1100/123, 128)	0
conv1d_7	(None, 1100/123, 128)	49,280
batch_normalization_5	(None, 1100/123, 128)	512
add_2	(None, 1100/123, 128)	0
activation_5	(None, 1100/123, 128)	0
max_pooling1d_1	(None, 1098/121, 128)	0
conv1d_8	(None, 1098/121, 256)	98,560
batch_normalization_6	(None, 1098/121, 256)	1024
activation_6	(None, 1098/121, 256)	0
conv1d_9	(None, 1098/121, 256)	196,864
batch_normalization_7	(None, 1098/121, 256)	1024
conv1d_10	(None, 1098/121, 256)	33,024
add_3	(None, 1098/121, 256)	0
activation_7	(None, 1098/121, 256)	0
global_average_pooling1d	(None, 256)	0
dense	(None, 128)	32,896
dense_1	(None, 64)	8256
dense_2	(None, 32)	2080
dense_3	(None, 3)	99

6.4.3.5 Feature Importance Extraction

In the classical chemometrics classifier (PCA, PLS-DA, LDA), feature importance can be represented using an eigenvalue matrix. PCA classifiers can use principle component (PC) loading, PLS-DA classifiers can utilize latent variable (LV) loading,

and LDA classifiers can employ linear discriminant (LD) loading (Chu *et al.*, 2022; Ozaki *et al.*, 2021).

In general, machine learning algorithms intrinsically do not provide a matrix that can be used as a feature of importance, especially algorithms that perform non-linearly, including SVM, KNN, and ANN. This is one of the disadvantages of a machine learning algorithm when it is used to carry out spectroscopy-based data processing, which requires information regarding the main features that can be used as fingerprints for the sample. Therefore, this study approaches it by analyzing each feature's contribution to the classification decision by observing the change in discrimination model performance when one feature is removed (Képeš *et al.*, 2022). This method is known as feature ablation (AiFI), defined by Equation (6.8). This method is quite similar to that reported by Breiman (2001) for extracting variable importance when using a random forest algorithm. This method calculates any difference between the average cross-validation score of all features (Bs_i) and the average cross-validation score if one feature is removed (As_i). A positive AiFI value indicates that removing the feature can reduce model performance, and a negative AiFI value indicates that removing the feature can improve model performance or, in other words, that the feature is suspected of noise. This method was chosen because it is general and can be used for all machine learning algorithms.

$$\text{AiFI} = Bs_i - As_i \tag{6.8}$$

For deep learning based on neural networks, the importance of features can be investigated from the weights used in each layer based on Equation (6.9), which was introduced by Cui and Fearn (2018). However, because it consists of many layers, it results in a pile of weights and overlapping. From that, this study modifies it to make it simpler by carrying out an average operation (Equation (6.10)) of each weight before carrying out a derivatives operation based on finite difference approximation using Equation (6.11).

$$f(x) = \left[\sum_{i=1}^{N}\left(w_{DL-i}x_i\right)\right] + b \tag{6.9}$$

$$w_{DL-i} = \frac{\sum_{i=1}^{N} w_{DL-i}}{N} \tag{6.10}$$

$$w_{DL-i} = \frac{f\left(x_1 + \varepsilon, \ldots x_i + \varepsilon, \ldots, x_n + \varepsilon\right) - f(x_1, \ldots x_i, \ldots, x_n)}{\varepsilon} \tag{6.11}$$

Where, $w_{DL\text{-}i}$ is weight to-*i*, *N* is the number of samples, x_1 to x_n is the absorbance of feature NIR spectra, *b* is the intercept, and ε is perturbation coefficient (10^{-6}).

6.4.3.6 Discrimination Model Evaluation

Three of geographical area coconut milk samples will be investigated qualitatively for classification, including CHP, SSK, and CHB. The confusion matrix will be used to evaluate the classification model. In the confusion matrix, each prediction result will be categorized into false positive (FP), false negative (FN), true positive (TP), and false negative (FN). After that, the parameters like precision (Pr_i), recall (Rc_i), and F1-score (Fs_i) will be calculated from the confusion matrix by Equation (6.12) to Equation (6.14). Next, The discrimination model performance was defined by classification accuracy (AC), weighted average of precision (Pr_{wa}), weighted average of recall (Rc_{wa}), and weighted average of F1-score (Fs_{wa}) calculated by Equation (6.15) to and Equation (6.18) with weighted each parameter (w_i) calculated by Equation (6.19). The ideal model is expected to exhibit a high Pr_{wa}, Rc_{wa}, Fs_{wa}, and Ac value. Finally, Kappa coefficient (Kc) can be used to measure the classification effect using Equation (6.20) with compare between accuracy to expected accuracy of model (E_{Ac}) (Lu *et al.*, 2020). However, according to Chu et al. (2022), Kc as a performance evaluation for classifier can be divided into five grades to represent the consistency of different classes including extremely low consistency under 0.2, general consistency from 0.2 to 0.4, moderate consistency from 0.4 to 0.6, high consistency from 0.6 to 0.8, almost complete consistency from 0.8 to 1.0.

$$Pr_i = \frac{TP}{TP + FP} \tag{6.12}$$

$$Rc_i = \frac{TP}{TP + FN} \tag{6.13}$$

$$Fs_i = \frac{2 \times Pr_i \times Rc_i}{Pr_i + Rc_i} \tag{6.14}$$

$$Ac = \frac{TP + TN}{TP + TN + FP + FN} \tag{6.15}$$

$$Pr_{wa} = \sum_{i=1}^{n} \left(w_i \times Pr_i \right) \tag{6.16}$$

$$Rc_{wa} = \sum_{i=1}^{n} \left(w_i \times Rc_i \right) \tag{6.17}$$

$$Fs_{wa} = \sum_{i=1}^{n} \left(w_i \times Fs_i \right) \quad (6.18)$$

$$w_i = \frac{\text{No. of samples in class-}i}{\text{Total number of samples}} \quad (6.19)$$

$$Kc = \frac{Ac - E_{Ac}}{1 - E_{Ac}} \quad (6.20)$$

In this study, all of the classifier algorithm were running on JupyterLab interface using open source platform Python version 6.5.4, Keras library version 2.13.1 with TensorFlow version 2.13.0 backend (Abadi *et al.*, 2016; Gulli and Pal, 2017). The CPU is Intel (R) core (TM) i9-13900H CPU @ 2.60 Ghz, and the graphics card is NVIDIA Geforce RTX 4060.

6.5 Results and Discussions

6.5.1 NIR Spectral Investigation

The coconut milk FT-NIR spectra were roughly compared visually to detect differences between the region-specific sample groups in the range 12,500–4000 cm^{-1} (800–2500 nm). All recorded coconut milk as raw FT-NIR spectra are shown in Figure 6.2a. The mean FT-NIR spectra of each sample group after preprocessing are shown in Figure 6.2b to Figure 6.2f to visualize the region-specific differences. On a visual basis, the overall shape of both the raw FT-NIR spectra and the preprocessed FT-NIR spectra appears to be comparable regardless of the geographical sample region. It can also be seen that the FT-NIR spectrum region above 10,000 cm^{-1} shows poor absorption, where this region is second-order overtone or higher than the vibration of bonds. The only small peak absorption band located above 10,000 cm^{-1} is at approximately 10,283 cm^{-1}, which corresponds to the second-order overtone vibration of the O-H stretching band. Other FT-NIR spectra show absorption peaks at approximately 8270 cm^{-1} and 6881 cm^{-1}, slightly shifting from 8264 cm^{-1} and 6897 cm^{-1} from references. These peaks correspond to the second-order overtones C-H stretching bonds related to fat and N-H stretching bonds related to water/protein. At about 5161 cm^{-1} are absorption maxima related to moisture in combination with O–H stretching and H–O–H deformation located. Lastly, between 6881 cm^{-1} and 5161cm^{-1}, there is a small peak corresponding to vibrations related to aliphatic residues of lipids in a range around 5670 cm^{-1}, which is the first-order overtone of C–H stretching;

$-CH_2-$ (Osborne *et al.*, 1993; Suksangpanomrung *et al.*, 2024; Workman Jr and Weyer, 2007). Due to the minor visual discrepancies between region-specific sample groups, the chemometrics was further explored.

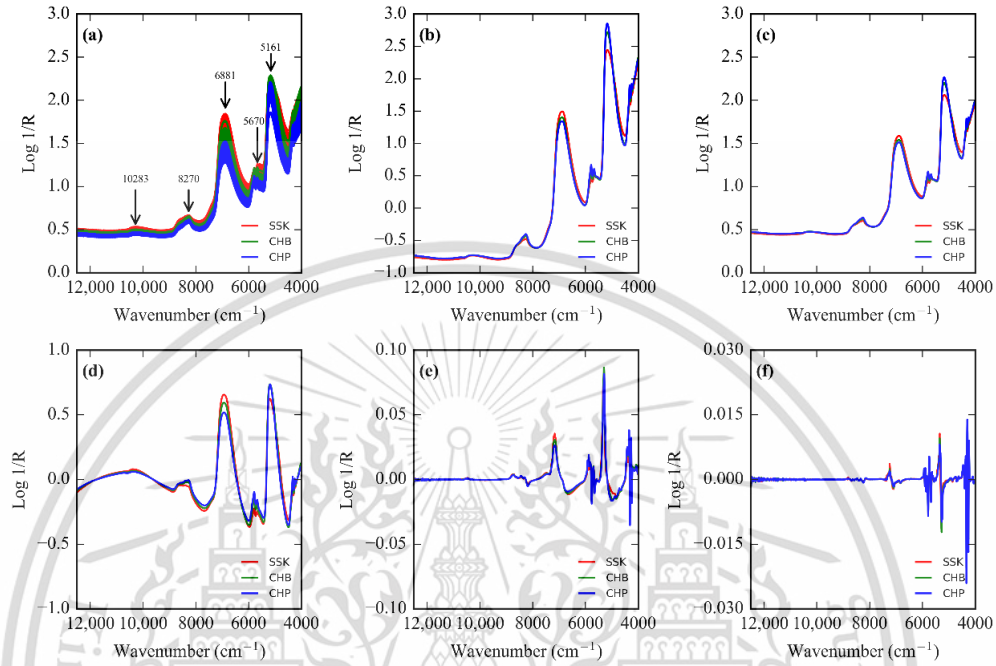


Figure 6.2. FT-NIR spectra of coconut milk in (a) Raw. The mean FT-NIR spectral for each group after preprocessing by (b) SNV, (c) MSC, (d) BSO, (e) FOD, and (f) SOD.

Figure 6.3a showcases the raw Micro-NIR spectra of coconut milk samples subjected to a geographical area source of coconut. Appropriate preprocessing is essential for subsequent modeling because NIR spectroscopy may be affected by undesirable scattering effects, including baseline shift and nonlinearity. In this work, several preprocessing methods, such as SNV, MSC, BSO, FOD, and SOD methods, were used individually to reduce the variability between samples due to the scatter effect and adjust baseline shifts between samples. After preprocessing, all NIR spectra from region-specific sample groups are shown in Figure 6.3b to Figure 6.3f. The NIR spectra display three clear absorption peaks at 980 nm, 1210 nm, and 1450 nm, corresponding to the second-order overtone vibration of the O-H stretching band, second-order overtones of the C-H stretching band related to fat, and first-order overtones of the O-H and N-H stretching bands connected to water/protein,

respectively (Osborne *et al.*, 1993; Workman Jr and Weyer, 2007). Notably, the 1210 nm peak emerging in this study aligned with the study by Suksangpanomrung *et al.* (2024), who reported that this peak corresponds to fatty acids from grated coconut for UHT coconut milk production.

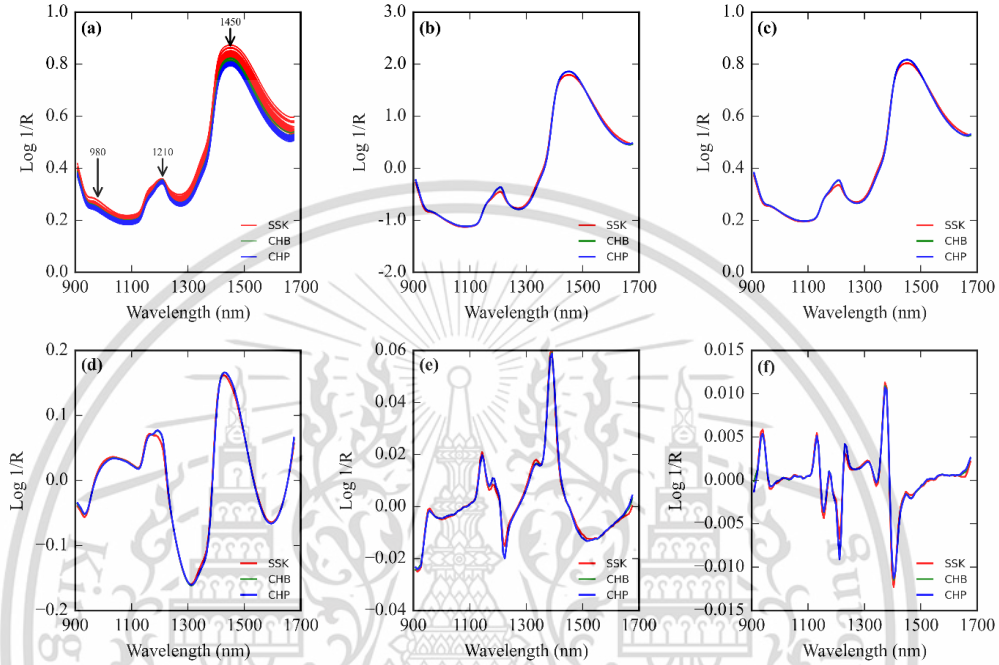


Figure 6.3. Micro-NIR spectra of coconut milk in (a) Raw. The mean Micro-NIR spectral for each group after preprocessing by (b) SNV, (c) MSC, (d) BSO, (e) FOD, and (f) SOD.

6.5.2 Discrimination Model Using FT-NIR

The determination of the geographical area of coconut milk was assessed qualitatively using FT-NIR spectra for a multiclass using classical to modern chemometrics classifiers. The corresponding performance results are listed in Table 6.5. Based on the nine classifiers coming from 3 group chemometrics classifiers (classical chemometrics, machine learning, deep learning), 45 models were trained, validated, and evaluated using the training and test sets processed with five different preprocessing methods. During the model training, an exhaustive GridsearchCV method was applied to find the optimal combination of hyperparameters for all models except the deep learning classifier. After that, iterations for all classifiers were executed until each model yielded the lowest error average of 5-fold cross-

validation of the training set. As is known and reported in many studies, cross-validation effectively reduces overfitting likelihood (Kabir *et al.*, 2021; Teye *et al.*, 2013)). Test sets were then used to evaluate the models' performance.

Table 6.5. Metric evaluation classifier using FT-NIR.

Chemo.	Reg.	Pre./Hyp.	Training (m=1102, n=216)				Testing (m=1102, n=54)				
			Pr_{wa}	Rc_{wa}	Fs_{wa}	Ac	Pr_{wa}	Rc_{wa}	Fs_{wa}	Ac	Kc
Classical	PCA	**SNV/PC=4**	**0.893**	**0.894**	**0.893**	**0.894**	**0.982**	**0.981**	**0.981**	**0.981**	**0.972**
		MSC/PC=6	0.894	0.894	0.893	0.894	0.949	0.944	0.943	0.944	0.917
		BSO/PC=3	0.922	0.921	0.920	0.921	0.928	0.926	0.926	0.926	0.889
		FOD/PC=2	0.872	0.829	0.812	0.829	0.843	0.704	0.631	0.704	0.556
		SOD/PC=2	0.822	0.713	0.639	0.713	0.500	0.667	0.556	0.667	0.500
	PLS-DA	SNV/LV=1	0.566	0.505	0.495	0.505	0.560	0.481	0.507	0.481	0.222
		MSC/LV=1	**0.565**	**0.500**	**0.491**	**0.500**	**0.565**	**0.500**	**0.521**	**0.500**	**0.250**
		BSO/LV=1	0.433	0.403	0.383	0.403	0.410	0.315	0.323	0.315	-0.028
		FOD/LV=1	0.469	0.505	0.477	0.505	0.421	0.370	0.388	0.370	0.056
		SOD/LV=1	0.494	0.528	0.503	0.444	0.433	0.444	0.438	0.444	0.167
	LDA	**SNV/LD=1**	**1.000**	**1.000**	**1.000**	**1.000**	**0.967**	**0.963**	**0.963**	**0.963**	**0.944**
		MSC/LD=1	1.000	1.000	1.000	1.000	0.967	0.963	0.963	0.963	0.944
		BSO/LD=1	1.000	1.000	1.000	1.000	0.967	0.963	0.963	0.963	0.944
		FOD/LD=1	0.986	0.986	0.986	0.986	0.907	0.870	0.865	0.870	0.806
		SOD/LD=1	0.986	0.986	0.986	0.986	0.907	0.870	0.865	0.870	0.806
Machine learning	SVM	**SNV/C=100, K=linear, D=2, G=scale**	**0.991**	**0.991**	**0.991**	**0.991**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
		MSC/C=100, K=linear, D=2, G=scale	0.968	0.968	0.968	0.968	0.982	0.981	0.981	0.981	0.972
		BSO/C=100, K=linear, D=2, G=scale	0.986	0.986	0.986	0.986	0.963	0.963	0.963	0.963	0.944
		FOD/C=100, K=poly, D=4, G=scale	0.995	0.995	0.995	0.995	0.982	0.981	0.981	0.981	0.972
		SOD/C=100, K=poly, D=3, G=scale	1.000	1.000	1.000	1.000	0.928	0.907	0.906	0.907	0.861
	KNN	SNV/nn=6	0.935	0.935	0.935	0.935	0.964	0.963	0.963	0.963	0.944
		MSC/nn=3	**0.968**	**0.968**	**0.967**	**0.968**	**0.982**	**0.981**	**0.981**	**0.981**	**0.972**
		BSO/nn=11	0.946	0.944	0.944	0.944	0.963	0.963	0.963	0.963	0.944
		FOD/nn=5	0.956	0.954	0.953	0.954	0.939	0.926	0.925	0.926	0.889
		SOD/nn=18	0.921	0.903	0.900	0.903	0.867	0.778	0.750	0.778	0.667
	ANN	SNV/H=(128)	0.861	0.806	0.802	0.806	0.905	0.889	0.891	0.889	0.833
		MSC/H=(128)	0.172	0.352	0.214	0.352	0.171	0.352	0.222	0.352	0.028
		BSO/H=(1024, 1024)	0.881	0.866	0.862	0.866	0.867	0.833	0.826	0.833	0.75
		FOD/H=(1024, 1024)	0.981	0.981	0.981	0.981	0.900	0.889	0.888	0.889	0.833

		SOD/H=(512, 512)	**1.00**	**1.00**	**1.00**	**1.00**	**0.939**	**0.926**	**0.925**	**0.926**	**0.889**
Deep learning	S-CNN	**SNV/e=205**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
		MSC/e=666	1.00	1.00	1.00	1.00	0.982	0.981	0.981	0.981	0.972
		BSO/e=568	0.995	0.995	0.995	0.995	1.00	1.00	1.00	1.00	1.00
		SOD/e=842	1.00	1.00	1.00	1.00	0.945	0.944	0.944	0.944	0.917
		FOD/e=74	1.00	1.00	1.00	1.00	0.982	0.981	0.981	0.982	0.972
	S-AlexNET	SNV/e=406	0.892	0.889	0.889	0.889	0.945	0.944	0.944	0.944	0.917
		MSC/e=395	0.898	0.898	0.898	0.898	0.926	0.926	0.925	0.926	0.889
		BSO/e=504	**0.948**	**0.944**	**0.944**	**0.948**	**0.945**	**0.944**	**0.944**	**0.944**	**0.917**
		SOD/e=296	0.940	0.940	0.940	0.940	0.952	0.944	0.944	0.944	0.917
		FOD/e=702	0.968	0.968	0.968	0.968	0.952	0.944	0.944	0.944	0.917
	ResNET	**SNV/e=366**	**0.991**	**0.991**	**0.991**	**0.991**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
		MSC/e=942	0.922	0.921	0.921	0.921	0.982	0.981	0.981	0.981	0.972
		BSO/e=181	0.981	0.981	0.981	0.981	0.981	0.981	0.981	0.981	0.972
		SOD/e=393	0.879	0.852	0.847	0.852	0.945	0.944	0.944	0.944	0.917
		FOD/e=857	0.850	0.833	0.829	0.833	0.861	0.852	0.850	0.852	0.778

Chemo.=chemometrics; Reg.=regressor; Pre.=preprocessing; Hyp.=hyperparameter; m=spectra features; n=samples spectra; Pr_{wa}=weighted average of precision; Rc_{wa}=weighted average of recall; Fs_{wa}=weighted average of f1-score; Ac=accuracy; Kc=Kappa coefficient; PCA=principal component analysis; PLS-DA=partial least squares discriminant analysis; LDA=linear discriminant analysis; SVM=support vector machine; KNN=*k*-nearest neighbors; ANN=artificial neural network; S-CNN=simple-convolutional neural network; SNV=standard normal variate; MSC=multiplicative scatter correction; BSO=baseline second order; FD=first-order derivatives; SD=second-order derivatives; PC=principal component; LD=linear discriminant variabel; LV=latent variable; C=penalty factor; K=kernel size; D=degree; G=gamma; nn=*n*-neighbor; H=hidden layer size; e=epoch.

6.5.2.1 By Classical Chemometric Classifier

The performances of models from classical chemometrics using FT-NIR spectroscopy were good in discriminating samples from a geographical area sources of coconut, with an accuracy range of 89.4% to 100% in training and 96.3% to 98.1% in testing, except for the PLS-DA classifier. According to the Kappa coefficient mentioned by Chu *et al.* (2022) as a performance evaluation for classifiers, this model has almost complete consistency because its coefficient is more than 80%. The best performance from both PCA and LDA classifiers should be supported with SNV preprocessing.

Opposite to other classical classifiers, among the many preprocessing techniques used in this investigation, only MSC preprocessing allowed the PLS-DA classifier to achieve the best performance. This matter arises because the number of latent variable (LV) from the PLS-DA classifier optimized via Gridsearch-5f-CV is just one variable, which explains the maximum of 89.1% of the variance in the data. In addition, the spectra visualization (Figure 6.2c) and 3D scatter scores of the PLS-DA

classifier supported by MSC preprocessing (Figure S4-1b) indicate that samples from province SSK are relatively clearly separated from samples from province CHP and CHB. However, samples from the CHP and CHB provinces have relatively smaller spaces and overlap. This meant that PLS-DA could not reveal the trend for geographical discrimination; advanced chemometrics should further explore information in the NIR spectra. This requires improving the PLS algorithm's decision-making boundary line when used for discrimination cases. At least, that is the concern of Dixon and Brereton (2009), where the classifier's performance should be improved by improving the boundary lines so that they work linearly and can be more adaptive towards quadratic, complex, linear, and complex, non-linear boundaries.

For visualization, Figure S4-1a and Figure S4-1c show the 3D PCA and LDA classifier score plots for the three relevant coconut milk geographical areas in the training dataset. A total of 90.47% of the variance is represented by the first principal component (PC1), 5.54% by PC2, and 2.91% by PC3. In contrast to the PCA classifier, a total of 94.74% of the variance for the LDA classifier is represented by the first linear discriminant (LD1) and 5.26% by LD2. This study has two LD variables as hyperparameter because the rule for the number of LD variables in the LDA classifier is determined by the number of classes to be classified is reduced by 1. Since these three PCs and LDs cover the total variance, the higher-order PCs and LDs could contain relevant information for determining geographical areas from coconut milk.

Several important wavenumbers from FT-NIR spectra using classical chemometrics can be identified with a high X-loading weight shown in Figure 6.4. Those wavenumbers are considered to be important for the differentiation between the classes. These wavenumbers are closely related to the absorbance of several important chemical components in coconut milk. In Figure 6.4a, several wavenumbers with high loading of PCs from PCA classifier at 8250, 7200, 6696, 5794, 5680, 5320, 5170, 4780, and 4335 cm^{-1} can be observed. For instance, the wavenumber at 8250, 7200 cm^{-1} is related to the absorbance of fat, while those at 6696 cm^{-1} are closely related to the absorbance of protein/cellulose (Suksangpanomrung *et al.*, 2024). From Figure 6.4b, several wavenumbers can be determined with high loading of LVs from the PLS-DA classifier, including

wavenumbers at 10,250, 8650, 8250, 7810, 7340, 5920, 5630, 5300, 4950, and 4400 cm^{-1}. The wavenumber at 10,250 cm^{-1} is related to the absorbance of water, while those at 8650, 8250, 7810, and 7340 cm^{-1} are closely related to the absorbance of fat (Suksangpanomrung *et al.*, 2024). Lastly, Figure 6.4c notes several wavenumbers with high loading of LDs from the LDA classifier at 8610, 7770, 7685, and 7340 cm^{-1}. The wavenumbers at 8610 and 7770 cm^{-1} are related to the absorbance of fat, while those at 7685 and 7340 cm^{-1} are closely related to the absorbance of water (Suksangpanomrung *et al.*, 2024). For assigning important functional groups spectral from classical chemometrics to identify geographical region-specific coconut milk based on NIR molecular overtones and combinations data as fully shown in Table S4-1.

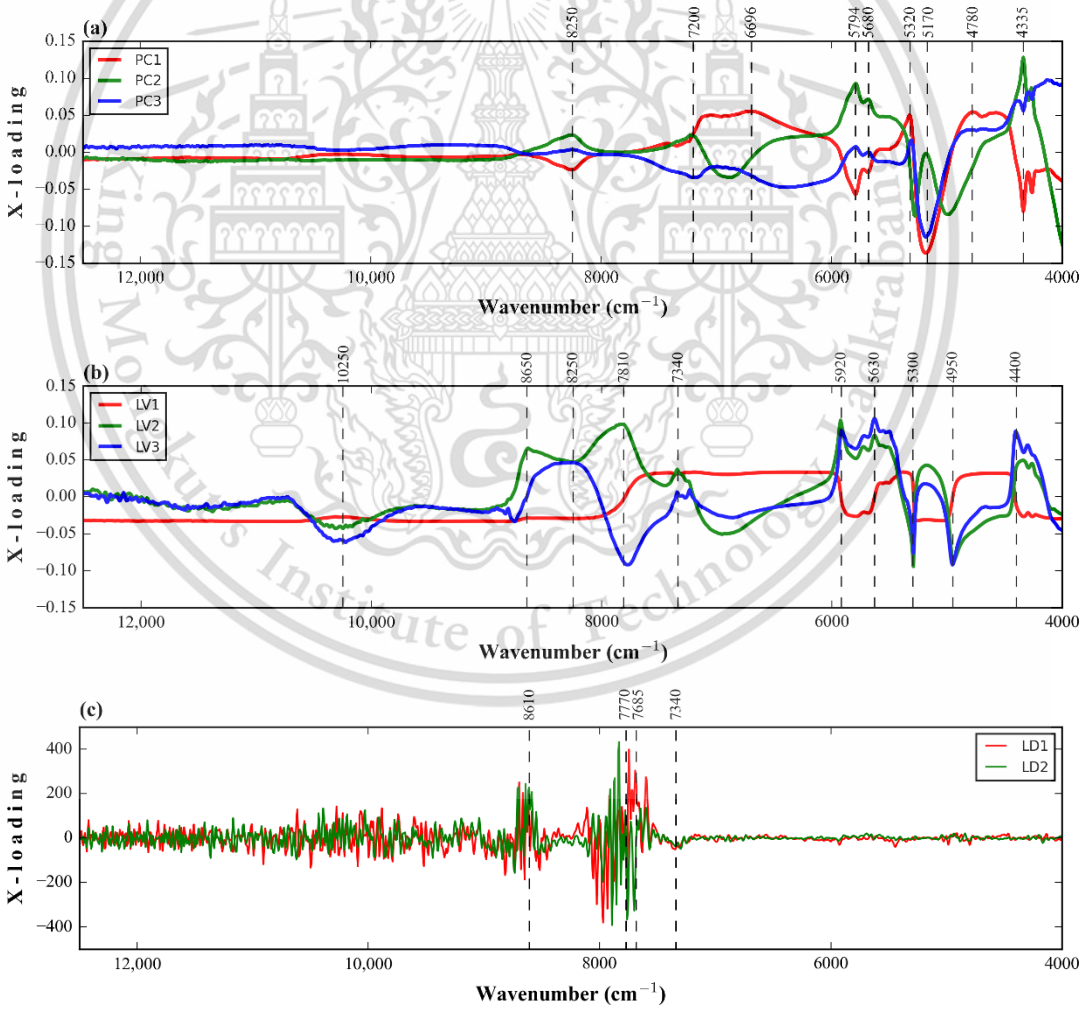


Figure 6.4. Loading from classical chemometrics classifier using FT-NIR. (a) PCA, (b) PLS-DA, (c) LDA.

6.5.2.2 By Machine Learning Classifier

The classifier from machine learning can distinguish relatively obvious using FT-NIR between region-specific sample groups compared to classical chemometrics. All classifiers have performance in the accuracy range of 96.8% to 100% in training and 92.6% to 100% in testing. With Kappa coefficient, all classifiers are excellent models for any application to discriminate region-specific coconut milk (Chu *et al.*, 2022). To achieve this performance, SVM, KNN and ANN classifiers need support from SNV, MSC, SOD preprocessing, respectively. Also, the SVM classifier needs a hyperparameter setting with a penalty factor of about 100, a degree of 2, a linear kernel, and a scale in gamma. KNN classifier needs three n-neighbors, and the ANN classifier needs two hidden layer sizes with each 512 neuron. The results in this study were consistent with findings from Chen *et al.* (2009) when employing an SVM, KNN, and ANN classifier to discriminate multi-class of the geographical origin from Chinese green tea via FT-NIR with performance accuracy in testing achieved 92.3%, 96.3%, 96.3%, respectively.

Figure 6.5 a highlights the ablation intensity from all machine learning classifiers for geographical region-specific coconut milk discrimination, revealing specific wavenumbers with a prominent influence. In ablation intensity, for identifying feature-important NIR spectra from machine learning classifiers, a positive value indicates that removing the feature can reduce model performance, and a negative value indicates that removing the feature can improve model performance (Képeš *et al.*, 2022). Therefore, in this study, several wavenumber ablation intensity values are noted with pink marks, which indicate a negative value as noise, and yellow marks, which indicate important features. When we overlap the machine learning classifier, it is known that the wavenumbers about 5238 and 5138 cm^{-1} are the most influential in the performance of the machine learning-based model. The wavenumbers are each related to the combination of H_2O (OH-stretching+OH-bending) and second-order overtone -COOH (2×C=O-stretching)(Conzen, 2006).

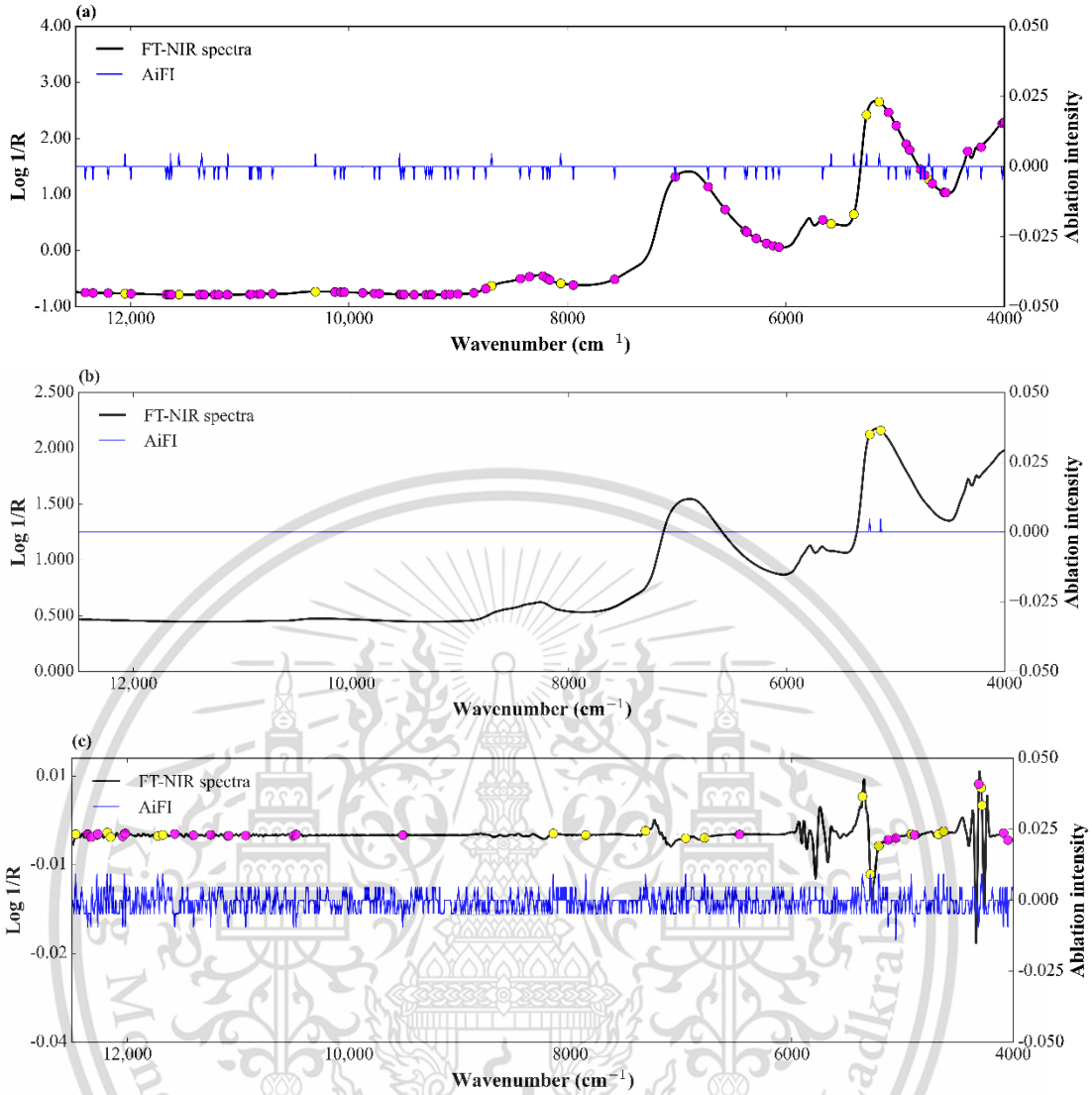


Figure 6.5. Ablation intensity from machine learning classifier using FT-NIR. (a) SVM (b) KNN, (c) ANN. ○ Feature importance; ● Feature noise.

6.5.2.3 By Deep Learning Classifier

The discrimination models employing deep learning can be further evaluated in-depth, as Table 6.5 details the performance metrics for all classifiers. The discrimination accuracy rate of all classifiers from the deep learning model is between 94.4% to 100% in the training set and 94.4% to 100% in the prediction set, respectively. According to the Kappa coefficient, classifiers from deep learning also perform excellently and can be categorized as being used for any application (Chu *et al.*, 2022). In order to obtain this level of performance, each architecture of deep learning classifiers (S-CNN, S-AlexNET, ResNET) requires the respective support of SNV,

BSO, and SNV preprocessing. The results of this study partially support the hypothesis of previous research (Cui and Fearn, 2018; Jiang *et al.*, 2020), which demonstrated that the deep learning algorithm in the regression case is highly adaptable in data processing and does not require any data preprocessing except normalization. The difference is that this study investigates discrimination cases and finds that the S-CNN and ResNET classifier architectures are sufficiently represented by SNV preprocessing to bring the best performance. Besides, for the S-AlexNET classifier architecture, if we look more closely using the kappa coefficient, we also don't find any major differences between SNV and BSO preprocessing in the testing set. The difference between both is only located in the training set where the model with BSO preprocessing (accuracy of 94.8%) is superior to SNV preprocessing (accuracy of 89.9%).

The outstanding performance of deep learning classifiers was reached by S-CNN models. This approach provided higher confusion matrix values for S-CNN classifier than the other predictive model. According to the Kappa coefficients, the S-CNN-based models had higher predictions for estimating most region-specific sample groups than other classifiers. Also, to support this evaluation, the proper approach for all deep learning classifiers is to catch the learning curve, and both the training and testing sets should converge (Géron, 2019). A significant distance between the training and testing curves indicates overfitting or underfitting. From the learning curve, we can check the predicted accuracy values of the training and test sets, which can indicate overfitting. Figure S4-2a indicates the S-CNN classifier could meaningfully escape overfitting chance and trap in the local optima.

The application of deep learning classifiers to the preprocessing dataset identifies different important features of deep learning for geographical region-specific coconut milk (see Figure 6.6). For the classifier using CNN and ResNET, because both utilize the same preprocessing (SNV), the important features are in the region 7267–7213 cm^{-1}. According to Conzen (2006), this area combines H_2O (OH-anti-symmetry stretching+OH-symmetry stretching).

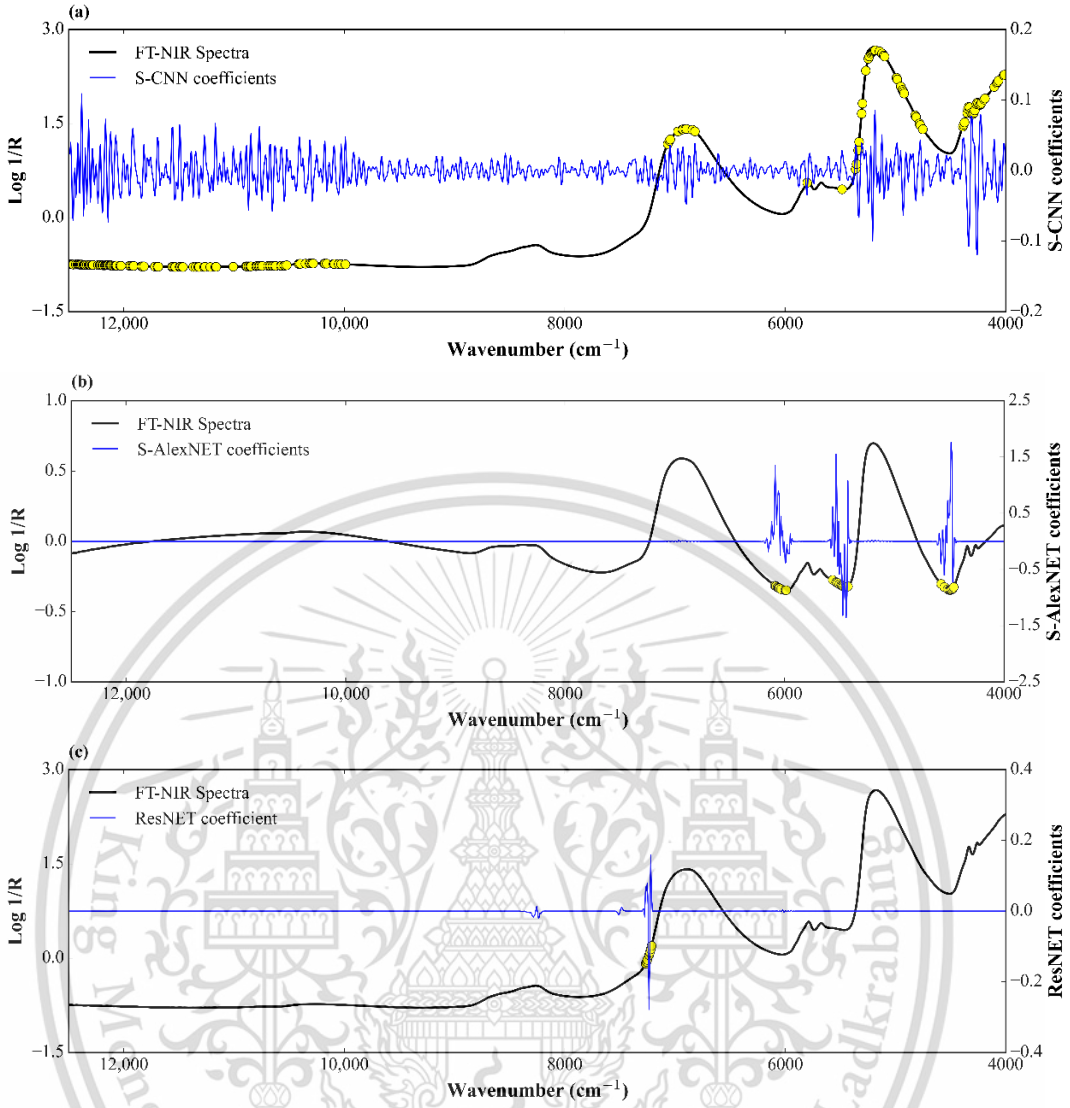


Figure 6.6. Regression coefficients from deep-learning classifier using FT-NIR. (a) S-CNN, (b) S-AlexNET, (c) ResNET. ○ Feature importance.

6.5.3 Discrimination Model Using Micro-NIR

This study tested and compared three approaches using classical chemometrics (PCA, PLS-DA, LDA), machine learning (SVM, KNN, ANN), and deep chemometrics (S-CNN, S-AlexNET, ResNET) for multiclass discrimination samples from geographical area sources of coconut (CHP, SSK, CHB) based on the Micro-NIR in the range 908–1676 nm (11,013–5967 cm^{-1}). Table 6.6 shows an overview of the performance of the discrimination models for all classifiers with whole preprocessing. All procedures for FT-NIR spectra analysis are also employed for Micro-NIR spectra. The identical procedure is used, initiating from the data splitting approach, optimizing

the classifier by GridsearchCV, and evaluating parameters. In general, the classifier model's accuracy performance in training and testing is between 33.3% and 100%. Moreover, comparing all classifier models based on the Kappa coefficient in the testing stage, as reported by Lu *et al.* (2020), we found that the classifiers from LDA, KNN, and ResNET work perfectly (100%) to differentiate coconut milk samples from their geographical source areas based on Micro-NIR data. This finding further confirms the reliability and effectiveness of the discrimination model proposed in this study. Several factors may influence the composition of coconut milk samples, so in this study, an analytical approach like NIR spectroscopy can be used to detect them. As we know, agricultural product differences can be influenced by crop maturity, session harvesting, irrigation, fertilizer applications, and other farming practices in a particular region (Kabir *et al.*, 2021; Szymczycha-Madeja *et al.*, 2014).

Table 6.6. Metric evaluation classifier using Micro-NIR.

Chemo.	Reg.	Pre./Hyp.	Training (m=125, n=216)				Testing (m=125, n=54)				
			Pr_{wa}	Rc_{wa}	Fs_{wa}	Ac	Pr_{wa}	Rc_{wa}	Fs_{wa}	Ac	Kc
Classical	PCA	SNV/PC=5	0.869	0.843	0.837	0.843	0.778	0.766	0.805	0.778	0.667
		MSC/PC=5	0.858	0.769	0.735	0.769	0.867	0.778	0.750	0.778	0.667
		BSO/PC=2	0.853	0.755	0.714	0.755	0.826	0.685	0.590	0.685	0.528
		FOD/PC=4	0.857	0.750	0.709	0.750	0.881	0.815	0.799	0.815	0.722
		SOD/PC=6	**0.846**	**0.843**	**0.842**	**0.843**	**0.854**	**0.852**	**0.851**	**0.852**	**0.778**
	PLS-DA	**SNV/LV=1**	**0.620**	**0.588**	**0.596**	**0.588**	**0.619**	**0.611**	**0.605**	**0.611**	**0.417**
		MSC/LV=1	0.620	0.588	0.596	0.588	0.619	0.611	0.605	0.611	0.417
		BSO/LV=1	0.591	0.560	0.560	0.560	0.618	0.574	0.555	0.574	0.361
		FOD/LV=1	0.582	0.546	0.545	0.546	0.424	0.389	0.384	0.389	0.083
		SOD/LV=1	0.634	0.602	0.605	0.602	0.550	0.537	0.543	0.537	0.306
	LDA	**SNV/LD=1**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
		MSC/LD=1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		BSO/LD=1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		FOD/LD=1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		SOD/LD=1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Machine learning	SVM	SNV/C=100, K=linear, D=2, G=scale	0.948	0.944	0.944	0.944	0.952	0.944	0.944	0.944	0.917
		MSC/C=100, K=poly, D=4, G=scale	0.969	0.968	0.968	0.968	0.945	0.944	0.944	0.944	0.917
		BSO/C=100, K=poly, D=4, G=scale	0.884	0.870	0.868	0.870	0.847	0.833	0.830	0.833	0.750
		FOD/C=100, K=poly, D=3, G=scale	0.941	0.940	0.940	0.940	0.929	0.926	0.926	0.926	0.889
		SOD/C=100, K=poly, D=4, G=scale	**0.995**	**0.995**	**0.995**	**0.995**	**1.000**	**1.00**	**1.00**	**1.00**	**1.00**
	KNN	**SNV/nn=1**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**

		MSC/nn=1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		BSO/nn=1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		FD/nn=1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		SD/nn=1	1.00	1.00	1.00	1.00	0.967	0.963	0.963	0.963	0.944
	ANN	SNV/H=(256, 256)	0.695	0.676	0.672	0.676	0.714	0.685	0.665	0.685	0.528
		MSC/H=(256)	0.111	0.333	0.167	0.111	0.111	0.333	0.167	0.333	0.00
		BSO/H=(512, 512)	0.849	0.796	0.781	0.796	0.781	0.759	0.750	0.759	0.639
		FOD/H=(1024, 1024)	**0.726**	**0.722**	**0.724**	**0.726**	**0.856**	**0.852**	**0.853**	**0.852**	**0.778**
		SOD/H=(1024, 1024)	0.111	0.333	0.167	0.333	0.111	0.333	0.167	0.333	0.00
Deep learning	S-CNN	SNV/e=946	0.820	0.792	0.782	0.792	0.805	0.778	0.766	0.778	0.667
		MSC/e=962	0.848	0.843	0.841	0.843	0.807	0.796	0.792	0.796	0.694
		BSO/e=387	**0.844**	**0.838**	**0.837**	**0.838**	**0.892**	**0.889**	**0.889**	**0.889**	**0.833**
		SOD/e=21	0.750	0.741	0.733	0.741	0.875	0.852	0.848	0.852	0.778
		FOD/e=838	0.111	0.333	0.167	0.333	0.111	0.333	0.167	0.333	0.00
	S-AlexNET	SNV/e=569	0.774	0.773	0.773	0.774	0.745	0.741	0.738	0.741	0.611
		MSC/e=642	**0.838**	**0.810**	**0.802**	**0.810**	**0.833**	**0.815**	**0.810**	**0.815**	**0.722**
		BSO/e=34	0.723	0.704	0.676	0.704	0.720	0.704	0.679	0.704	0.556
		SOD/e=304	0.867	0.778	0.750	0.778	0.874	0.796	0.775	0.796	0.694
		FOD/e=896	0.841	0.829	0.826	0.841	0.819	0.796	0.788	0.796	0.694
	ResNET	SNV/e=912	0.978	0.977	0.977	0.977	0.945	0.944	0.944	0.944	0.917
		MSC/e=859	0.859	0.806	0.791	0.806	0.874	0.796	0.775	0.796	0.694
		BSO/e=777	**0.972**	**0.972**	**0.972**	**0.972**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
		SOD/e=969	0.787	0.764	0.757	0.764	0.763	0.704	0.684	0.704	0.556
		FOD/e=387	0.724	0.681	0.692	0.681	0.763	0.704	0.714	0.704	0.556

Chemo.=chemometrics; Reg.=regressor; Pre.=preprocessing; Hyp.=hyperparameter; m=spectra features; n=samples spectra; Pr_{wa}=weighted average of precision; Rc_{wa}=weighted average of recall; Fs_{wa}=weighted average of f1-score; Ac=accuracy; Kc=Kappa coefficient; PCA=principal component analysis; PLS-DA=partial least squares discriminant analysis; LDA=linear discriminant analysis; SVM=support vector machine; KNN=*k*-nearest neighbors; ANN=artificial neural network; S-CNN=simple-convolutional neural network; SNV=standard normal variate; MSC=multiplicative scatter correction; BSO=baseline second order; FD=first-order derivatives; SD=second-order derivatives; PC=principal component; LD=linear discriminant variabel; LV=latent variable; C=penalty factor; K=kernel size; D=degree; G=gamma; nn=*n*-neighbor; H=hidden layer size; e=epoch.

6.5.3.1 By Classical Chemometric Classifier

Using Micro-NIR in tandem with classical chemometrics, the performances of classifier models can be categorized as middle-good in discriminating samples from a geographical area of coconut milk sources, with the best accuracy of each preprocessing in the range of 58.8% to 100% in training and 61.1% to 100% in testing (Table 6.6). Based on the Kappa coefficient parameter mentioned by Chu *et al.* (2022), the LDA classifier in this study is the best classifier model from classical chemometrics (100%). The LDA classifier algorithm can operate all types of preprocessing used in the study to perform excellently with only one LD component hyperparameters. Therefore, the preprocessing SNV is selected to support the LDA

classifier for further analysis in the 3D score scatter plot in Figure S4-3c. The variance of LD1 and LD2 is 90.1% and 9.9%, respectively. This value shows that samples from three geographical areas of coconut milk sources can be discriminated excellent based on Micro-NIR, which performs as well as using FT-NIR spectra. On the other hand, scatter in the 3D plot clearly visualizes the PCA classifier's support by second-order derivatives and 6 PC hyperparameters based on Micro-NIR spectra in discriminating three relevant coconut milk geographical areas in the training dataset shown in Figure S4-3a. The total variance from the PCA classifier using Micro-NIR spectra is 88.57%, which consists of PC1, P2, and PC3 at 58.73%, 25.42%, and 4.42%, respectively. This total variance is lower than that found when using FT-NIR (98.91%), which lowers the performance of the Micro-NIR PCA classifier model compared to FT-NIR.

The weak classifier is the PLS-DA classifier, supported by SNV preprocessing and has a kappa coefficient performance in the testing of 41.7%. It is little bit better than using the FT-NIR dataset (25.0%) supported by MSC preprocessing with the same number of LVs. This can be explained even though the samples from the SSK province have been significantly separated from the samples from the CHP and CHB provinces, which were overlapped using both FT-NIR and Micro-NIR. However, the scatter of these three samples tends to be better when visualizing the 3D scores scatter data using Micro-NIR (Figure S4-3b). This is confirmed by the LV2 variance value (8.47%) in the Micro-NIR sample being greater than the LV2 variance value (5.52%) in the FT-NIR sample. From this value, the total variance up to the second LV from using Micro-NIR (96.12%) is greater than FT-NIR (94.62%). This is in line with the statement from Brereton and Lloyd (2018), which states that the faster a variance reaches its maximum on the lowest variable, the greater the variance will be, which can explain the contents of the entire data. This is proven in the case of the classification of samples from a geographical area sources of coconut milk in this study.

The X-loading parameter is utilized to interpret several important wavelengths corresponding to discriminatory chemical properties of the geographical area of coconut milk using classical chemometrics. It can be seen that several wavelengths that appear as important features in the application of the PCA classifier include 975,

1125, 1210, 1330, 1380, 1395, and 1425 nm (Figure 6.7a). The peaks seen at wavelengths of 975 and 1425 nm are closely associated with water absorption, therefore verifying their significant contribution to the correctness of the discrimination model. Meanwhile, other wavelengths at 1125, 1210, 1330, 1380, and 1395 nm are specifically correlated to fat, protein, and cellulose. For the PLS-DA classifier, an important feature represented by wavelengths at 1140, 1150, and 1220 nm is closely associated with fat; about 1410 nm is related to water, and 1520 nm corresponds to protein. Lastly, the LDA classifier's important feature represented by wavelengths 1090, 1125, 1135, and 1275 nm shows fat; peaks about 1325 and 1465 nm correspond to water; peaks about 1350 and 1380 nm associate with acid (Conzen, 2006; Suksangpanomrung *et al.*, 2024; Workman Jr and Weyer, 2007). The wave correlation and absorption regions of major peaks found in coconut milk are shown in Table S4-1.

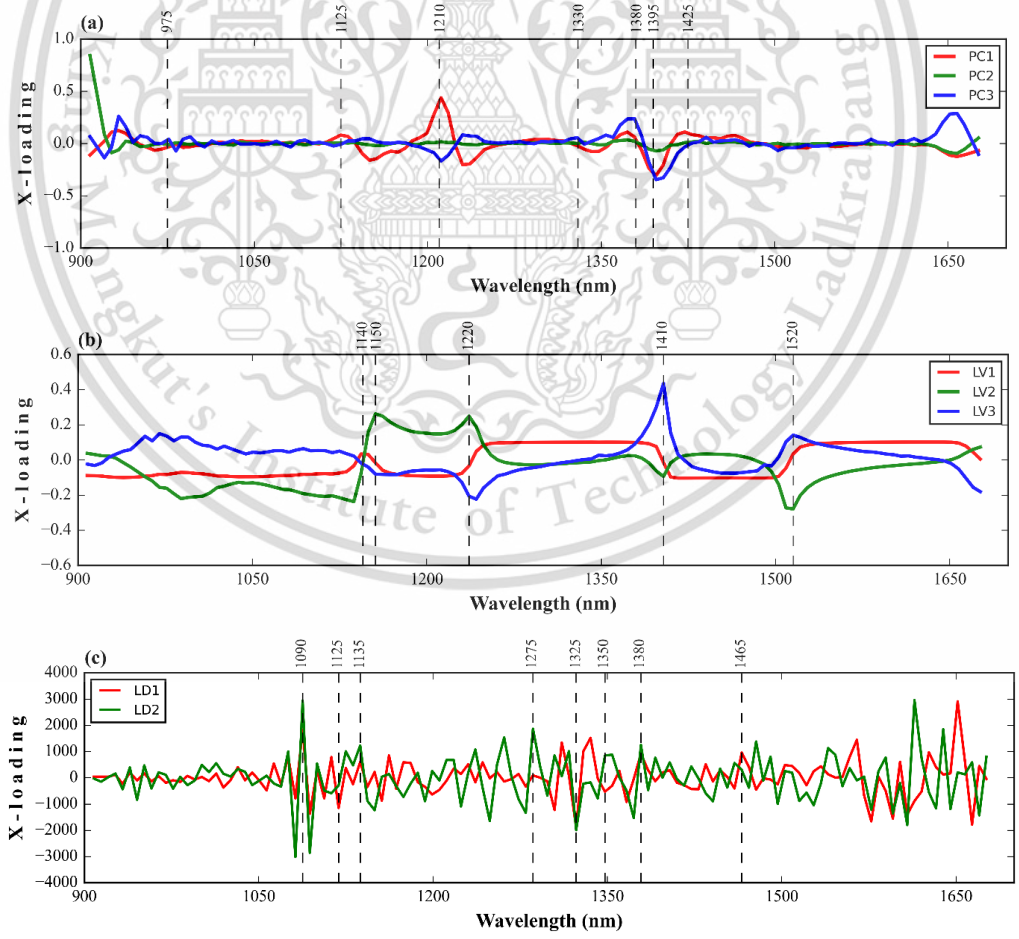


Figure 6.7. Loading from classical chemometrics classifier using Micro-NIR. (a) PCA, (b) PLS-DA, (c) LDA.

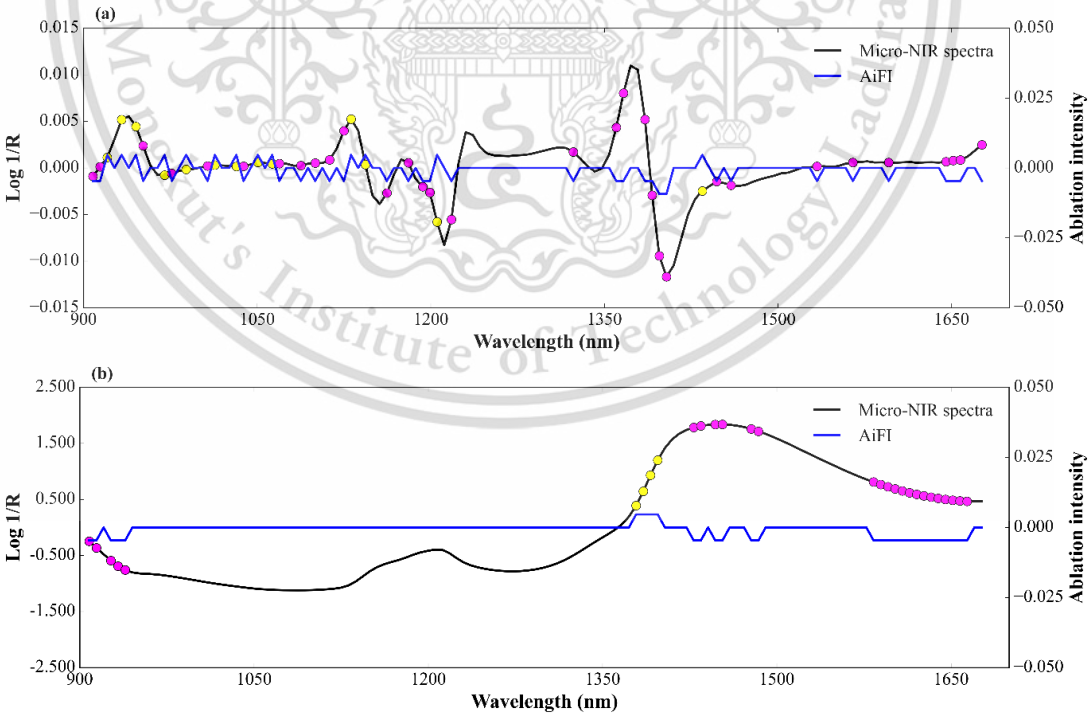
6.5.3.2 By Machine Learning Classifier

Three classifier models using a machine learning algorithm based on Micro-NIR were also found to discriminate samples from a geographical area of coconut milk sources with the best accuracy in the range of 72.6% to 100% in training and 85.2% to 100% in testing (Table 6.6). According to statistical performance classification using the Kappa coefficient from Chu *et al.* (2022), all classifiers machine learning in this case are excellent models for any application to discriminate geographical region-specific coconut milk (Kc of 100%), except the ANN classifier (Kc of 77.8%). The performance of each classifier from machine learning is supported by preprocessing SOD for SVM, SNV for KNN, and FOD for ANN. Additionally, every machine learning classifier requires hyperparameter tuning where for an SVM classifier with a penalty factor of about 100, a degree of 4, a polynomial kernel, and a scale in gamma; for KNN classifier with n-neighbors of 1; for ANN classifier with two hidden layer sizes with 1024 neurons for each layer.

When comparing the performance of classifier models using the same Micro-NIR data set, it was found that the SVM and KNN classifiers from machine learning were as good as the LDA classifier from classical chemometrics in discriminating geographical region-specific regions of coconut milk. The performance of the LDA and KNN classifiers was slightly decreased when using the dataset from FT-NIR with kappa coefficients of 94.4% and 97.2%, respectively. However, the SVM classifier shows equally good performance with a perfect kappa coefficient using both the dataset from FT-NIR and the dataset from Micro-NIR. However, the difference is in the preprocessing conditions and hyperparameters used to achieve the best discrimination model performance. This is in line with the report by Kabir *et al.* (2021) in the case of discrimination of geographic origin from millet using Vis-NIR (350 to 2500 nm) which shows that a classifier that works non-linearly (SVM) can still perform reasonably for cases where the probability of the dataset is linear which is probably more suitable for linear classifiers such as KNN and LDA.

Several important wavelengths by ablation intensity approach corresponding to discriminatory chemical properties of the geographical area of coconut milk using machine learning were highlighted in Figure 6.8. At least positive ablation intensity values (yellow marks) are spread between 908 and 1676 nm as much as for the SVM

classifier with 13 values, KNN classifier with 4 values, and ANN classifier with 6 values. Also, negative ablation intensity values (pink marks) for the SVM, KNN, and ANN classifiers were 32 values, 25 values, and 64 values, respectively. The differences in information retrieval related to features important wavelengths from each classifier are influenced by the differences in each preprocessing used by the classifier algorithm. In particular, the SVM classifier model supported by SOD preprocessing implies the features important from wavelengths about 932 and 1205 nm, which correspond to water and fat, respectively (Pandiselvam *et al.*, 2022; Suksangpanomrung *et al.*, 2024). However, in the KNN classifier supported by SNV preprocessing, the important features range from approximately 1385 to 1397 nm, corresponding to the CH_2 combination 2×CH-stretching+3×CH-bending (Conzen, 2006; Pandiselvam *et al.*, 2022). In contrast to the ANN classifier, which is supported by FOD preprocessing where the features important from wavelengths about 1075 and 1273 nm are related to fat, 1391 nm corresponding to acid and wavelength at approximately 1422 and 1428 nm correlated to water/protein (Suksangpanomrung *et al.*, 2024).



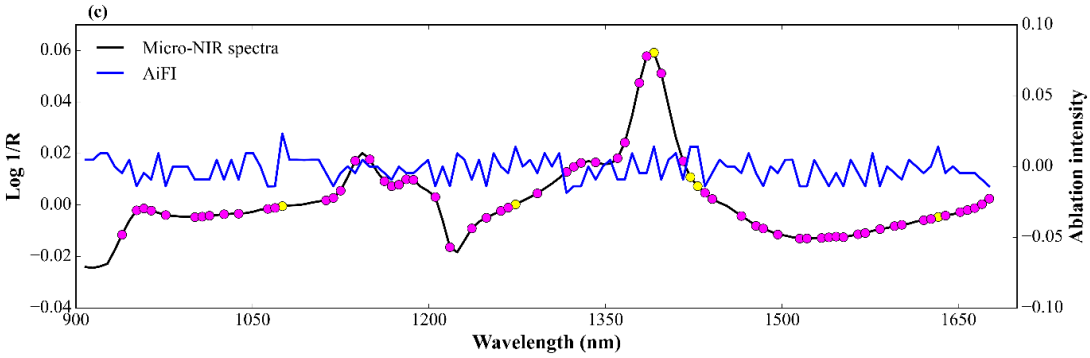


Figure 6.8. Ablation intensity from machine learning classifier using Micro-NIR. (a) SVM (b) KNN, and (c) ANN. ○ Feature importance, ● Feature noise.

6.5.3.3 By Deep Learning Classifier

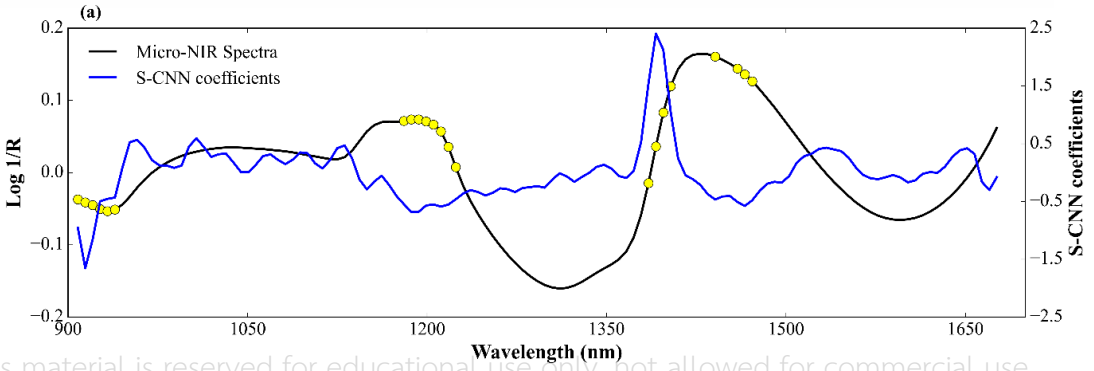
The discrimination model performance based on Micro-NIR for each deep learning classifier was tabulated in Table 6.6. The discrimination accuracy rate of all classifiers from the deep learning model is between 81.0% to 97.2% in the training set and 81.5% to 100% in the prediction set, respectively. The best classifier based on the Kappa coefficient from deep learning classifier to discriminate samples of coconut milk from its geographical area sources using Micro-NIR in order from the best to worst is ResNET > S-CNN > S-AlexNET. The ResNET and S-CNN classifiers are included in the category for use in any application because their performance based on the Kappa coefficient is more than 80%. Meanwhile, the S-AlexNET classifier is included in the substantive performance category with high consistency because it is between 60 and 80% (Chu *et al.*, 2022). In order to achieve this performance, each S-CNN, S-AlexNET, and ResNET classifier is supported by BSO, MSC, and BSO preprocessing.

In the case of using the Micro-NIR dataset, this is inverse to the hypothesis from previous research results reported by Cui and Fearn (2018) and Jiang *et al.* (2020), which states that deep learning algorithms are highly adaptable in data processing and do not require any data preprocessing anymore except normalization. In fact, this study shows that preprocessing can still be proven to improve the performance of the discrimination model. Irregularities in one such study related to or without preprocessing were also found by Acquarelli *et al.* (2017) when developing a CNN algorithm and testing it on several spectroscopy datasets. The study states that

appropriate preprocessing can still improve the accuracy of a deep-learning model, particularly from CNN.

When comparing the performance of classifier models using the same Micro-NIR data set, it was found that the ResNET classifier from deep learning was as good as the LDA classifier from classical chemometrics and SVM and KNN classifiers from machine learning in discriminating geographical region-specific regions of coconut milk. Also, the performance of the ResNET classifier from deep learning did not change much when using the FT-NIR dataset in the testing stages (Ks=100%). However, if we observe the learning curve from ResNET, it tends to have a lot of noise throughout the epoch between the training and testing curve and is different from the S-CNN classifier, which tends to be smooth and stable (Figure S4-4). This is believed to be because the ResNET architecture is much deeper and more complex in extracting information from the spectrum than the S-CNN classifier (Wu *et al.*, 2019).

Figure 6.9 highlights the identification of several important wavelengths by weights used in each layer network corresponding to discriminatory chemical properties of the geographical area of coconut milk using deep learning base on Micro-NIR. In the S-CNN classifier, at least 22 important features are scattered across several wavelength regions, from 908 to 939 nm, 1180 to 1224 nm, 1385 to 1403 nm, and 1440 to 1471 nm. Meanwhile, in the S-AlexNET classifier, important features are found in 5 wavebands starting from 945 nm and 1211–1230 nm. Finally, the ResNET classifier found 14 important features spread from wavelength 914–963 nm and 1378–1403 nm. This absorbance wavelength range is commonly related to water (1425 nm) and fat (1125, 1180, 1200, 1220, and 1390 nm) (Suksangpanomrung *et al.*, 2024).



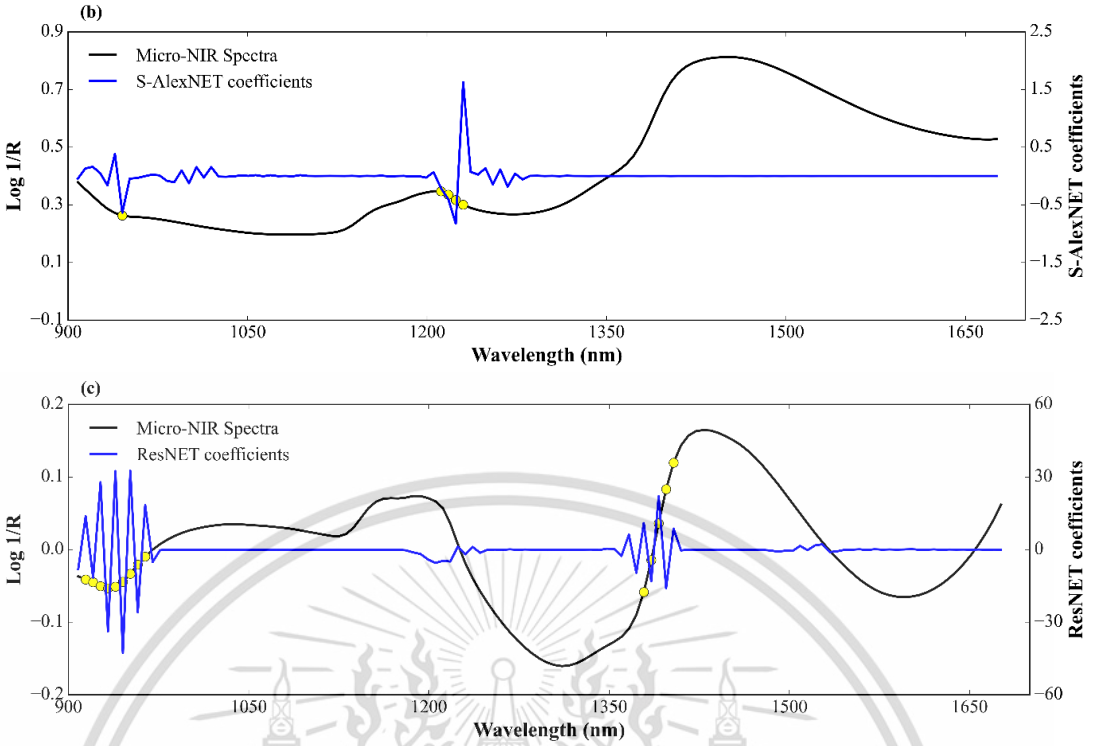


Figure 6.9. Regression coefficients from deep-learning classifier using Micro-NIR. (a) S-CNN, (b) S-AlexNET, and (c) ResNET. ○ Feature importance.

6.6 Conclusions

The present study demonstrated that NIR spectroscopy (FT-NIR and Micro-NIR) combined with a classical to modern chemometrics classifier could successfully discriminate the geographical area of coconut milk from three different province regions in Thailand. The NIR spectra effectively differentiate coconut milk from different geographic areas because the environmental differences in different regions may also affect the content of coconut milk. Also, the approach of x-loading for classical chemometrics classifier, as known before, the ablation feature for machine learning classifier, and the weight approach for deep learning classifier showed how it could be easily applied to select feature important NIR spectral in both FT-NIR and Micro-NIR.

For FT-NIR, the overall accuracy and kappa coefficient of the best model from each group chemometrics classifier identification was in the range of 96.3–100% and 99.4–100%, respectively. From classical representation with the LDA classifier, machine learning is represented by the SVM classifier, and deep learning is defined

by the S-CNN classifier. The Micro-NIR method achieved overall accuracy and kappa coefficient in testing 1.00 for the best model in each group of chemometrics classifier identification. In the context of classical chemometrics represented by the LDA classifier, machine learning is described by the SVM and KNN classifiers and deep learning is defined by the ResNET classifier.

This work not only proves that the coconut milk samples from different regions could be identified rapidly and efficiently but also provides a novel reference for geographical area discrimination of this valuable agricultural product. Moreover, this study confirms that market regulators can use NIR reliability to improve the authenticity of coconut milk in the marketplace. Last but not least, to improve the predictive ability and robustness of the discrimination models, more coconut milk samples testing with different harvesting seasons and in different regions and provinces should be added to the analysis.

6.7 References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., & Isard, M. (2016). *TensorFlow: a system for Large-Scale machine learning.* Paper presented at the 12th USENIX symposium on operating systems design and implementation (OSDI 16).

Acquarelli, J., van Laarhoven, T., Gerretzen, J., Tran, T. N., Buydens, L. M. C., & Marchiori, E. (2017). Convolutional neural networks for vibrational spectroscopic data analysis. *Analytica Chimica Acta, 954*, 22-31.

Arndt, M., Rurik, M., Drees, A., Ahlers, C., Feldmann, S., Kohlbacher, O., & Fischer, M. (2021). Food authentication: Determination of the geographical origin of almonds (Prunus dulcis Mill.) via near-infrared spectroscopy. *Microchemical Journal, 160*, 105702.

Barnes, R. J., Dhanoa, M. S., & Lister, S. J. (1989). Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra. *Applied spectroscopy, 43*(5), 772-777.

Breiman, L. (2001). Random forests. *Machine learning, 45*, 5-32.

Brereton, R. G. (2022). Numerical introduction to principal components analysis. *Journal of Chemometrics, 36*(8), e3405.

Brereton, R. G., & Lloyd, G. R. (2010). Support Vector Machines for classification and regression. *Analyst, 135*(2), 230-267.

Brereton, R. G., & Lloyd, G. R. (2014). Partial least squares discriminant analysis: taking the magic away. *Journal of Chemometrics, 28*(4), 213-225.

Brereton, R. G., & Lloyd, G. R. (2018). Partial least squares discriminant analysis for chemometrics and metabolomics: How scores, loadings, and weights differ according to two common algorithms. *Journal of Chemometrics, 32*(4).

Chen, Q., Zhao, J., & Lin, H. (2009). Study on discrimination of Roast green tea (Camellia sinensis L.) according to geographical origin by FT-NIR spectroscopy and supervised pattern recognition. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 72*(4), 845-850.

Chu, X., Huang, Y., Yun, Y.-H., & Bian, X. (2022). *Chemometric methods in analytical spectroscopy technology*: Springer.

CODEX-STAN-240. (2003). Standard for Aqueous Coconut Products-Coconut Milk and Coconut Cream.: FAO/WHO Food Standards Programme.

Conzen, J.-P. (2006). *Multivariate calibration*.

Cui, C., & Fearn, T. (2018). Modern practical convolutional neural networks for multivariate regression: Applications to NIR calibration. *Chemometrics and Intelligent Laboratory Systems, 182*, 9-20.

De Girolamo, A., Cortese, M., Cervellieri, S., Lippolis, V., Pascale, M., Logrieco, A. F., & Suman, M. (2019). Tracing the Geographical Origin of Durum Wheat by FT-NIR Spectroscopy. *Foods, 8*(10), 450.

Delwiche, S. R., & Reeves, J. B. (2010). A graphical method to evaluate spectral preprocessing in multivariate regression calibrations: Example with Savitzky–Golay filters and partial least squares regression. *Applied spectroscopy, 64*(1), 73-82.

Dixon, S. J., & Brereton, R. G. (2009). Comparison of performance of five common classifiers represented as boundary methods: Euclidean Distance to Centroids, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Learning Vector Quantization and Support Vector Machines, as dependent on data structure. *Chemometrics and Intelligent Laboratory Systems, 95*(1), 1-17.

Dong, J.-E., Wang, Y., Zuo, Z.-T., & Wang, Y.-Z. (2020). Deep learning for geographical discrimination of Panax notoginseng with directly near-infrared spectra image. *Chemometrics and Intelligent Laboratory Systems, 197*, 103913.

Géron, A. (2019). *Hands-on machine learning with scikit-learn, keras, and tensorflow: concepts*.

Gulli, A., & Pal, S. (2017). *Deep learning with Keras*: Packt Publishing Ltd.

Hosseinpour-Zarnaq, M., Omid, M., Sarmadian, F., & Ghasemi-Mobtaker, H. (2023). A CNN model for predicting soil properties using VIS–NIR spectral data. *Environmental Earth Sciences, 82*(16), 382.

Jiang, D., Qi, G., Hu, G., Mazur, N., Zhu, Z., & Wang, D. (2020). A residual neural network based method for the classification of tobacco cultivation regions using near-infrared spectroscopy sensors. *Infrared Physics & Technology, 111*, 103494.

Kabir, M. H., Guindo, M. L., Chen, R., & Liu, F. (2021). Geographic Origin Discrimination of Millet Using Vis-NIR Spectroscopy Combined with Machine Learning Techniques. *Foods, 10*(11), 2767.

Képeš, E., Vrábel, J., Adamovsky, O., Střítežská, S., Modlitbová, P., Pořízka, P., & Kaiser, J. (2022). Interpreting support vector machines applied in laser-induced breakdown spectroscopy. *Analytica Chimica Acta, 1192*, 339352.

Lu, Y., Wang, W., Huang, M., Ni, X., Chu, X., & Li, C. (2020). Evaluation and classification of five cereal fungi on culture medium using Visible/Near-Infrared (Vis/NIR) hyperspectral imaging. *Infrared Physics & Technology, 105*, 103206.

Manley, M. (2014). Near-infrared spectroscopy and hyperspectral imaging: non-destructive analysis of biological materials. *Chemical Society Reviews, 43*(24), 8200-8214.

Mat, K., Abdul Kari, Z., Rusli, N. D., Che Harun, H., Wei, L. S., Rahman, M. M., Mohd Khalid, H. N., Mohd Ali Hanafiah, M. H., Mohamad Sukri, S. A., Raja Khalif, R. I. A., Mohd Zin, Z., Mohd Zainol, M. K., Panadi, M., Mohd Nor, M. F., & Goh, K. W. (2022). Coconut Palm: Food, Feed, and Nutraceutical Properties. *Animals, 12*(16), 2107.

Ni, L.-J., Zhang, L.-G., Xie, J., & Luo, J.-Q. (2009). Pattern recognition of Chinese flue-cured tobaccos by an improved and simplified K-nearest neighbors

classification algorithm on near infrared spectra. *Analytica Chimica Acta, 633*(1), 43-50.

Osborne, B. G., Fearn, T., & Hindle, P. H. (1993). *Practical NIR spectroscopy with applications in food and beverage analysis*: Longman scientific and technical.

Ozaki, Y., Huck, C., Tsuchikawa, S., & Engelsen, S. B. (2021). *Near-Infrared Spectroscopy: Theory, Spectral Analysis, Instrumentation, and Applications*: Springer.

Pandiselvam, R., Mahanti, N. K., Manikantan, M. R., Kothakota, A., Chakraborty, S. K., Ramesh, S. V., & Beegum, P. P. S. (2022). Rapid detection of adulteration in desiccated coconut powder: vis-NIR spectroscopy and chemometric approach. *Food Control, 133*, 108588.

Rocha, W. F. d. C., Prado, C. B. d., & Blonder, N. (2020). Comparison of Chemometric Problems in Food Analysis using Non-Linear Methods. *Molecules, 25*(13), 3025.

Schütz, D., Riedl, J., Achten, E., & Fischer, M. (2022). Fourier-transform near-infrared spectroscopy as a fast screening tool for the verification of the geographical origin of grain maize (Zea mays L.). *Food Control, 136*, 108892.

Silva, C. S., Borba, F. d. S. L., Pimentel, M. F., Pontes, M. J. C., Honorato, R. S., & Pasquini, C. (2013). Classification of blue pen ink using infrared spectroscopy and linear discriminant analysis. *Microchemical Journal, 109*, 122-127.

Srinuttrakul, W., Mihailova, A., Islam, M. D., Liebisch, B., Maxwell, F., Kelly, S. D., & Cannavan, A. (2021). Geographical Differentiation of Hom Mali Rice Cultivated in Different Regions of Thailand Using FTIR-ATR and NIR Spectroscopy. *Foods, 10*(8), 1951.

Suksangpanomrung, P., Ritthiruangdej, P., Hiriotappa, A., & Therdthai, N. (2024). Rapid, non-destructive prediction of coconut composition for sustainable UHT milk production via near-infrared spectroscopy. *Journal of Food Composition and Analysis, 128*, 106009.

Sun, Y., Liu, N., Kang, X., Zhao, Y., Cao, R., Ning, J., Ding, H., Sheng, X., & Zhou, D. (2021). Rapid identification of geographical origin of sea cucumbers Apostichopus japonicus using FT-NIR coupled with light gradient boosting machine. *Food Control, 124*, 107883.

Szymczycha-Madeja, A., Welna, M., Jedryczko, D., & Pohl, P. (2014). Developments and strategies in the spectrochemical elemental analysis of fruit juices. *TrAC Trends in Analytical Chemistry, 55*, 68-80.

Teye, E., Huang, X., Dai, H., & Chen, Q. (2013). Rapid differentiation of Ghana cocoa beans by FT-NIR spectroscopy coupled with multivariate classification. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 114*, 183-189.

Tulashie, S. K., Amenakpor, J., Atisey, S., Odai, R., & Akpari, E. E. A. (2022). Production of coconut milk: A sustainable alternative plant-based milk. *Case Studies in Chemical and Environmental Engineering, 6*, 100206.

Workman Jr, J., & Weyer, L. (2007). *Practical guide to interpretive near-infrared spectroscopy*: CRC press.

Wu, D., Liu, X., Bai, B., Li, J., Wang, R., Zhang, Y., Deng, Q., Huang, H., & Wu, J. (2023). Determining farming methods and geographical origin of chinese rice using NIR combined with chemometrics methods. *Journal of Food Measurement and Characterization, 17*(4), 3695-3708.

Wu, Z., Shen, C., & van den Hengel, A. (2019). Wider or Deeper: Revisiting the ResNet Model for Visual Recognition. *Pattern Recognition, 90*, 119-133.

Xu, L., Zhou, Y.-P., Tang, L.-J., Wu, H.-L., Jiang, J.-H., Shen, G.-L., & Yu, R.-Q. (2008). Ensemble preprocessing of near-infrared (NIR) spectra for multivariate calibration. *Analytica Chimica Acta, 616*(2), 138-143.

Yan, T., Duan, L., Chen, X., Gao, P., & Xu, W. (2020). Application and interpretation of deep learning methods for the geographical origin identification of Radix Glycyrrhizae using hyperspectral imaging. *RSC advances, 10*(68), 41936-41945.

Zareef, M., Chen, Q., Hassan, M. M., Arslan, M., Hashim, M. M., Ahmad, W., Kutsanedzie, F. Y. H., & Agyekum, A. A. (2020). An Overview on the Applications of Typical Non-linear Algorithms Coupled With NIR Spectroscopy in Food Analysis. *Food Engineering Reviews, 12*(2), 173-190.

Zhang, X., Lin, T., Xu, J., Luo, X., & Ying, Y. (2019). DeepSpectra: An end-to-end deep learning approach for quantitative spectral analysis. *Analytica Chimica Acta, 1058*, 48-57.

Zhou, D., Yu, Y., Hu, R., & Li, Z. (2020). Discrimination of Tetrastigma hemsleyanum according to geographical origin by near-infrared spectroscopy combined with

a deep learning approach. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 238*, 118380.

Zou, L., Liu, W., Lei, M., & Yu, X. (2021). An Improved Residual Network for Pork Freshness Detection Using Near-Infrared Spectroscopy. *Entropy, 23*(10), 1293.

CHAPTER 7 – DEPLOYMENT MODEL

7.1 Deployment Model from Chapter 3

In Chapter 3, model development along with automatic preprocessing and tuning hyperparameter from ML using data acquired on 13 December 2021 for fresh coconut milk (FCM) and instant coconut milk (ICM), and 23 March 2022 for samples of fresh coconut milk (FCM) and coconut milk adulterated by distilled water (ADW) with levels 10, 20, 30, 40, 50, 60, 70, 80, 90% (w/w). From this case study, we generated a discrimination model to classify coconut milk, including FCM, ICM, and ADW. Furthermore, this case study also generates a calibration model to predict the adulteration level of distilled water in fresh coconut milk in the range of 10 to 90%, with a uniform increase of 10%. Each model generated by the ML algorithm (classifier and regressor) will be saved in *.Joblib file so that it can be recalled during the deployment testing process using an unknown dataset.

7.1.1 Classification

An unknown dataset for deployment testing of the Chapter 3 discrimination model is described in Table 7.1. The dataset comes from cross-over data in other chapters, including Chapters 4, 5, and 6, and data acquired specifically for this test. This discrimination model was generated using fresh coconut milk (FCM), instant coconut milk (ICM), and adulterated fresh coconut milk by distilled water (ADW) data using FT-NIR. Unknown data as a cross-over from another chapter in this work have been filtered to deal with the discrimination model that was developed. For example, the discrimination model in Chapter 3 only classifies samples with FCM, ICM, and ADW using FT-NIR. ADW samples acquired on 20 May 2023 consisted of coconut milk samples that were adulterated with distilled water at levels 10, 20, 30, 40, and 50% (w/w), and samples acquired on 27 January 2024 consisted of samples that were adulterated at level 10 and 50 % (w/w). Besides, the unknown dataset for this chapter consists of different multi-sessions because it comes from several experimental timelines from 2023 to 2024.

The three classifier algorithms from ML (LDA, SVM MLP) in Chapter 3 produce the same good performance of the discrimination model. Therefore, the 3 ML will be tested using the unknown dataset. Visualization of the unknown dataset FT-NIR

spectra for testing the discrimination model, which has been adjusted to the preprocessing of each classifier, is presented in Figure 7.1.

Table 7.1. NIRs information for classification model deployment in Chapter 3.

Data from	Date	FT-NIR	
		Class	Total
Chapter 4	20 May 2023	FCM	10
		ADW	50
Chapter 5	6 January 2024	FCM	90
Chapter 6	27 January 2024	FCM	90
-	27 January 2024	ADW	60

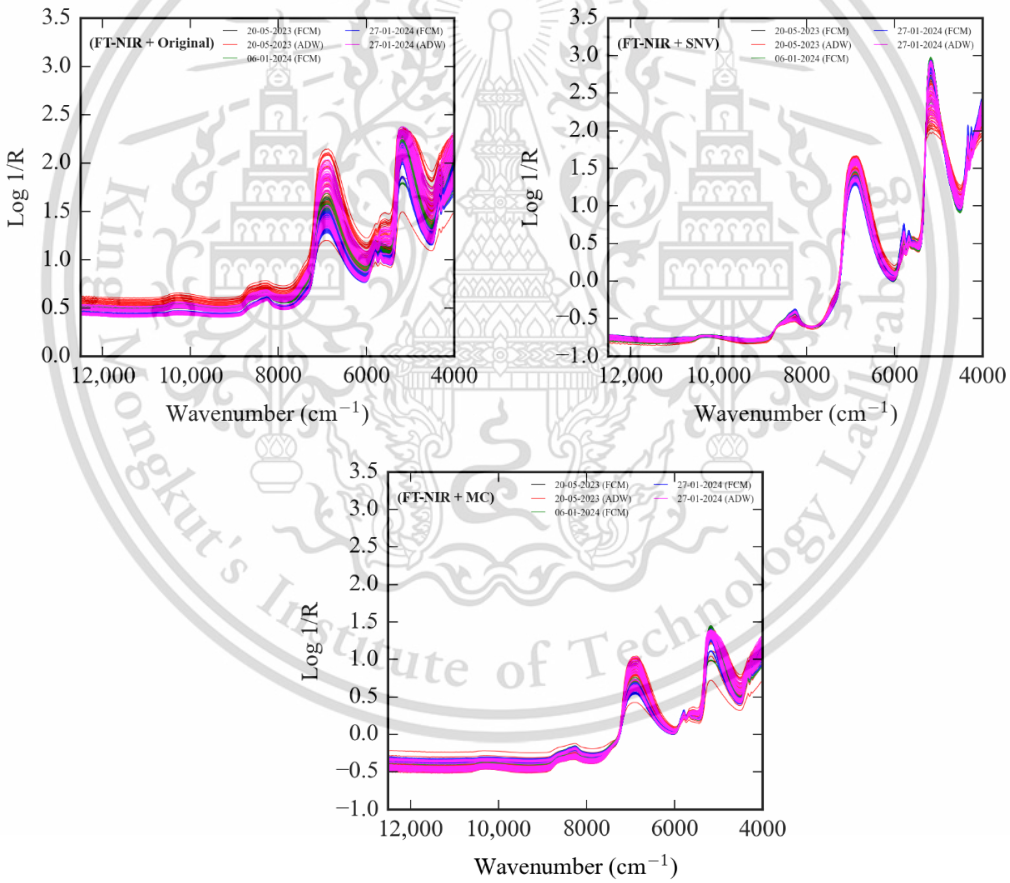
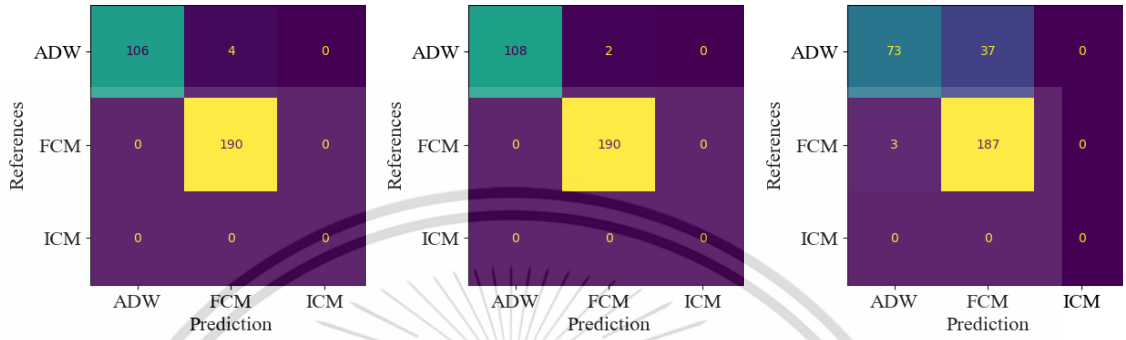


Figure 7.1. NIRs for deployment of a classification model in Chapter 3.

The performance results of the discrimination model generated from Chapter 3 for each ML algorithm in a confusion matrix are presented in Figure 7.2. All the ML classifiers used in the chapter can perfectly identify FCM samples from 3 different

sessions except the MLP classifier, where the MLP classifier identifies 3 FCM samples adulterated by distilled water. However, to identify ADW samples, of the 110 samples used and coming from 2 different sessions, each LDA, SVM, and MLP algorithm found misprediction of 4, 2, and 37 samples, respectively.



FT-NIR + LDA + Original FT-NIR + SVM + SNV FT-NIR + MLP + MC

Figure 7.2. The result of deploying a classification model in Chapter 3.

7.1.2 Regression

An unknown dataset for deployment testing of the Chapter 3 calibration model is described in Table 7.2. In this chapter, as much as 1 calibration model is generated to predict the adulteration level of coconut milk from distilled water using FT-NIR. This calibration model was generated using adulteration levels of 10, 20, 30, 40, 50, 60, 70, 80, and 90% (w/w) from distilled water and acquired using FT-NIR. The unknown dataset comes from the cross-over data in Chapter 4 and data acquired specifically for this test. Cross-over data from other chapters has been filtered to fit the calibration model developed. For example, the Chapter 3 calibration model only uses deployment testing to predict adulteration levels of distilled water (ADW) from 10, 20, 30, 40, and 50% (w/w). Besides, the unknown dataset for deployment testing in this chapter consists of the experiment timeline for 2023 and 2024.

Of the three ML regressor algorithms (LDA, SVM, MLP) in Chapter 3, it shows that the SVM regressor produces the best calibration model performance. Therefore, the SVM regressor will be tested using the unknown dataset. Visualization of the unknown dataset FT-NIR spectra for deployment testing the calibration model, which has been adjusted to the preprocessing of the regressor, is presented in Figure 7.3.

Table 7.2. NIRs information for regression model deployment in Chapter 3.

Data from	Date	FT-NIR		
		Class	Adulteration level (%)	Total
Chapter 4	20 May 2023	ADW	10	10
			20	10
			30	10
			40	10
			50	10
-	27 January 2024	ADW	10	30
			50	30

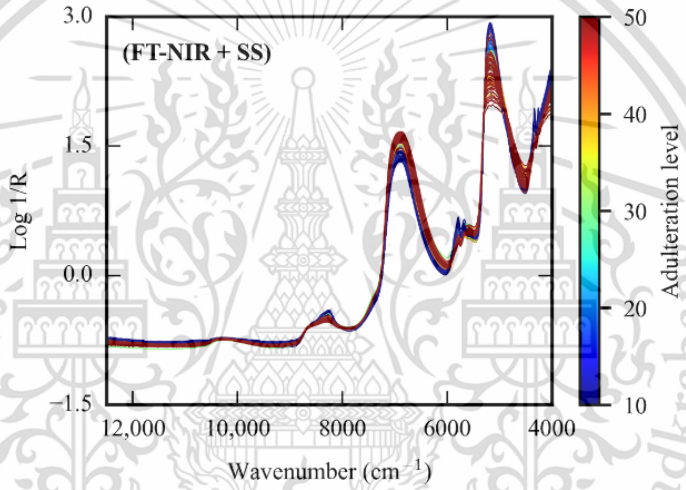


Figure 7.3. Spectra used for deployment of a regression model in Chapter 3.

The performance results of the calibration model generated from Chapter 3 for the FT-NIR dataset using SVM regressors in a scatter plot are presented in Figure 7.4. It can be seen that the SVM regressor with the unknown data set FT-NIR can predict the adulteration level of distilled water in coconut milk with accuracy via the R^2 and RMSE evaluators of 96.9% and 3.099%, respectively. In other words, the performance of on-deployment testing using FT-NIR with SVM regressor was between training (R^2=98%, RMSE=3.74%) and testing (R^2=93%, RMSE=8.30%).

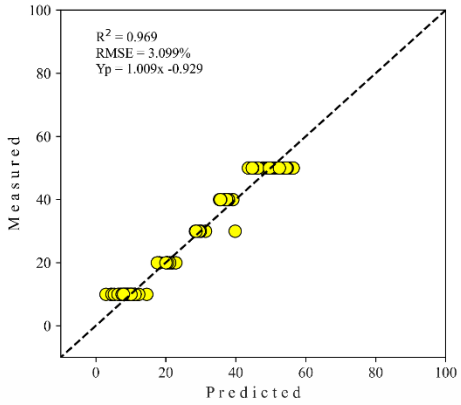


Figure 7.4. Result from the deployment of a regression model in Chapter 3.

7.2 Deployment Model from Chapter 4

In Chapter 4, model development with automatic multi-preprocessing and tuning hyperparameter from ML using data acquired on 20 May 2023 for fresh coconut milk (FCM), coconut milk adulterated by distilled water (ADW) and coconut milk adulterated by mature coconut water (ACW) with levels 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 30, 40, and 50% (w/w). We generated a discrimination model from this case study to classify coconut milk, including FCM, ADW, and ACW. Furthermore, this case study also generates 2 calibration models, first to predict the adulteration level of distilled water in coconut milk, and second to predict the adulteration level of mature coconut water in coconut milk in the range of 1 to 50% with non-uniform increment. The 3 models (1 discrimination model and 2 calibration models) generated by the ML algorithm will be saved in *.Joblib file so that they can be recalled in the deployment testing using an unknown dataset.

7.2.1 Classification

An unknown dataset for deployment testing of the Chapter 4 discrimination model is described in Table 7.3. The dataset comes from a cross-over of data in other chapters, including Chapters 3, 5, and 6, and a dataset acquired specifically for this test. This discrimination model was generated using FCM, ADW, and ACW data using FT-NIR and Micro-NIR. Unknown data, as a cross-over dataset from another chapter in this work, has been filtered to fit with the discrimination model that was developed. For example, the discrimination model in Chapter 4 only classifies FCM, ADW, and ACW samples using FT-NIR and Micro-NIR. Besides, the unknown dataset

for this chapter consists of multiple sessions because it comes from several experimental timelines starting from 2021, 2022, and 2024.

Table 7.3. NIRs information for classification model deployment in Chapter 4.

Data from	Date	FT-NIR		Micro-NIR	
		Class	Total	Class	Total
Chapter 3	13 December 2021	FCM	10	-	-
	23 March 2022	FCM	10	-	-
		ADW	50	-	-
Chapter 5	6 January 2024	FCM	90	FCM	90
Chapter 6	27 January 2024	FCM	90	FCM	90
-	27 January 2024	ADW	90	ADW	90
		ACW	90	ACW	90

Of the three ML classifier algorithms (LDA, KNN MLP) in Chapter 4, the best discrimination model performance is the MLP classifier for the FT-NIR dataset and the KNN classifier for the Micro-NIR dataset. Therefore, the 2 ML classifiers will be tested using the unknown dataset. Visualization of the unknown dataset of FT-NIR and Micro-NIR spectra for deployment testing of the discrimination model, which has been adapted to the preprocessing of each classifier, is presented in Figure 7.5.

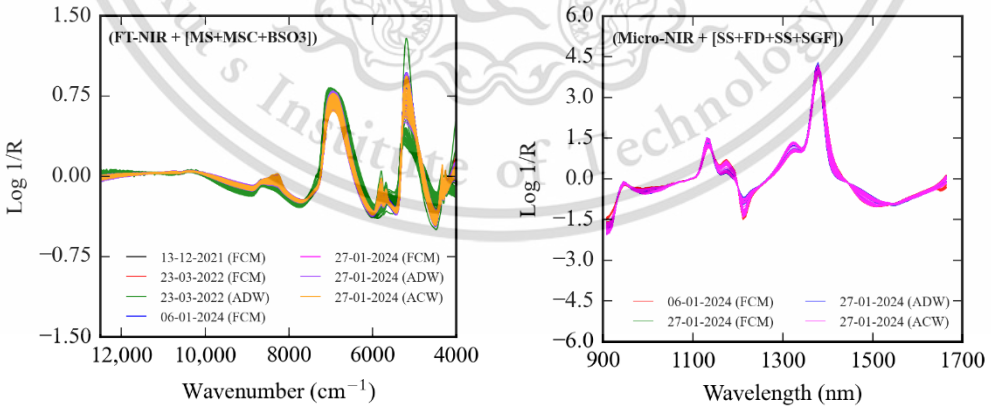


Figure 7.5. NIRs for deployment of a classification model in Chapter 4.

The performance results of the discrimination model generated from Chapter 4 for the FT-NIR dataset using the MLP classifier for the Micro-NIR dataset using the KNN classifier in the confusion matrix are presented in Figure 7.6. It can be seen that

the MLP classifier with the FT-NIR data set can correctly identify 82, 123, and 165 ACW, ADW, and FCM samples from a total of 90 ACW samples, 140 ADW samples, and 200 FCM samples, respectively. In other words, the accuracy of deployment testing using FT-NIR with MLP classifier is 86.05%, decreasing performance compared to training (accuracy of 90%) and testing (accuracy of 92%). The KNN classifier with the Micro-NIR dataset can correctly identify 90/90 ACW, 84/90 ADW, and 180/180 FCM samples. In other words, accuracy on deployment testing using Micro-NIR with KNN classifier is 98.33%, decreasing performance compared to training (accuracy of 100%) and higher compared to testing (accuracy of 97%).

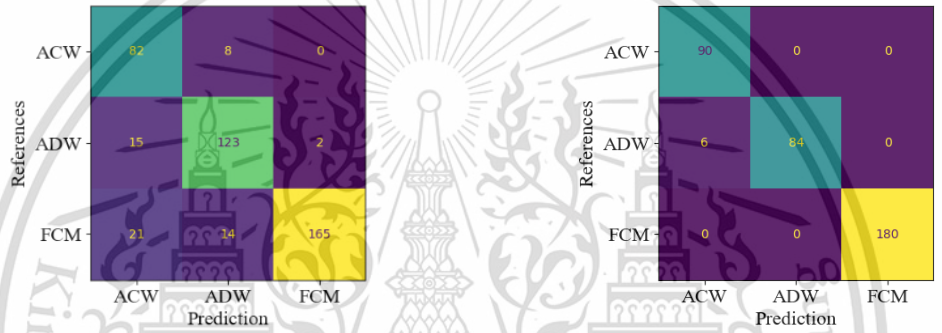


Figure 7.6. The result of deploying a classification model in Chapter 4.

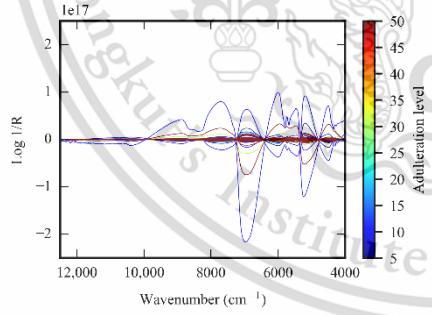
7.2.2 Regression

An unknown dataset for deployment testing of the Chapter 4 calibration model is described in Table 7.4. In this chapter, as much as 2 calibration models are generated, including a calibration model for the case of predicting the adulteration level of coconut milk from distilled water and mature coconut water using FT-NIR and Micro-NIR. This model was generated using adulteration levels of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 30, 40, and 50% (w/w). The unknown dataset here comes from the cross-over dataset from Chapter 3 and a dataset acquired specifically for this test. Cross-over data from other chapters has been filtered to fit the calibration model developed. For example, the dataset from Chapter 3 for deployment testing is data with adulteration levels of 5, 10, 20, 30, 40, and 50% (w/w) from distilled water and mature coconut water using FT-NIR. Besides, the unknown dataset for deployment testing in this chapter consists of the experiment timeline for 2022 and 2024.

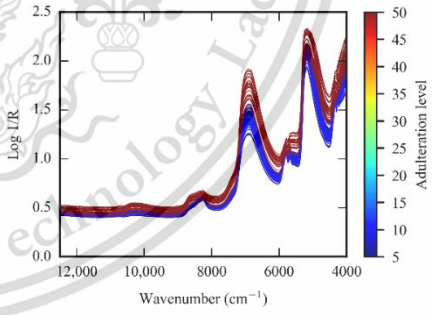
Table 7.4. NIRs information for regression model deployment in Chapter 4.

Data from	Date	FT-NIR				Micro-NIR			
		ADW		ACW		ADW		ACW	
		Adulteration level (%)	Total	Adulteration level (%)	Total	Adulteration level (%)	Total	Adulteration level (%)	Total
Chapter 3	23 March 2022	10	10	-	-	-	-	-	-
		20	10	-	-	-	-	-	-
		30	10	-	-	-	-	-	-
		40	10	-	-	-	-	-	-
		50	10	-	-	-	-	-	-
-	27 January 2024	5	30	5	30	5	30	5	30
		10	30	10	30	10	30	10	30
		50	30	50	30	50	30	50	30

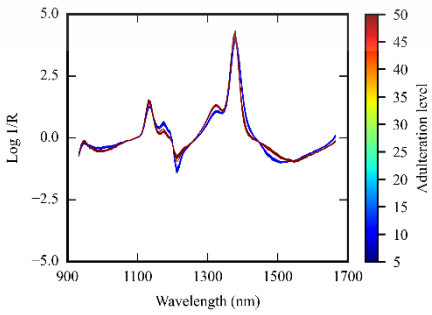
Of the three ML regressor algorithms (PLS, KNN, MLP) in Chapter 4, it shows that the PLS regressor produces the best performance of the calibration model to predict the case of ADW using the FT-NIR dataset and the KNN regressor to predict the case of ACW using the Micro-NIR dataset. Therefore, the PLS and KNN regressors will be tested using unknown datasets. Visualization of the unknown dataset FT-NIR and Micro-NIR spectra for deployment testing the calibration model, which has been adjusted to the preprocessing of the regressor, are presented in Figure 7.7.



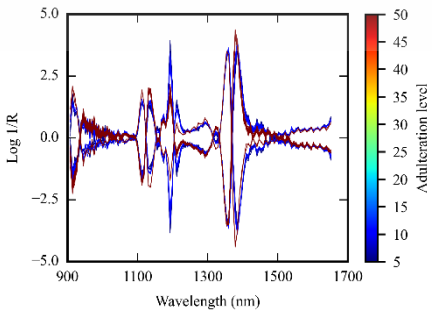
ADW + FT-NIR + [MS+SNV+MS+BSO3] + PLSR



ACW + FT-NIR + [SGF+SGF+SGF+SGF] + PLSR



ADW + Micro-NIR + [FD+MS+SD+SNV] + KNN



ACW + Micro-NIR + [BSO3+MS+SD+SS] + KNN

Figure 7.7. Spectra used for deployment of a regression model in Chapter 4.

The performance results of the calibration model generated from Chapter 4 for the FT-NIR and Micro-NIR datasets using ML regressors in a scatter plot are presented in Figure 7.8. It can be seen that the PLSR regressor with the unknown data set FT-NIR can predict the adulteration level of distilled water in coconut milk with accuracy through the R^2 and RMSE evaluators of 82.8% and 7.738%, respectively. In other words, the performance on deployment testing using FT-NIR with PLSR regressor is lower than during training (R^2=99.5%, RMSE=1.042%) and testing (R^2=98.2%, RMSE=1.988%). Meanwhile, to predict the level of adulteration of mature coconut water in coconut milk, the PLSR regressor has an R^2 performance of 93.6% and an RMSE of 5.081% or lower than during training (R^2=98.8%, RMSE=1.534%) and testing (R^2=95.2%, RMSE=3.221%).

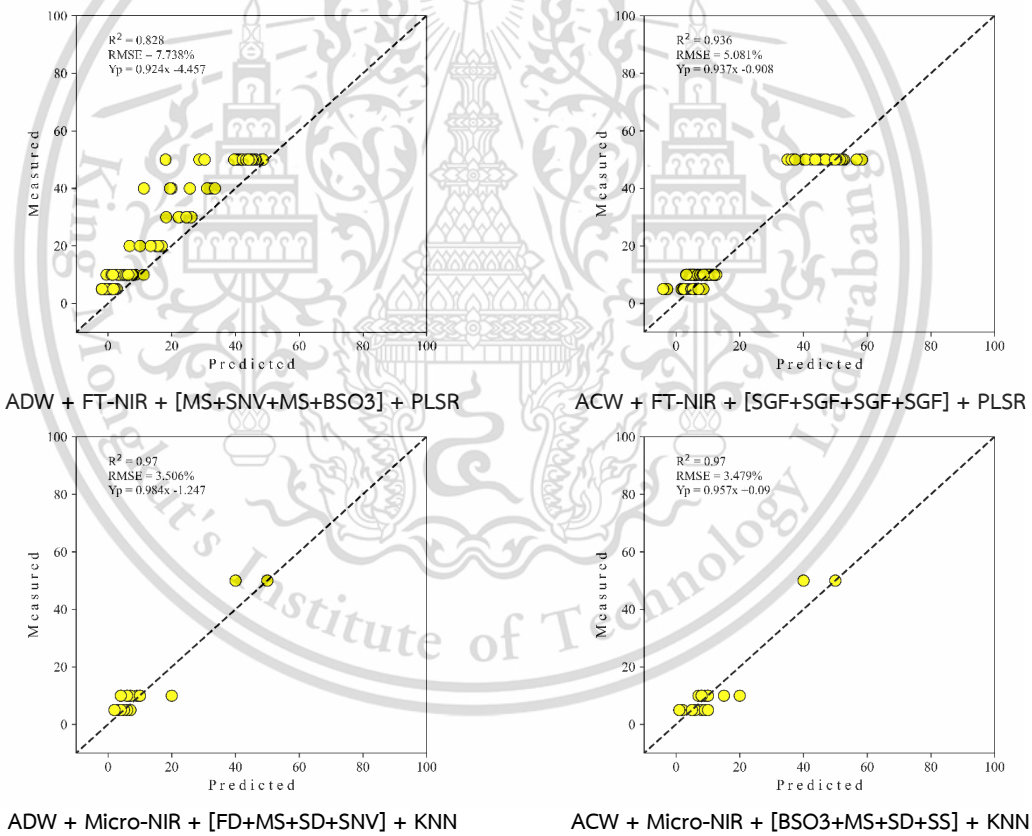


Figure 7.8. Result from the deployment of a regression model in Chapter 4.

For cases using Micro-NIR, the KNN regressor with an unknown dataset from Micro-NIR can predict the adulteration level of distilled water in coconut milk with accuracy via R^2 and RMSE of 97.0% and 3.506%, respectively. The performance on

deployment testing using Micro-NIR with KNN regressor is lower than during training (R^2=100%, RMSE=0%) and testing (R^2=99.5%, RMSE=0.962%). Meanwhile, to predict the level of adulteration of mature coconut water in coconut milk, the KNN regressor has an R^2 performance of 97.0% and an RMSE of 3.479% or lower than during training (R^2=100%, RMSE=0%) and testing (R^2=99.6%, RMSE=0.908%).

7.3 Deployment Model from Chapter 5 – Regression

An unknown dataset for deployment testing of the Chapter 5 calibration model is described in Table 7.5. This chapter generates a calibration model to predict the adulteration level of coconut milk from corn flour (CF) and tapioca starch (TS) using FT-NIR and Micro-NIR. This model was generated using adulteration levels of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 30, 40, and 50% (w/w). The calibration model generated by the DL algorithm will be saved in a *.h5 file so it can be recalled in the deployment testing process using an unknown dataset. The unknown dataset comes from data acquired specifically for this test on 27 January 2024. In Chapter 5, the calibration model uses deployment testing to predict adulteration levels from levels 5, 10, and 50% (w/w) of CF and TS. Besides, the unknown dataset for deployment testing in this chapter only consists of 1 experiment timeline (2024).

Table 7.5. NIRs information for regression model deployment in Chapter 5.

Data from	Date	FT-NIR				Micro-NIR			
		ADW		ACW		ADW		ACW	
		Adulteration level (%)	Total	Adulteration level (%)	Total	Adulteration level (%)	Total	Adulteration level (%)	Total
-	27 January 2024	5	30	5	30	5	30	5	30
		10	30	10	30	10	30	10	30
		50	30	50	30	50	30	50	30

Of the four DL regressor architectures (S-CNN, S-AlexNET, ResNET, GoogleNET) in Chapter 5 shows that the regressor from the GoogleNET architecture produces the best calibration model performance for predicting the case of adulteration of coconut milk with corn flour and tapioca starch using the FT-NIR dataset. When using the Micro-NIR dataset, the regressor from the GoogleNET architecture was the best in predicting the level of corn flour adulteration in coconut milk, and the ResNET

architecture was the best for the case of tapioca starch. Therefore, the next regressors will be tested using the unknown dataset. Visualization of the unknown dataset FT-NIR and Micro-NIR spectra for testing the calibration model with SNV preprocessing is presented in Figure 7.9.

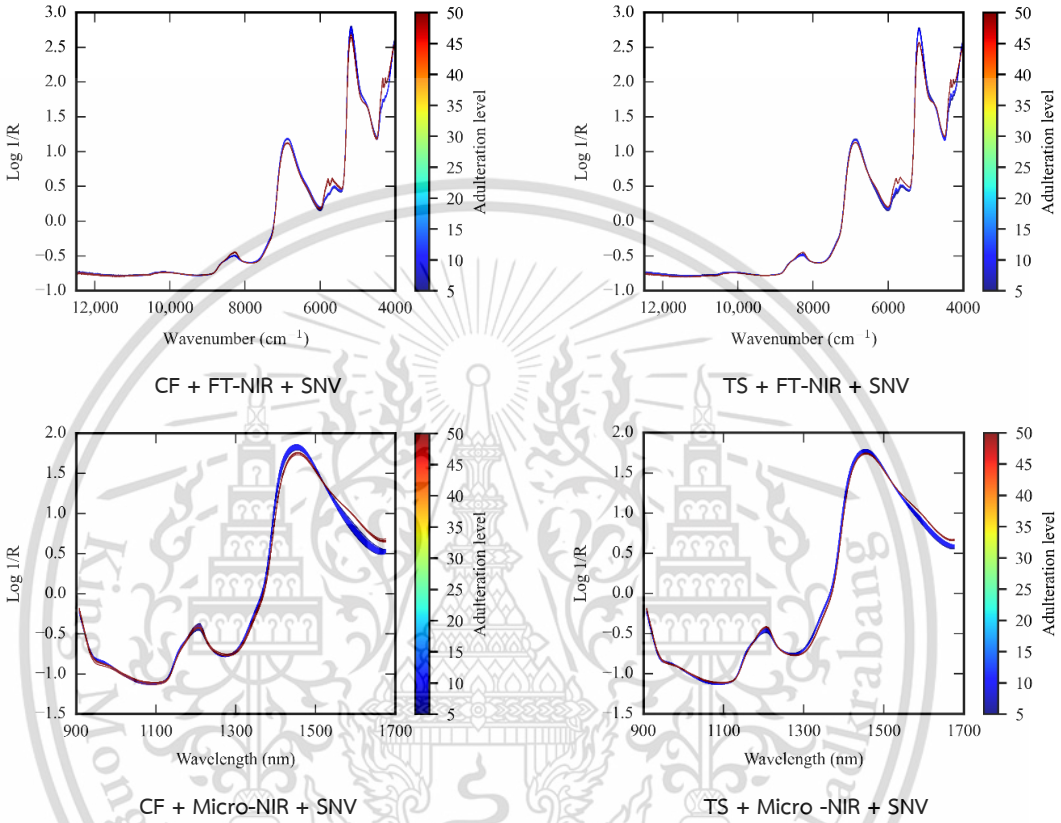


Figure 7.9. Spectra used for deployment of a regression model in Chapter 5.

The performance results of the calibration model generated from Chapter 5 for the FT-NIR and Micro-NIR datasets using the DL regressor in a scatter plot are presented in Figure 7.10. It can be seen that the regressor of GoogleNET with the FT-NIR unknown dataset can predict the level of corn flour adulteration in coconut milk with accuracy through the R^2 and RMSE evaluators of 96.6% and 3.736%, respectively. In other words, the performance on deployment testing using FT-NIR with regressor GoogleNET architecture is slightly lower than during training (R^2=99.8%, RMSE=0.601%) and testing (R^2=99.8%, RMSE=0.686%). Meanwhile, to predict the level of tapioca starch adulteration in coconut milk, the regressor of GoogleNET has

an R^2 performance of 95.4% and an RMSE of 4.334% or lower than during training (R^2=99.9%, RMSE=0.482%) and testing (R^2=99.8%, RMSE=0.670%).

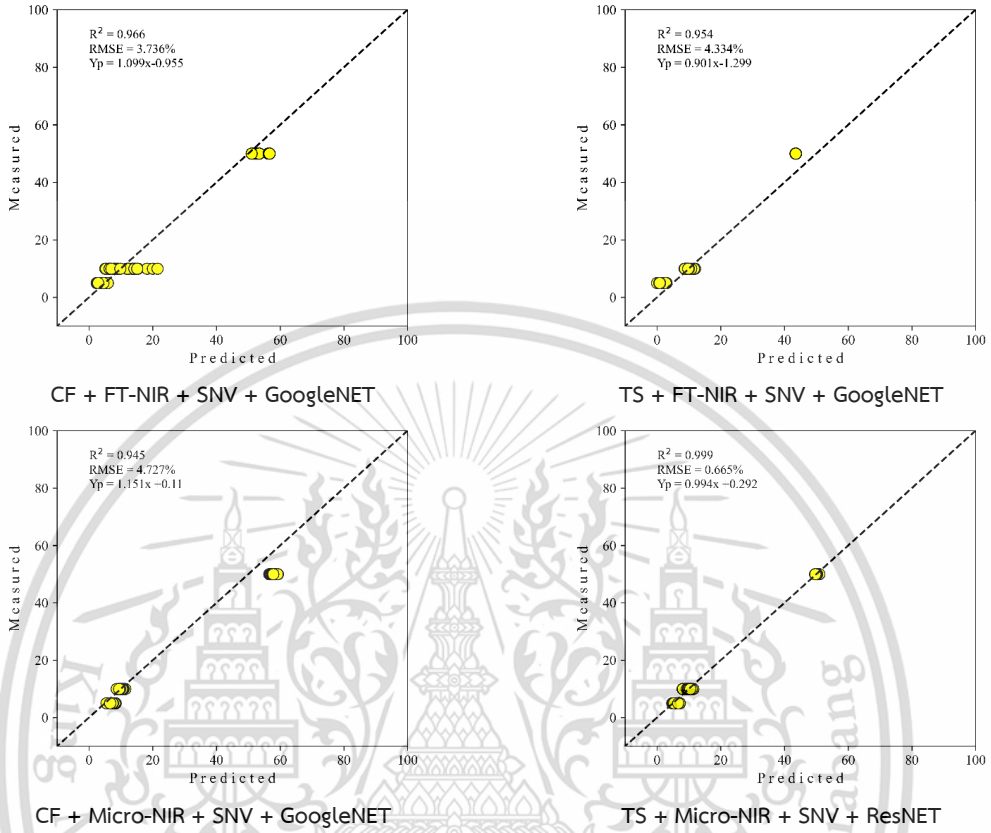


Figure 7.10. Result from the deployment of a regression model in Chapter 5.

For cases using Micro-NIR, the GoogleNET architecture regressor with an unknown dataset from Micro-NIR can predict the level of corn flour adulteration in coconut milk with accuracy via the R^2 evaluator and RMSE of 94.5% and 4.727%, respectively. The performance on deployment testing using Micro-NIR with regressor GoogleNET architecture is slightly lower than during training (R^2=99.9%, RMSE=0.414%) and testing (R^2=99.9%, RMSE=0.463%). Meanwhile, to predict the level of tapioca starch adulteration in coconut milk, the ResNET architecture regressor has an R^2 performance of 99.9% and an RMSE of 0.665% or almost as good as during training (R^2=99.9%, RMSE=0.431%) and testing (R^2=99.9%, RMSE=0.461%).

7.4 Deployment Model from Chapter 6 – Classification

In Chapter 6, model development comes from classical chemometrics (CC), ML, and DL approaches with minimum preprocessing using data acquired on 27

January 2024 for coconut milk originating from 3 provinces, including Chumphon (CHP), Samut Songkhram (SSK), and Chonburi (CHB) using FT-NIR and Micro-NIR. From this case study, we generated a discrimination model to classify coconut milk based on its origin, including CHB, CHP, and SSK. Furthermore, this case study generated 6 best discrimination models, 3 models each from the CC, ML, and DL groups and 2 models from 2 instruments, including FT-NIR and Micro-NIR. The classifier model generated from CC and ML will be saved in the form of a *.Joblib file, and the classifier model generated from DL will be stored in the form of a *.h5 file so that it can be recalled in the deployment testing process using unknown dataset.

An unknown dataset for deployment testing of the regional discrimination model of coconut milk in Chapter 6 is described in Table 7.6. The dataset comes from cross-over data in other chapters, including Chapters 3, 4, and 5. Cross-over data from other chapters has been filtered to fit the discrimination model developed. The unknown dataset for this chapter consists of different multi-sessions because it comes from several experimental timelines (2021 to 2024).

Table 7.6. NIRs information for classification model deployment in Chapter 6.

Data from	Date	FT-NIR		Micro-NIR	
		Class	Total	Class	Total
Chapter 3	13 December 2021	CHP	10	-	-
	23 March 2022	CHP	10	-	-
Chapter 4	20 May 2023	CHP	10	CHP	10
Chapter 5	6 January 2024	CHP	90	CHP	90

The three best classifier algorithms from CC, ML, and DL for the FT-NIR dataset in Chapter 6 are represented by PCA, SVM, and S-CNN with SNV preprocessing for all classifiers. Meanwhile, the Micro-NIR dataset is represented by LDA and KNN with SNV preprocessing and ResNET with BSO preprocessing. Visualization of the unknown dataset FT-NIR and Micro-NIR spectra for testing the discrimination model, which has been adapted to the preprocessing of each classifier, is presented in Figure 7.11.

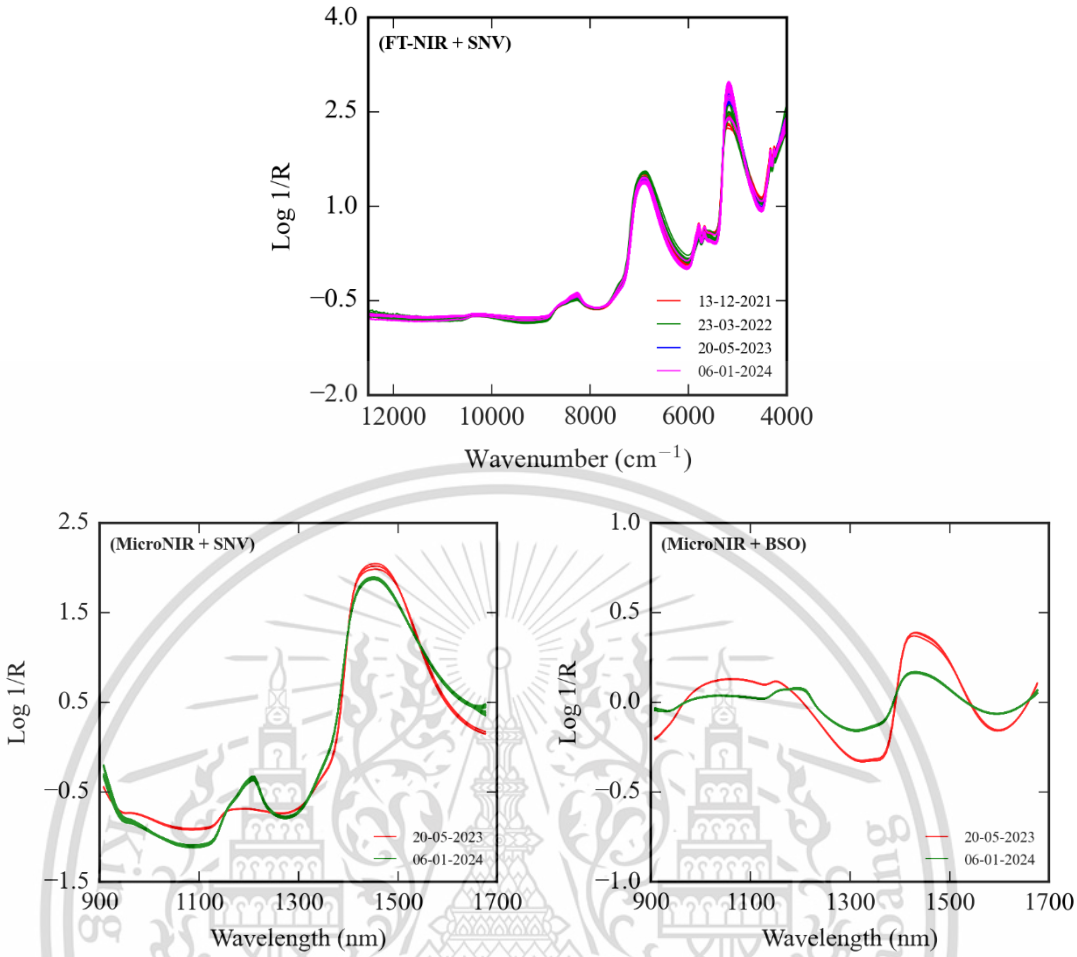


Figure 7.11. NIRs for deployment of a classification model in Chapter 6.

The performance results of the discrimination model generated from Chapter 6 using the FT-NIR and Micro-NIR dataset in the confusion matrix are presented in Figure 7.12. It can be seen that when using the FT-NIR dataset, each of the PCA, SVM, and S-CNN classifiers can predict 107, 118, and 199 samples from the 120 CHP of samples. Meanwhile, when using the Micro-NIR dataset, all the best classifiers from each class can discriminate 100 CHP samples perfectly, except for the ResNET classifier, which can predict 97 samples from the 100 CHP samples fed.

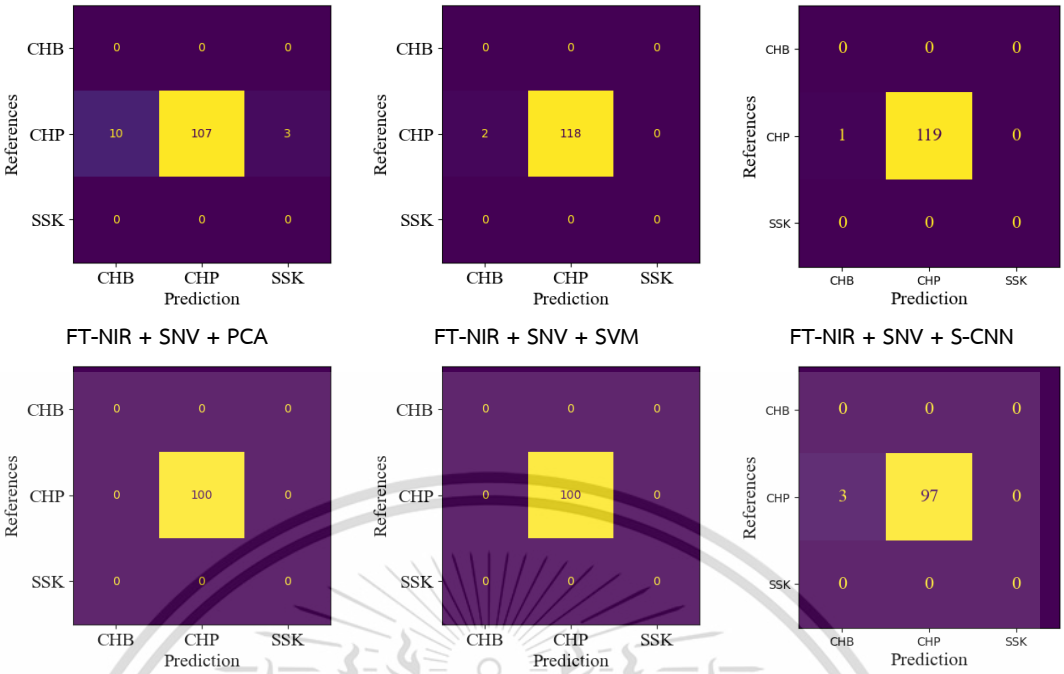


Figure 7.12. The result of deploying a classification model in Chapter 6.

CHAPTER 8 – CONCLUSIONS, RECOMMENDATIONS AND FUTURE WORKS

8.1 Conlusions

The need to process NIR spectroscopy-based data to generate a fast and robust prediction model has led to exploration efforts in the area of artificial intelligence. The development of machine learning and deep learning, which are part of artificial intelligence, has been studied for some time and is claimed to be suitable for development as a method for modeling various purposes with NIR spectroscopy descriptors. It is believed that NIR spectroscopy, which can carry out data acquisition quickly and non-destructively, will work wonders with the advances in modeling techniques currently available.

Previously, NIR spectroscopy was very dependent on chemometric techniques, which involved several steps, including preprocessing, due to the noise that came with the NIR data. Therefore, many types of preprocessing and their variants have been developed, depending on the presence of noise that will be encountered. This results in the need for methods and expertise to be able to adjust the noise present. In addition, several popular software/libraries widely used in NIR spectroscopy data processing show that the preprocessing and optimization processes of the modeling algorithm are always carried out in separate conditions. Therefore, based on the findings of this thesis report, several leading algorithms from machine learning and deep learning adapted to FT-NIR and Micro-NIR dataset conditions are suitable for combining preprocessing and hyperparameter optimization of these algorithms.

To prove this concept, testing was carried out on the FT-NIR and Micro-NIR datasets from cases of coconut milk adulteration. The object of adulteration in agricultural and food products was chosen because, based on the results of a literature review, this has the potential to occur in these products in the future. Moreover, unsatisfactory results were also seen in this study when NIRs detected tiny adductors. This is possible due to the limitations of the available modeling approach itself.

For the first qualitative study, case samples were used from fresh coconut milk (FCM), instant coconut milk (ICM), and coconut milk that had been adulterated with

distilled water (A-FCM or ADW) and acquired using FT-NIR. The second qualitative model uses datasets acquired using FT-NIR and Micro-NIR from fresh coconut milk (FCM), coconut milk adulterated with distilled water (ADW), and coconut milk adulterated with mature coconut water (ACW) at several non-uniform levels. The third qualitative model uses coconut milk from 3 different provincial areas (Chumphon, Samut Songkhram, and Chonburi Province) and is acquired using FT-NIR and Micro-NIR. Meanwhile, for the quantitative study case, 4 NIR datasets were used from coconut milk adulteration at various levels, including distilled water (ADW), mature coconut water (ACW), corn flour (CF), and tapioca starch (TS).

The performance of the machine learning algorithm, which has been modified to work automatically to select proper preprocessing for itself along with its hyperparameters, is reported in Chapter 3 and Chapter 4. The difference is that Chapter 3 focuses on concept development by selecting a single appropriate preprocessing with the FT-NIR dataset. In contrast, Chapter 4 involves multiple preprocessing and its hyperparameters with datasets from FT-NIR and Micro-NIR. The classifier algorithms explored include LDA, SVM, MLP, and KNN with FCM, ICM, and A-FCM datasets. All classifiers (LDA, SVM, MLP) using FT-NIR obtained the same satisfactory results with all the precision, recall, F1-score, and perfect accuracy (100%) in both calibration and prediction with single preprocessing. The same performance was also found when multiple preprocessing methods, such as FT-NIR and Micro NIR, were used to classify FCM, ADW, and ACW.

Regressors from machine learning that are explored include PLS, SVM, MLP, and KNN. In case regression A-FCM using FT-NIR and single preprocessing, SVM obtained acceptable results, with a determination coefficient of calibration and prediction all over 0.93, root mean square error of calibration and prediction all below 8.30%, and ratio of prediction to deviation over 3.80. The highest regression model performance to predict ADW level using FT-NIR is to combine the PLS regressor (LVs of 11) and a quadruple-preprocessing step. The KNN regressor supported by quadruple-preprocessing steps and one n-neighbor as a hyperparameter is the best combination by the Mirco NIR. For the ACW case, the highest regression model performance using FT-NIR is to combine the PLS regressor (with 9 LVs) and a quadruple-preprocessing step. Besides, the KNN regressor

supported by quadruple-preprocessing steps and one n-neighbor of hyperparameter is the best using the Mirco-NIR. The performance strategy proposed in Chapters 3 and 4 effectively addressed and produced satisfactory outcomes in classifying and regression challenges and problems from coconut milk adulteration. These results demonstrate that our proposed approach can more deeply discover the best classification and regression model, particularly for the coconut milk adulteration NIR spectroscopy.

The application only uses standardization for preprocessing (SNV preprocessing) to perform modeling for NIR spectroscopy datasets, and it has also been explored using algorithms from deep learning. By using 4 types of deep learning architecture (CNN, S-AlexNET, ResNET, and GoogleNET) and 2 NIR instruments (FT-NIR and Micro-NIR), the adulteration level of coconut milk from corn flour (CF) and tapioca starch (TS) can be predicted well. The results confirmed the feasibility of deep learning algorithms for predicting the degree of coconut milk adulteration with reliable performance (R^2 of 0.886–0.999, RMSE of 0.370–6.108%, and Bias of −0.176–1.481). Furthermore, the RPD of all algorithms with all types of NIR spectrophotometers indicates an excellent capability for quantitative predictions for any application (RPD > 8.1) except for case predicting tapioca starch, using FT-NIR by ResNET (RPD < 3.0). This chapter demonstrated the feasibility of using deep learning algorithms and NIR spectral data as a rapid, accurate, robust, and non-destructive way to evaluate coconut milk adulterants. Micro-NIR is more promising than FT-NIR in predicting coconut milk adulteration from solid adulterants, and it is portable for in situ measurements in the future.

Finally, there is a qualitative case when using classical chemometrics (CC), machine learning, and deep learning algorithms to detect the origin of coconut milk. Our findings showed that a classifier from the machine learning (SVM) and deep learning (ResNET) groups could yield the optimal performance for discriminating the geographical source area of coconut milk, with an accuracy of 99.1% for the training and 100% for the testing using FT-NIR. Furthermore, when using Micro-NIR, the classifier from group classical (LDA), machine learning (SVM, KNN), and deep learning (ResNET) delivered the highest accuracy of 99.5% for the training and 100% for the testing. This chapter concluded that both FT-NIR and Micro-NIR, supported by

classical to modern chemometric classifiers, could be used to evaluate the geographical area source from coconut milk. Also, the method in this chapter includes a strategy for discovering feature-important NIR spectra for interpretability purposes, thereby facilitating the qualitative interpretation of results for all types of classifiers.

8.2 Recommendations and Future Works

A practical application of the concept to make it uncomplicated to carry out proper preprocessing and proper hyperparameter optimization of machine learning algorithms for NIR data processing has been successfully achieved in this thesis. The first way is to combine automation between preprocessing and hyperparameters from machine learning algorithms, both for qualitative and quantitative cases. The second way is to use a deep learning algorithm that does not require or at least minimal use of preprocessing. These two solutions will help NIR practitioners or chemometricians reduce the burden of finding the best model for their dataset. Keep in mind that the complexity and computational cost of finding the proper preprocessing and proper hyperparameter optimized for a particular problem is very high, as several tests and trials have to be performed. However, once both were discovered, implementation and training became effortless. Developing proper preprocessing and proper hyperparameters and applying them to some problems are two different research subjects that should not be confused.

On one side, the idea concept of combining to find proper preprocessing automation with hyperparameters still needs improved display so that the preprocessing and hyperparameter automation concept of ML can be more user-friendly and should involve the graphical user interface (GUI). The developed GUI should implement automated preprocessing with hyperparameters from machine learning that is more accessible so that it can also be used by practitioners who do not have programming skills or are unfamiliar with the Python programming language. Also, the option to add more variety of preprocessing should be prepared in this GUI.

On the other hand, coconut milk adulteration samples from various materials and origins are used in this study; increasing the quantity and diversity of the dataset to a more significant extent is needed in the future to produce a more robust model

before it is released to practical application. Currently, the discrimination and calibration models developed are individual models, not global ones. In other words, the resulting discrimination and calibration models only work with terms and conditions according to the individual conditions in which the model was developed. When the dataset that has been collected reaches high diversity and quantity in the future, the development of NIR-based global models that can detect various types and levels of adulteration has the potential to be developed further.

***

Appendix

International Published Papers

Author Biography

Appendix 1. Supplementary Material Chapter 1 – General Introduction

Table S1-1. Effect of type of liquid adulteration and level of adulteration to moisture content of coconut milk with ANOVA.

Source of variation	Sum of square	DoF	Mean square	P-value	P<0.05
Adulteration level	2067.96	3	689.32	0.0000	Sig.
Adulterant type	4.28	1	4.28	0.0399	Sig.
Interaction	19.21	3	6.40	0.0024	Sig.
Within	13.68	16	0.85		
Total	2105.12	23			

Table S1-2. Effect of type of liquid adulteration and level of adulteration to fat content of coconut milk with ANOVA.

Source of variation	Sum of square	DoF	Mean square	P-value	P<0.05
Adulteration level	1309.00	3	436.33	0.0000	Sig.
Adulterant type	0.88	1	0.88	0.5248	No sig.
Interaction	6.91	3	2.30	0.3764	No sig.
Within	33.37	16	2.09		
Total	1350.15	23			

Table S1-3. Effect of type of liquid adulteration and level of adulteration to protein content of coconut milk with ANOVA.

Source of variation	Sum of square	DoF	Mean square	P-value	P<0.05
Adulteration level	14.85	3	4.95	0.0000	Sig.
Adulterant type	0.04	1	0.04	0.0266	Sig.
Interaction	0.13	3	0.04	0.0054	Sig.
Within	0.11	16	0.01		
Total	15.13	23			

Table S1-4. Effect of type of liquid adulteration and level of adulteration to ash content of coconut milk with ANOVA.

Source of variation	Sum of square	DoF	Mean square	P-value	P<0.05
Adulteration level	0.64	3	0.21	0.0000	Sig.
Adulterant type	0.04	1	0.04	0.0000	Sig.
Interaction	0.07	3	0.02	0.0000	Sig.
Within	0.00	16	0.00		
Total	0.74	23			

Table S1-5. Effect of type of liquid adulteration and level of adulteration to carbohydrate content of coconut milk with ANOVA.

Source of variation	Sum of square	DoF	Mean square	P-value	P<0.05
Adulteration level	37.04	3	12.35	0.0000	Sig.
Adulterant type	0.88	1	0.88	0.2682	No sig.
Interaction	3.16	3	1.05	0.2350	No sig.
Within	10.72	16	0.67		
Total	51.81	23			

Table S1-6. Effect of type of solid adulteration and level of adulteration to moisture content of coconut milk with ANOVA.

Source of variation	Sum of square	DoF	Mean square	P-value	P<0.05
Adulteration level	401.92	3	133.97	0.0000	Sig.
Adulterant type	44.99	1	44.99	0.0004	Sig.
Interaction	25.64	3	8.55	0.0326	Sig.
Within	36.49	16	2.28		
Total	509.04	23			

Table S1-7. Effect of type of solid adulteration and level of adulteration to fat content of coconut milk with ANOVA.

Source of variation	Sum of square	DoF	Mean square	P-value	P<0.05
Adulteration level	1839.30	3	613.10	0.0000	Sig.
Adulterant type	3.64	1	3.64	0.0966	No sig.
Interaction	1.50	3	0.50	0.7362	No sig.
Within	18.71	16	1.17		
Total	1863.15	23			

Table S1-8. Effect of type of solid adulteration and level of adulteration to protein content of coconut milk with ANOVA.

Source of variation	Sum of square	DoF	Mean square	P-value	P<0.05
Adulteration level	17.16	3	5.72	0.0000	Sig.
Adulterant type	0.01	1	0.01	0.0543	No sig.
Interaction	0.01	3	0.00	0.1119	No sig.
Within	0.03	16	0.00		
Total	17.22	23			

Table S1-9. Effect of type of solid adulteration and level of adulteration to ash content of coconut milk with ANOVA.

Source of variation	Sum of square	DoF	Mean square	P-value	P<0.05
Adulteration level	1.23	3	0.41	0.0000	Sig.
Adulterant type	0.00	1	0.00	0.0723	No sig.
Interaction	0.00	3	0.00	0.0433	Sig.
Within	0.00	16	0.00		
Total	1.23	23			

Table S1-10. Effect of type of solid adulteration and level of adulteration to carbohydrate content of coconut milk with ANOVA.

Source of variation	Sum of square	DoF	Mean square	P-value	P<0.05
Adulteration level	4456.87	3	1485.62	0.0000	Sig.
Adulterant type	21.93	1	21.93	0.0001	Sig.
Interaction	21.40	3	7.13	0.0015	Sig.
Within	13.84	16	0.87		
Total	4514.04	23			

Appendix 2. Supplementary Material Chapter 3 – Case study 1

Table S2-1. Full preprocessing and hyperparameters using LDA classifier.

Link: https://ars.els-cdn.com/content/image/1-s2.0-S0026265X23010809-mmc1.xlsx
Sheet : Table S1

Table S2-2. Full preprocessing and hyperparameters using SVM classifier.

Link: https://ars.els-cdn.com/content/image/1-s2.0-S0026265X23010809-mmc1.xlsx
Sheet : Table S2

Table S2-3. Full preprocessing and hyperparameters using MLP classifier.

Link: https://ars.els-cdn.com/content/image/1-s2.0-S0026265X23010809-mmc1.xlsx
Sheet : Table S3

Table S2-4. Full preprocessing and hyperparameters using PLS regressor.

Link: https://ars.els-cdn.com/content/image/1-s2.0-S0026265X23010809-mmc1.xlsx
Sheet : Table S4

Table S2-5. Full preprocessing and hyperparameters using SVM regressor.

Link: https://ars.els-cdn.com/content/image/1-s2.0-S0026265X23010809-mmc1.xlsx
Sheet : Table S5

Table S2-6. Full preprocessing and hyperparameters using MLP regressor.

Link: https://ars.els-cdn.com/content/image/1-s2.0-S0026265X23010809-mmc1.xlsx
Sheet : Table S6

Appendix 3. Supplementary Material Chapter 4 – Case study 2

Table S3-1. Hyperparameters range of tuning from machine learning algorithms.

Case	Algorithm	Hyperparameters	Tuning range
Classifi-cation	LDA	Linear discriminant (LD)	1, 3, 5, 7, 9, 11
	KNN	n-neighbors (n)	1, 3, 5, 7, 9, 11
	MLP	Hidden layer sizes (HLS)	(3, 5, 7), (3, 5), (5, 7), (5)
		Activation function (AF)	Identity, Logistic, Tanh, ReLU
		Learning rate initial (LRI)	0.01, 0.1
Regression	PLS	Latent variable (LV)	1, 3, 5, 7, 9, 11
	KNN	n-neighbors (n)	1, 3, 5, 7, 9, 11
	MLP	Hidden layer sizes (HLZ)	(3, 5), (5, 7), (100, 100), (1000)
		Activation function (AF)	Identity, Logistic, ReLU
		Learning rate initial (LRI)	0.01, 0.1

Table S3-2. Statistics of the training and testing dataset.

Case	Subset	Training			Testing		
Classification using FT-NIR and Micro-NIR	**Samples**	**FCM**	**ADW**	**ACW**	**FCM**	**ADW**	**ACW**
	Samples per class	8	111	113	2	39	37
	Total samples	232			78		
Regression of ADW and ACW using FT-NIR and Micro-NIR	**Item**	**Range**	**Mean**	**SD**	**Range**	**Mean**	**SD**
	Range of samples	0 - 50	13.30	14.36	0 - 50	12.60	14.32
	Total samples	120			40		

Table S3-3. The vibration bands with a high X-loading from the PCA.

NIR spectra in cm^{-1} (nm)	References of NIR spectra in cm^{-1} (nm)	Functional group and band vibration	References
10,582 (945)	12,260 – 10,150 (816 – 985)	H_2O (2^{nd} overtone)	Conzen (2006)
	10,661 (938)	CH_2 (3^{rd} overtone of CH-str.)	Osborne *et al.* (1993)
8299 (1205)	8230 (1215)	CH_2 (2^{nd} overtone of CH-str.)	Osborne *et al.* (1993)
8230 (1215)			
7246 (1380)	7270 – 7220 (1376 – 1385)	H_2O (combination of OH-anti-sym.str. and OH-sym.str.)	Conzen (2006)
6900 – 6897 (1449 – 1450)	6940 – 6900 (1441 – 1449)	CH (combination of 2×CH-str. and CH-bending)	Conzen (2006); Osborne *et al.* (1993)
	6900 – 6850 (1449 – 1460)	H_2O (1st overtone of OH-str.)	
	6900 – 6780 (1449 – 1475)	NH_2 (1st overtone of $ArNH_2$)	
	6940 – 6850 (1441 – 1460)	$CONH_2$ (2^{nd} overtone), NH_2 (1^{st} overtone of NH_2-str.)	
5797 (1725)	5850 – 5780 (1709 – 1730)	CH3 (1^{st} overtone of CH-str.)	Conzen (2006)
	5797 (1725)	CH2 (1^{st} overtone of CH-str.)	Osborne *et al.* (1993)
5315 (1881)	5549 – 4550 (1802 – 2198)	Combination of OH-str. and OH-bending	Workman Jr and Weyer (2007)
5175 (1932)	5180 – 5150 (1931 – 1942)	H_2O (combination of OH-str. and OH-bending)	Conzen (2006)
	5180 – 5130 (1931 – 1949)	COOH, COOR (2^{nd} overtone of 2×C=O-str.)	
4825 (2073)	4850 – 4780 (2062 – 2092)	OH (combination of OH-str. and OH-bending)	Conzen (2006)
4500 (2222)	4460 (2242)	Amino acid (Combination NH-str. and NH_3-bending)	Osborne *et al.* (1993)
4340 (2304)	4340 – 4320 (2304 – 2315)	CH_2 (combination of CH-str. and CH-bending)	Conzen (2006)
	4329 (2310)		Osborne *et al.* (1993)
4260 (2347)	4270 – 4260 (2342 – 2347)	$HC{=}{=}CHCH_2$ (combination of CH_2-str. and CH_2-bending)	Conzen (2006); Osborne *et al.* (1993)

Table S3-4. Comparison of performance models to predict adulteration level of ADW in FCM using FT-NIR.

Regressor	The best preprocessing	Hyperparameter	Training		Testing			
			R^2	RMSE	R^2	RMSE	Bias	RPD
PLS	MS+SNV+MS+BSO3	LV=11	0.995	1.042	0.982	1.988	0.268	7.409
KNN	MSC+MS+BSO3	n=7	0.954	3.120	0.969	2.616	-0.143	5.689
MLP	SNV+SS	HLS=(3, 5, 7), AF=identity, LRI=0.01	0.943	6.590	0.950	6.358	0.640	4.487

Table S3-5. Comparison of performance models to predict adulteration level of ACW in FCM using FT-NIR.

Regressor	The best preprocessing	Hyperparameter	Training		Testing			
			R^2	RMSE	R^2	RMSE	Bias	RPD
PLS	SGF+SGF+SGF+SGF	LV=9	0.988	1.534	0.952	3.221	0.013	4.576
KNN	BSO3+MS+MSC+SS	n=3	0.968	2.549	0.938	3.687	0.550	4.017
MLP	MS+SGF+BSO3	HLS =(5,7), AF=identity, LRI=0.1	0.982	1.940	0.943	3.648	-0.135	4.182

Table S3-6. Comparison of performance models to predict adulteration level of ADW in FCM using Micro-NIR.

Regressor	The best preprocessing	Hyperparameter	Training		Testing			
			R^2	RMSE	R^2	RMSE	Bias	RPD
PLS	BSO3+BSO3+MS+BSO3	LV=11	0.992	1.311	0.988	1.585	0.070	8.990
KNN	FD+MS+SD+SNV	n=1	1.00	0.00	0.995	0.962	-0.025	14.769
MLP	BSO3+SD+ SNV+SGF	HLS =(1000), AF=logistic, LRI=0.01	0.998	0.587	0.988	1.571	-0.051	9.195

Table S3-7. Comparison of performance models to predict adulteration level of ACW in FCM using Micro-NIR.

Regressor	The best preprocessing	Hyperparameter	Training		Testing			
			R^2	RMSE	R^2	RMSE	Bias	RPD
PLS	FD+BSO3+SS+FD	LV=11	0.974	2.300	0.896	4.762	0.549	3.103
KNN	BSO3+MS+SD+SS	n=1	1.00	0.00	0.996	0.908	0.225	16.108
MLP	MS+SD+SS+SGF	HLS =(100, 100), AF=Logistic, LRI=0.01	0.998	0.741	0.981	2.024	0.593	7.317

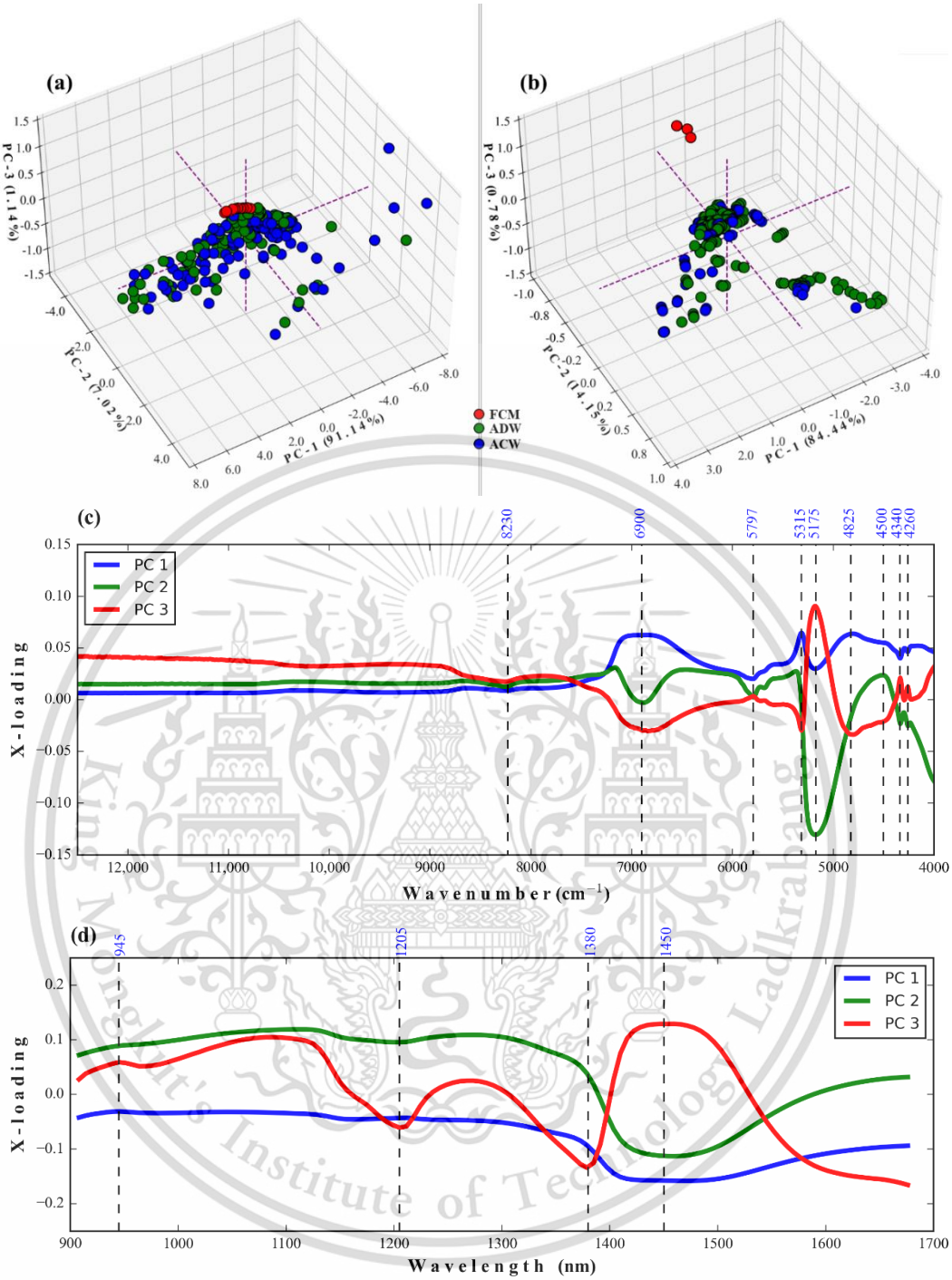


Figure S3-1. Projection of NIR spectra in 3D scores scatter plots of PCs from (a) FT-NIR and (b) Micro-NIR and loading of PC values from (c) FT-NIR (d) Micro-NIR.

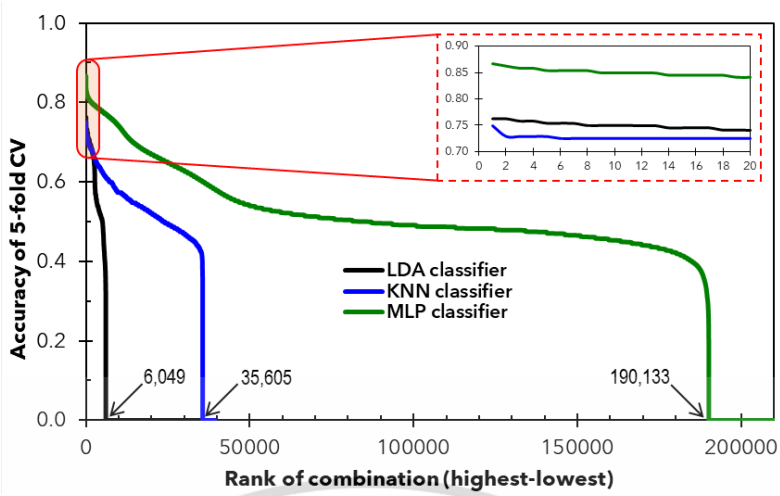


Figure S3-2. Descending list of preprocessing and hyperparameters using FT-NIR for classification problems.

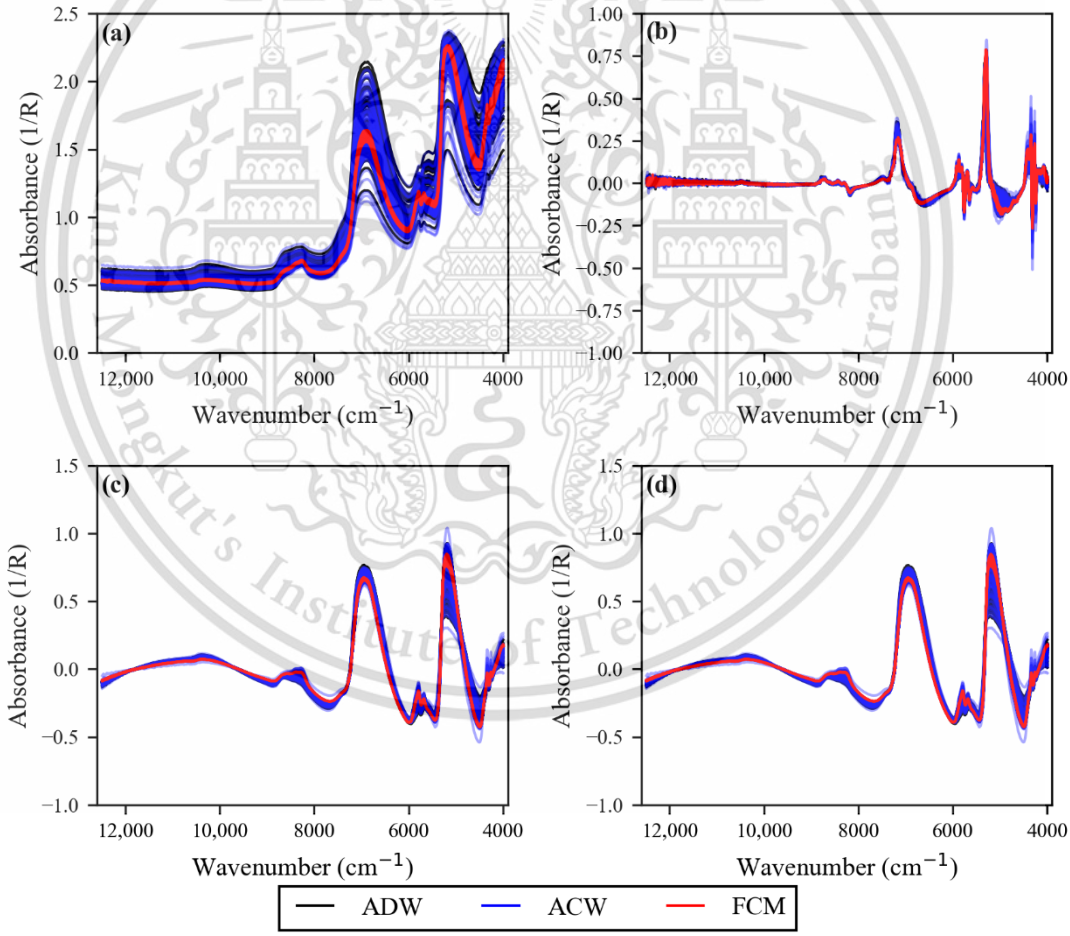


Figure S3-3. Preprocessing of FT-NIR spectra for classification case, (a) Raw (b) BSO3+SNV+BSO3+FD, (c) MSC+MSC+MS+BSO3, (d) MS+MSC+BSO3.

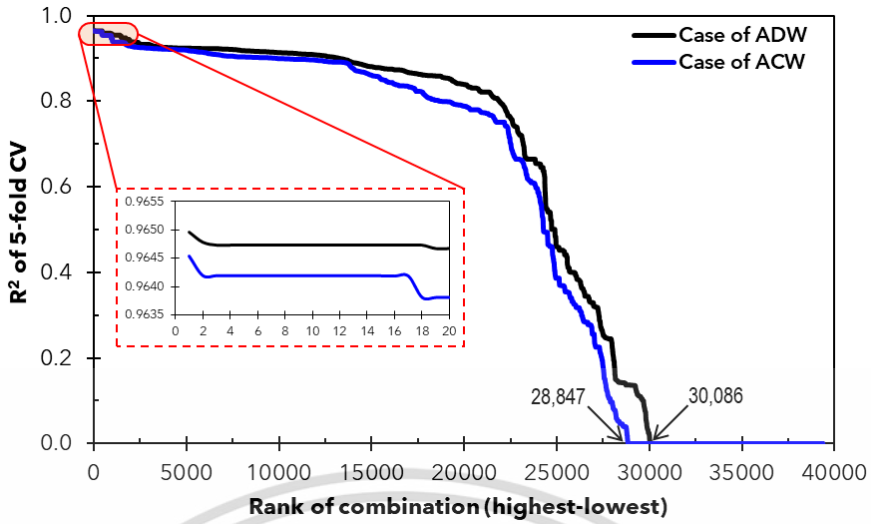


Figure S3-4. Descending list of preprocessing and hyperparameters using FT-NIR for PLS regressor.

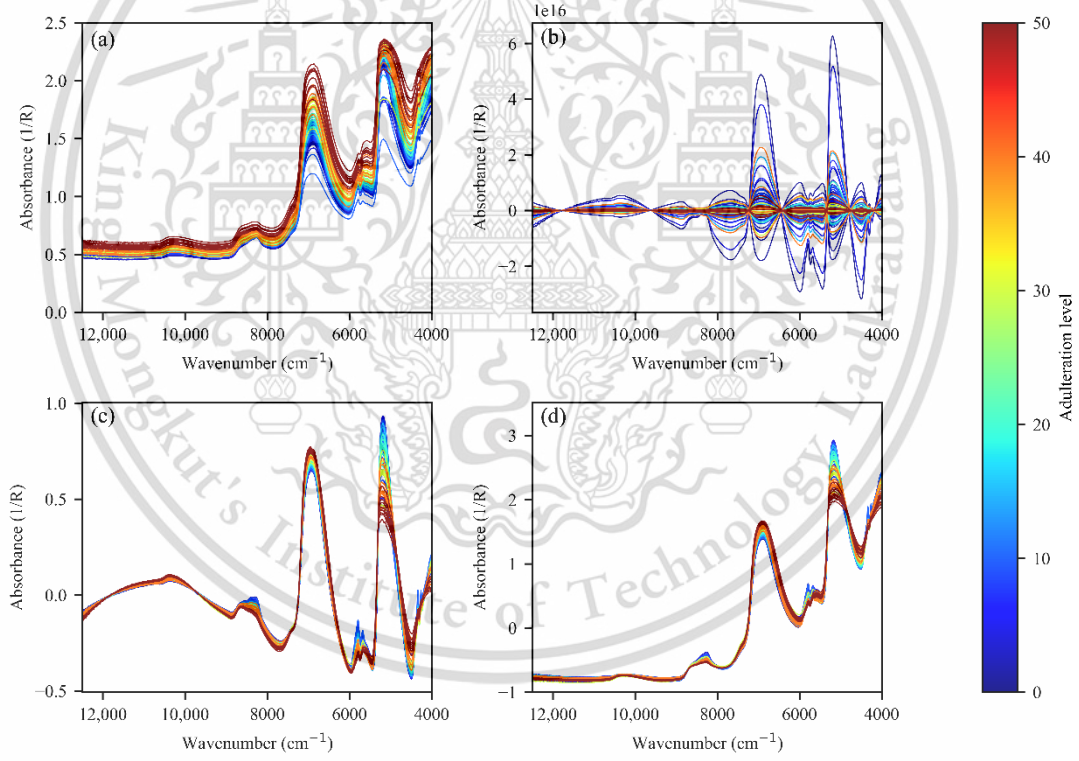


Figure S3-5. Preprocessing of FT-NIR spectra for prediction level of ADW, (a) raw (b) MS+SNV+MS+BSO3, (c) MSC+MS+BSO3, (d) SNV+SS.

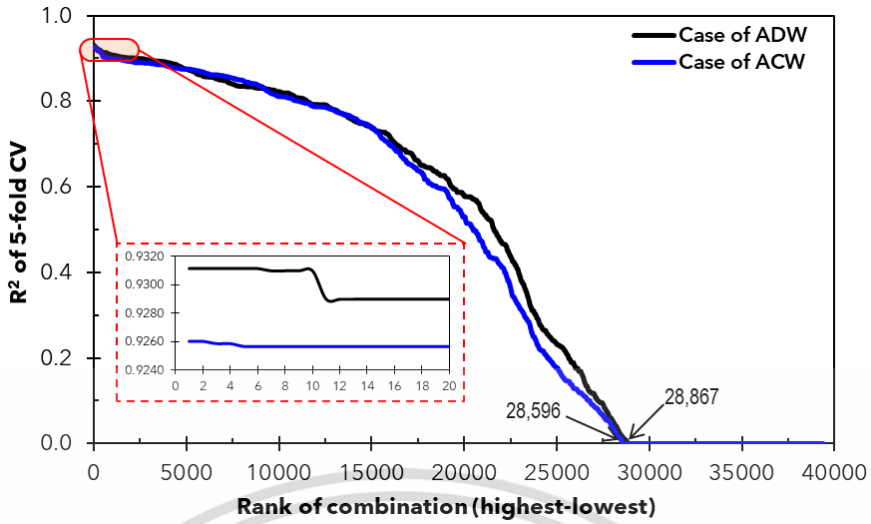


Figure S3-6. Descending list of preprocessing and hyperparameters using FT-NIR for KNN regressor.

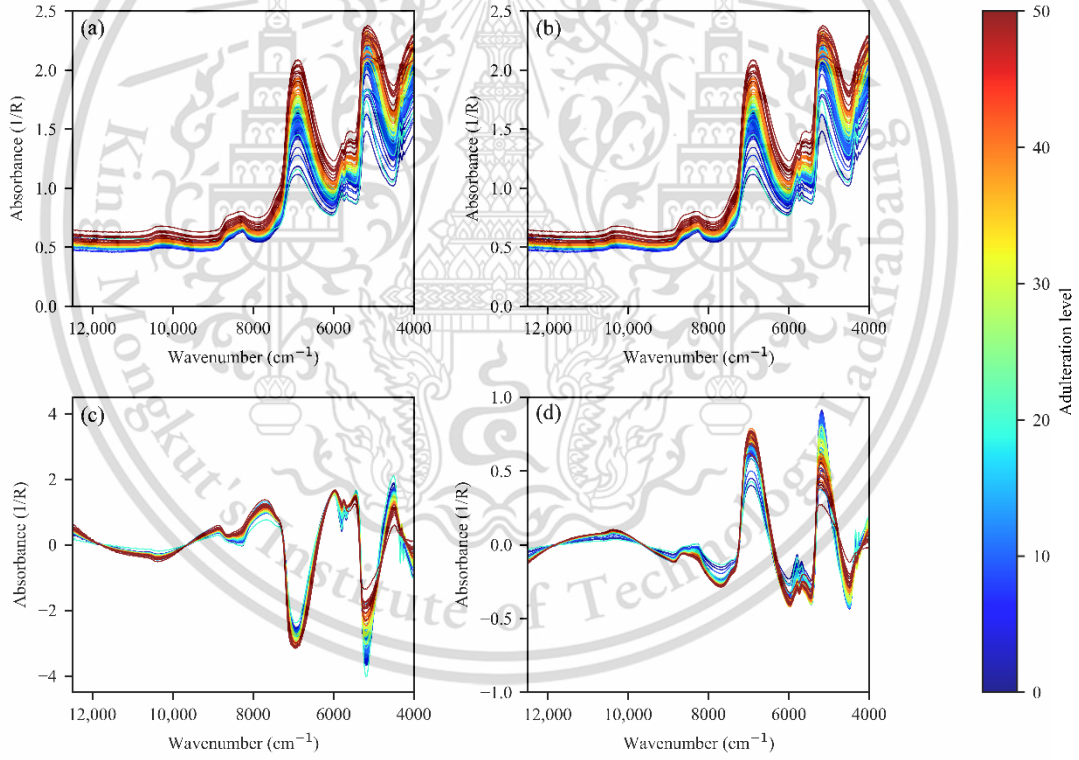


Figure S3-7. Preprocessing of FT-NIR spectra for prediction level of ACW, (a) raw, (b) SGF+SGF+SGF+SGF, (c) BSO3+MS+MSC+SS, (d) MS+SGF+BSO3.

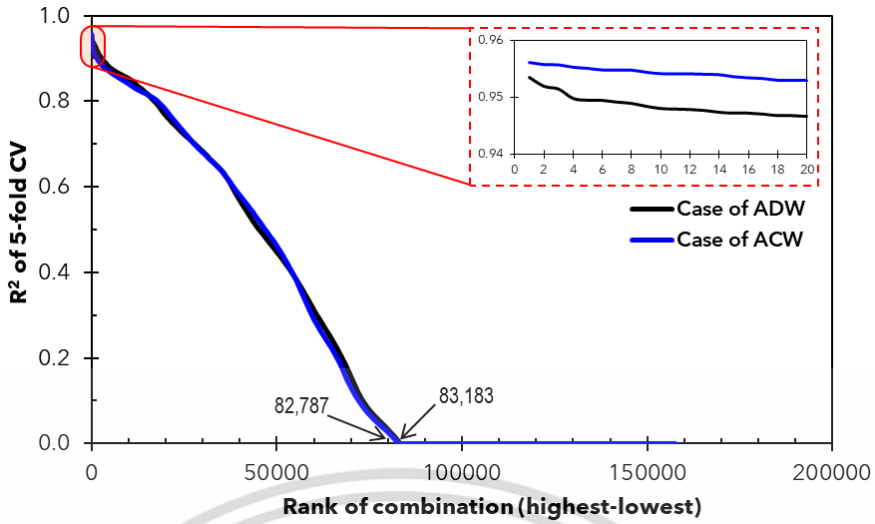


Figure S3-8. Descending list of preprocessing and hyperparameters using FT-NIR for MLP regressor.

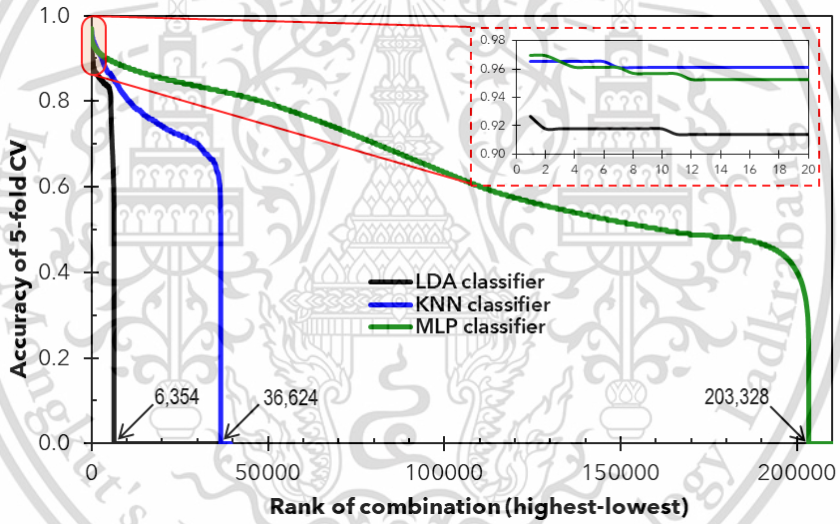


Figure S3-9. Descending list of preprocessing and hyperparameters using Micro-NIR for classification problem.

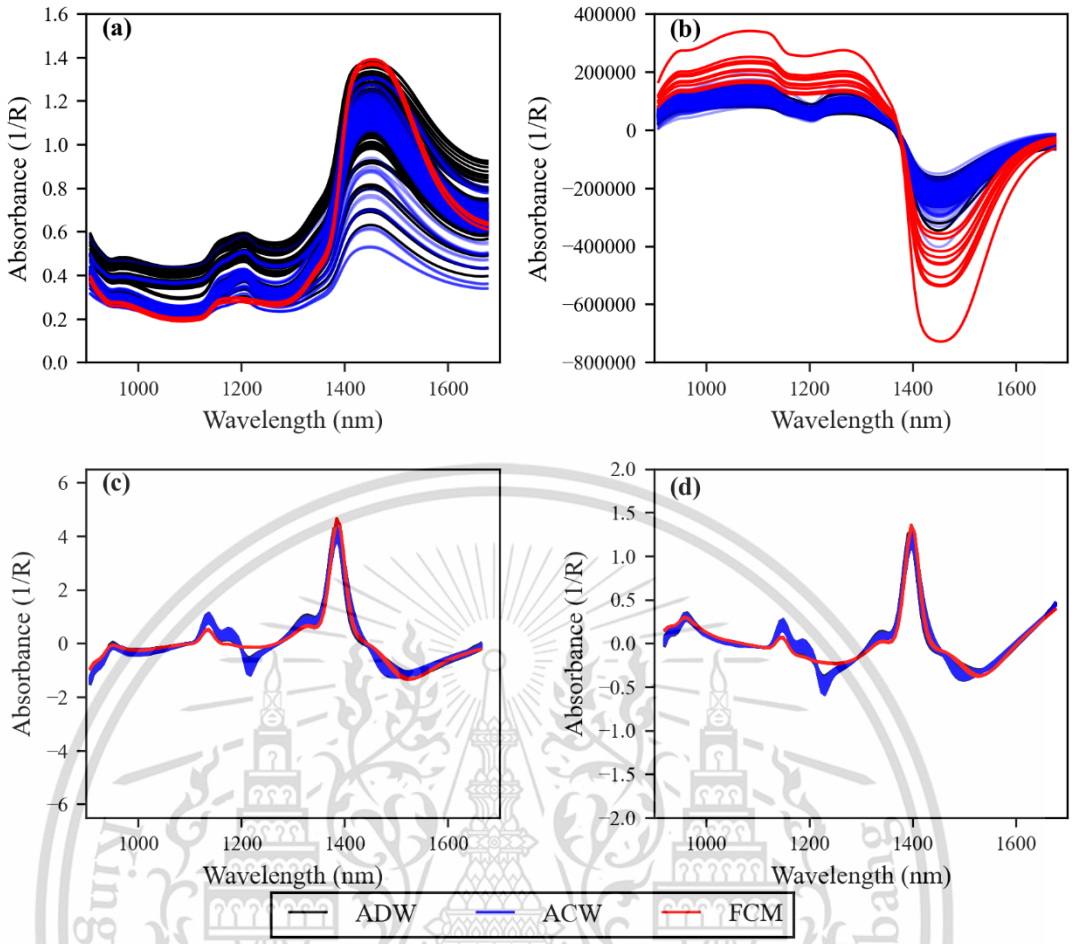


Figure S3-10. Preprocessing of spectra from Micro-NIR for classification case, (a) raw, (b) SNV+SGF+MS+SGF, (c) SS+FD+SS+SGF, (d) SGF+BSO3+SS+FD.

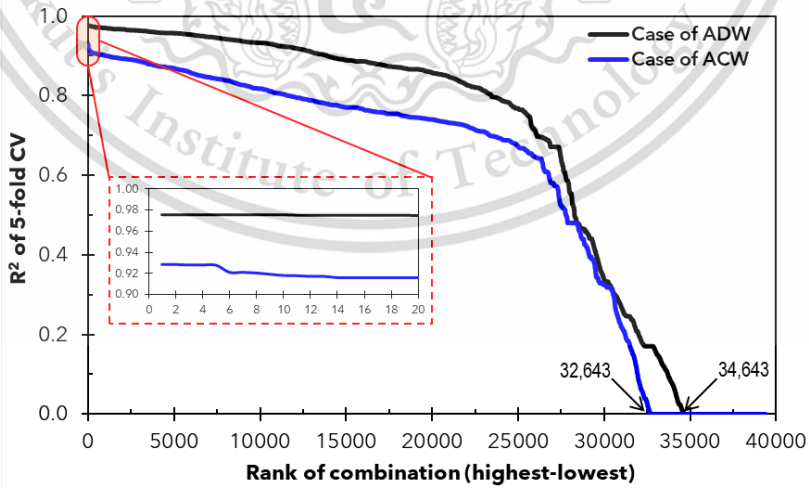


Figure S3-11. Descending list of preprocessing and hyperparameters using Micro-NIR for the PLS regression.

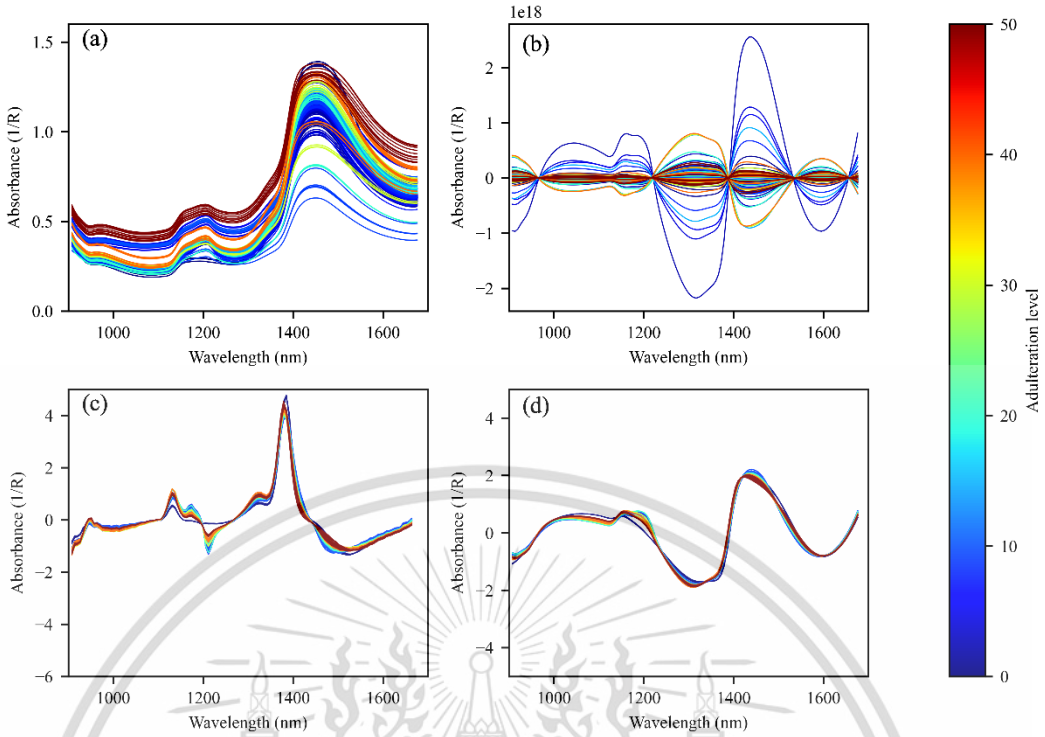


Figure S3-12. Preprocessing of spectra from Micro-NIR for prediction level of ADW, (a) raw, (b) BSO3+BSO3+MS+BSO3, (c) FD+MS+SD+SNV, (d) BSO3+SD+SNV+SGF.

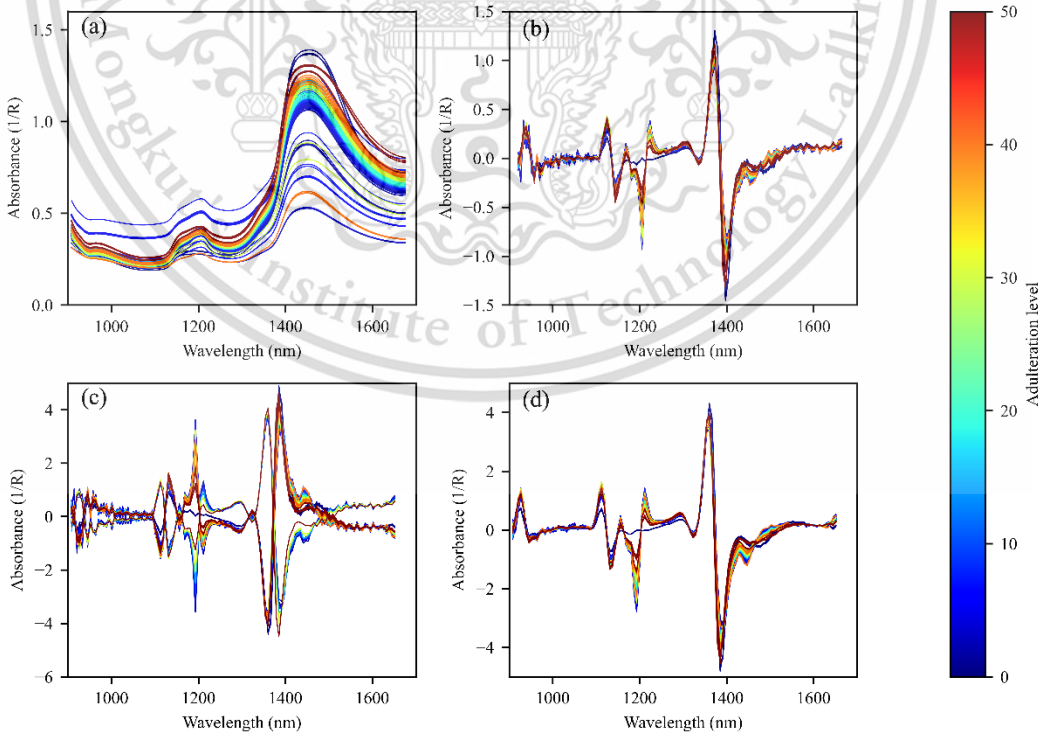


Figure S3-13. Preprocessing of spectra from Micro-NIR for prediction level of ACW, (a) raw, (b) FD+BSO3+SS+FD, (c) BSO3+MS+SD+SS, (d) MS+SD+SS+SGF.

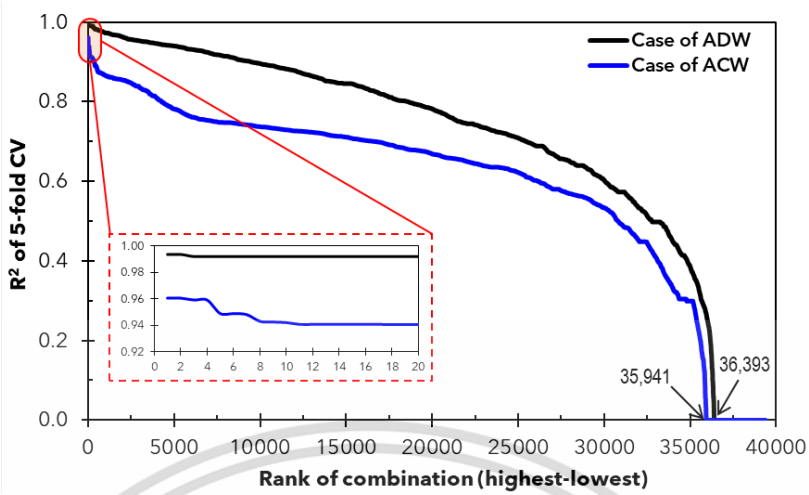


Figure S3-14. Descending list of preprocessing and hyperparameters using Micro-NIR for the KNN regressor.

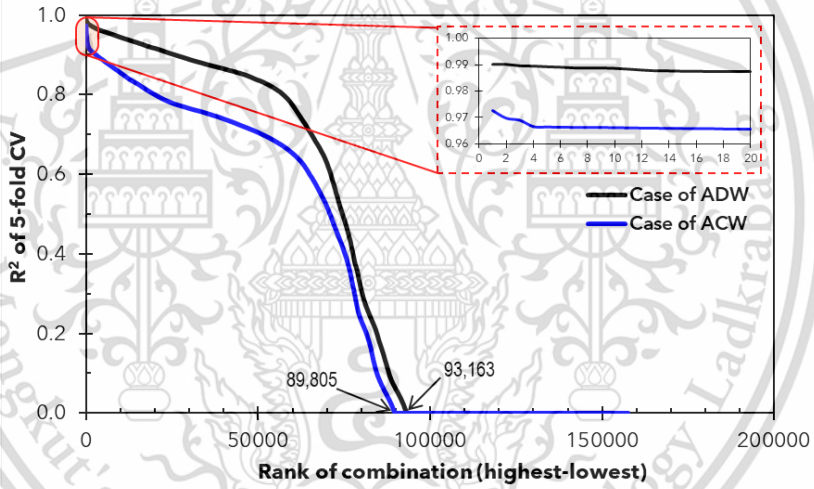


Figure S3-15. Descending list of preprocessing and hyperparameters using Micro-NIR for the MLP regressor.

References

Conzen, J. (2006). *Multivariate Calibration: A practical guide for developing methods in the quantitative analytical chemistry*.

Osborne, B. G., Fearn, T., & Hindle, P. H. (1993). *Practical NIR spectroscopy with applications in food and beverage analysis*: Longman scientific and technical.

Workman Jr, J., & Weyer, L. (2007). *Practical guide to interpretive near-infrared spectroscopy*: CRC press.

Appendix 4. Supplementary Material Chapter 6 – Case study 4

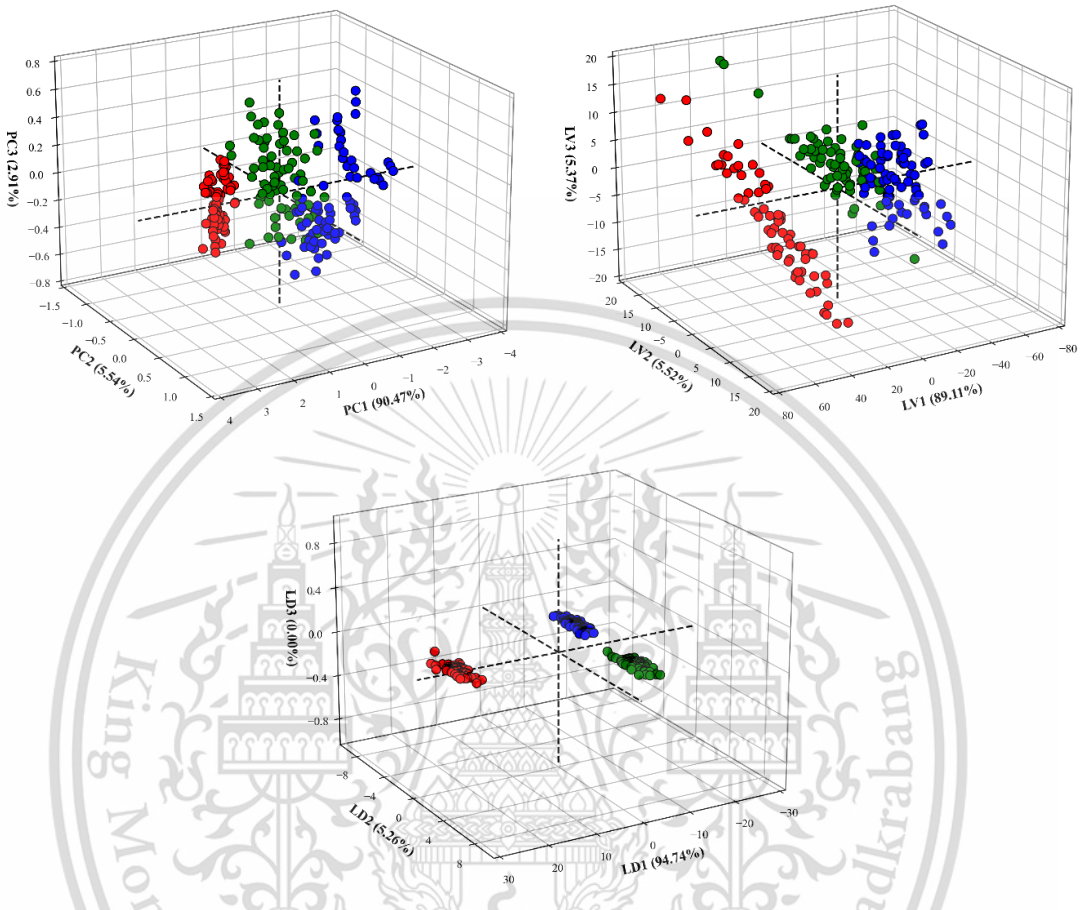


Figure S4-1. 3D scores scatter of classical chemometrics classifier using FT-NIR. (a) PCA, (b) PLS-DA and (c) LDA. ● SSK, ● CHP, ● CHB.

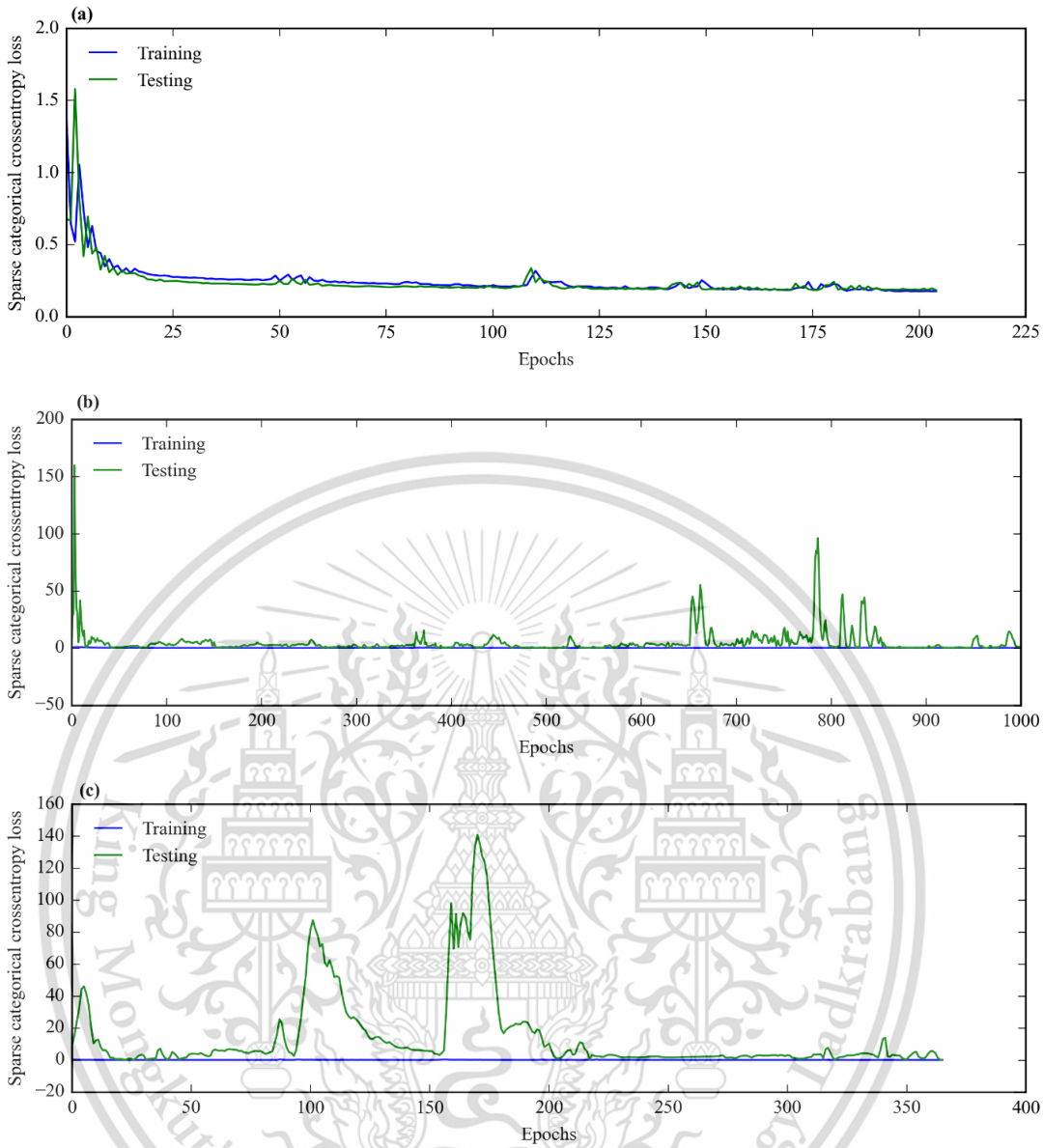


Figure S4-2. Loss curve of training and testing dataset from deep learning classifier using FT-NIR. (a) S-CNN, (b) S-AlexNET, and (c) ResNET.

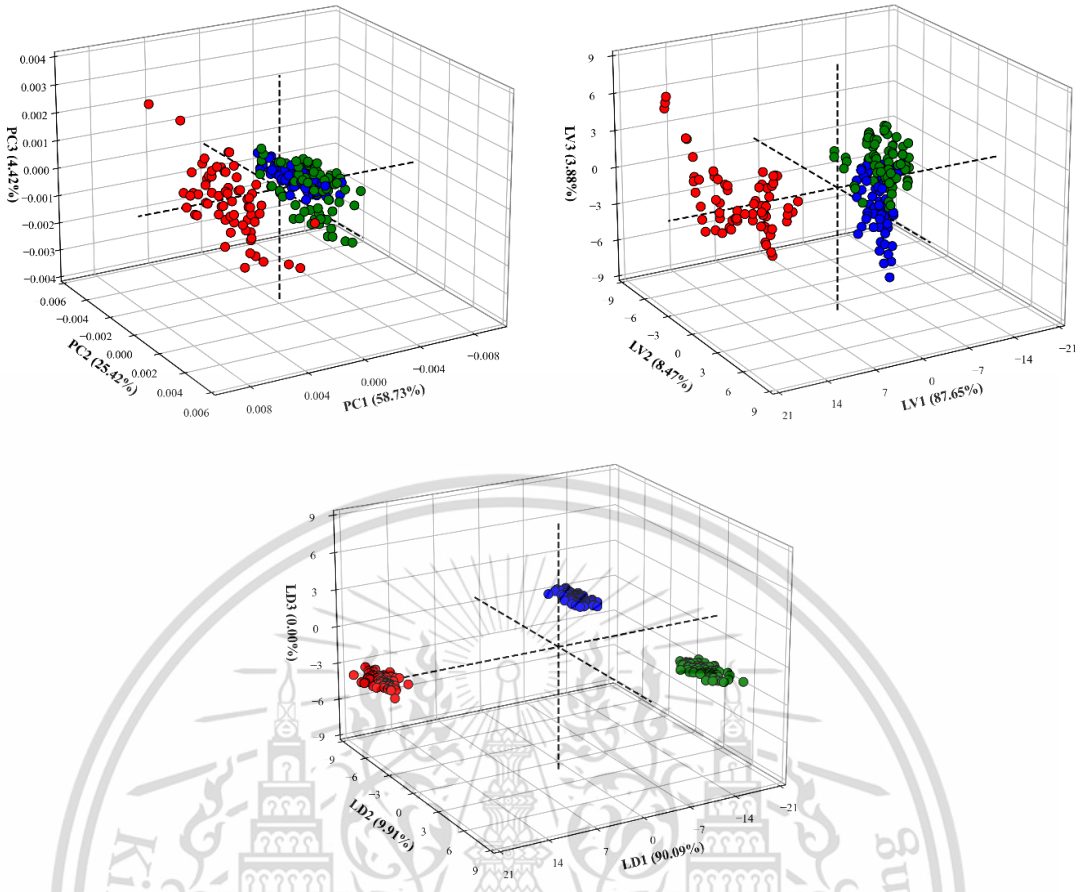


Figure S4-3. 3D scores scatter from of classical chemometrics classifier using Micro-NIR. (a) PCA, (b) PLS-DA, (c) LDA. ● SSK, ● CHP, ● CHB.

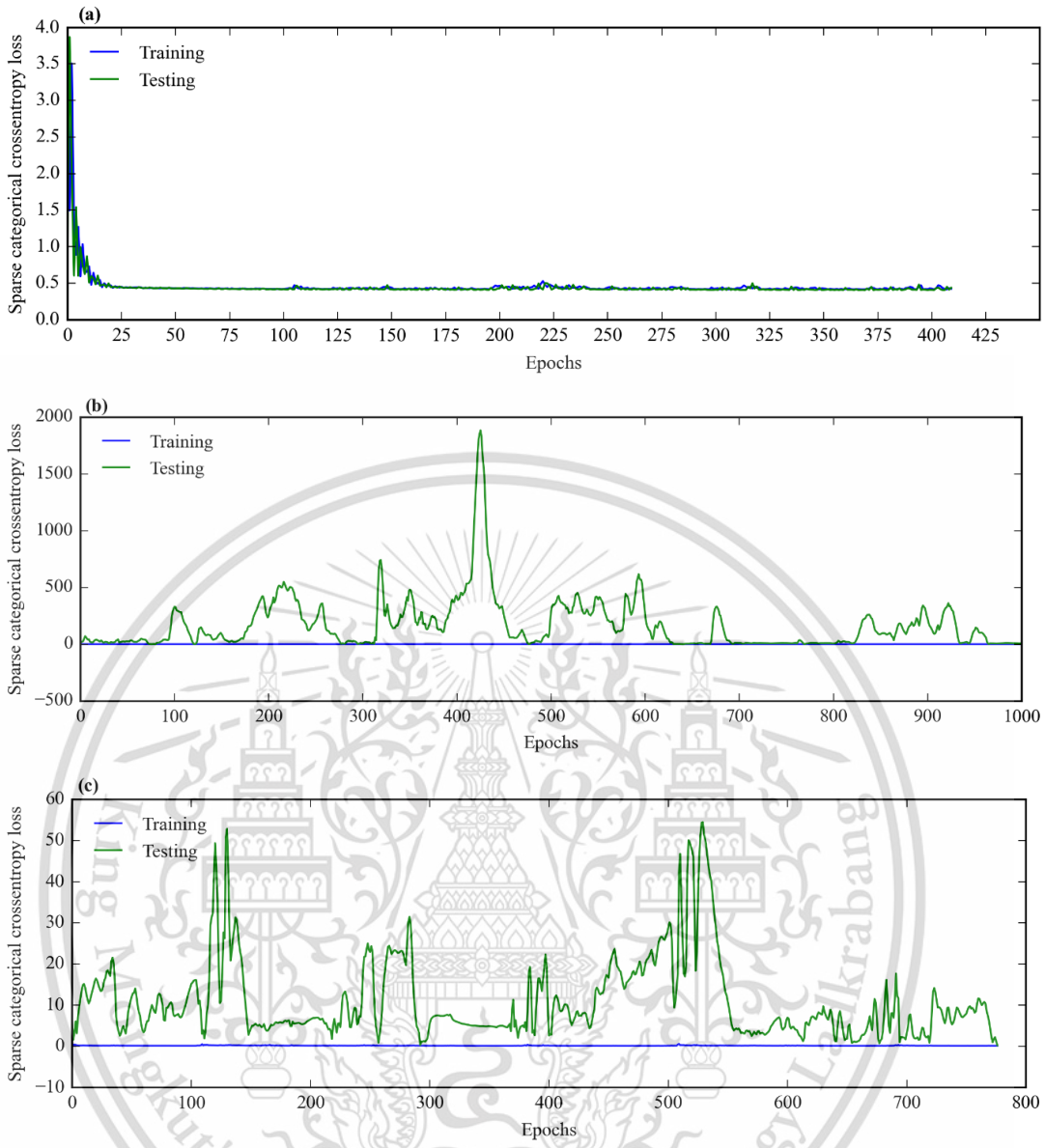


Figure S4-4. Loss curve of training and testing dataset from deep learning classifier using Micro-NIR. (a) S-CNN, (b) S-AlexNET, (c) ResNET.

Table S4-1. Spectra–structure correlation and absorption regions of major peaks found in coconut milk.

NIR spectra in cm^{-1} (nm)	References of NIR spectra in cm^{-1} (nm)	Functional group and band vibration	Ref.
803 (12,451)	803 (12,453)	C-H methyl C-H, associated with aromatic (ArCH3)	*
804 (12,435)	813 (12,300)	C-H methyl C-H, associated with branched aliphatic RC(CH3)3 or RCH(CH3)2	*
863 – 882 (11,587 – 11,340)	863 – 882 (11,587 – 11,340)	Aldehydes, Carbonyl C-H	**
892-912 (11,209-10,970)	884 – 914 (11,312 – 10,941)	Alkyne C-C.	**
908 (11,012)	908 (11,013)	C-H methyl (CH_3)	*
914 (10,937)	915 (10,929)	C-H methyl C-H (CH_3)	*
927 (10,791)	930 (10,753)	C-H methylene (CH_2)	*
964 (10,375)	962 (10,395)	O-H alkyl alcohols O-H with no hydrogen bonding (R-C-OH) in CCl_4	*
989 (10,115)	990 (10,101)	O-H phenols, dilute phenol in CCl_4	*
1018 (9820)	1019 (9818)	N-H primary aromatic amine (o-OCH3)	*
1143 (8745)	1143 (8749)	C-H (aromatic C-H)	*
1193 (8382)	1194 (8375)	C-H methyl C-H, (CH_3)	*
1212 (8253)	1211 (8258)	C-H methylene (CH_2)	*
1218 (8211)	1215 (8230)	C-H methylene (CH_2)	*
1224 (8710)	1225 (8163)	C-H secondary or tertiary carbon (CH)	*
1360 (7352)	1360 (7353)	C-H methyl (CH_3)	*
1376 – 1386 (7267 – 7213)	1370 – 1390 (7299 – 7194)	Beta, gamma-unsaturated gamma-lactone	**
1391 (7188)	1390 (7194)	SiOH	*
1397 (7156)	1395 (7168)	C-H methylene (CH_2)	*
1404 (7124)	1404 – 1422 (7123 – 7032)	O-H from alcohols in CCl_4 (first overtone (2**V**S) of the O-H alcoholic stretching modes	*
1410 (7093)			*
1422 (7032)			*
1441 (6941)	1440 (7092)	#N-H primary aromatic amine (m-NO_2), #O-H from sugar as crystalline sucrose	*
1447 (6911)	1446 (6916)	#C-H aromatic (ArC-H) #N-H primary aromatic amine (p-Cl)	*
1459 (6852)	1459 (6854)	N-H primary aromatic amine (p-NH2)	*
1466 (6823)	1465 (6826)	N-H for secondary amine as (R-NH-R)	*
1472 (6794)	1472 (6793)	N-H primary aromatic amine (o-NO_2)	*
1509 (6627)	1510 (6623)	N-H amide NH or NH_2	*
1515 (6600)	1515 (6601)	N-H bonded from polyamide 11	*
1571 (6366)	1570 (6369)	# N-H amide NH # N-H bonded combination from polyamide 11	*
1620 (6171)	1620 (6173)	C-H alkene, =CH_2	*
1653 (6048)	1654 (6045)	C-H methyl C-H, nitro (CH_3NO_2)	*
1670 (5986)	1671 (5985)	C-H aromatic C-H (aryl)	*
1677 (5963)	1678 (5960)	C-H methyl C-H, carbonyl adjacent as (C=OCH_3)	*
1688 (5925)	1688 (5925)	C-H methyl C-H, OH associated as ($ROHCH_3$)	*
1689 (5920)	1689 (5920)	C-H aromatic C-H (aryl)	*
1726 (5794)	1725 (5797)	C-H methylene (.CH_2)	*
1728 (5786)	1728 (5787)	C-H methylene (CH_2) (asymmetric)	*
1731 (5778)	1732 (5773)	C-H methyl C-H, OH associated as ($ROHCH_3$)	*

1738 (5755)	1738 (5755)	$CONH_2$ specifically due to C=O hydrogen bonded to the N-H of the peptide link termed the α-helix structure	*
1740 (5747)	1740 (5747)	S-H thiol (S-H)	*
1742 (5739)	1744 (5735)	C-H methyl C-H, aromatic associated ($ArCH_3$)	*
1790 (5585)	1790 (5587)	O-H from water	*
1798 – 1906 (5562 – 5246)	1783 – 1903 (5609 – 5255)	C-O saturated, 5-member ring	**
1909 (5238)	1908 (5241)	P-OH phosphate (P-OH)	*
1920 (5207)	1920 (5208)	C=O amide (C=ONH)	*
1929 (5184)	1930 (5181)	O-H (O-H and HOH)	*
1943 (5145)	1940 (5155)	OH — classic filter instrument	*
1964 (5091)	1965 (5090)	O-H and CH combination from methanol	*
1976 (5061)	1976 (5062)	N-H combination, primary aromatic amine ($p\text{-}CH_3$)	*
1979 (5053)	1980 (5051)	N-H amide II ($CONH_2$)	*
2092 (4780)	2090 (4785)	N-H from gamma-valerolactam	*
2201 (4544)	2200 (4545)	CHO carbohydrate (CHO)	*
2231 (4482)	2230 (4484)	CHO — classic filter instrumen	*
2270 (4405)	2270 (4405)	O-H/C-H cellulose (OH and C-O)	*
2273 (4400)	2273 (4400)	O-H/C-O from glucose	*
2278 (4389)	2280 (4386)	C-H starch (C-H and CH_2)	*
2302 (4343)	2302 (4343)	C-H (C-H bending)	*
2307 (4335)	2307 (4333)	C-H methylene C-H, associated with linear aliphatic $R(CH_2)NR$	*
2311 (4328)	2310 (4329)	C-H (C-H bending)	*
2319 (4312)	2322 (4307)	C-H methylene C-H, associated with branched aliphatic $RC(CH_3)_3$ or $RCH(CH_3)_2$	*
2323 (4305)	2330 (4292)	C-H (C-H and CH_2)	*
2327 (4297)	2330 (4292)	C-H (C-H and CH_2)	*
2336 (4281)	2336 (4281)	CHO — classic filter instrument	*
2344 (4266)	2345 (4265)	C-H methylene C-H, associated with ovalbumin protein side chains seen at pH 5.0	*
2361 (4235)	2363 (4232)	C-H methylene C-H, associated with branched aliphatic $RC(CH_3)_3$ or $RCH(CH_3)_2$	*
2387 (4189)	2387 (4190)	C-H aromatic C-H (aryl)	*
2405 (4158)	2407 (4155)		*
2441 (4096)	2440 (4099)		*
2488 (4019)	2488 (4019)	C-H/C-C (C-H and C-C)	*

* Workman Jr, J., & Weyer, L. (2007). Practical guide to interpretive near-infrared spectroscopy: CRC press.

** Workman, J. (2016). The Concise Handbook of Analytical Spectroscopy: Theory, Applications, and Reference Materials: Volume 3: Near Infrared Spectroscopy.

International Published Papers

1. **Sitorus, A.**, and Lapcharoensuk, R. (2024). Exploring Deep Learning to Predict Coconut Milk Adulteration Using FT-NIR and Micro-NIR Spectroscopy. Sensors, 24(7), 2362. Doi: https://doi.org/10.3390/s24072362.
2. **Sitorus, A.**, and Lapcharoensuk, R. (2023). A rapid method to predict type and adulteration of coconut milk by near-infrared spectroscopy combined with machine learning and chemometric tools. Microchemical Journal, 195, 109461. Doi: https://doi.org/10.1016/j.microc.2023.109461.
3. **Sitorus, A.**, and Lapcharoensuk, R. (2022). A comprehensive overview of near infrared and infrared spectroscopy for detecting the adulteration on food and agroproducts—a critical assessment. INMATEH-Agricultural Engineering, 67(2). Doi: https://doi.org/10.35633/inmateh-67-47.

Author Biography

Name-Surname : Agustami Sitorus

Date of birth : 28 November 1989

Email address : (THA) 64601001@kmitl.ac.th; (IND) agustami.sitorus@brin.go.id

Contact number : (THA) +66 93 720 4957; (IND) +62 812 8343 9334

Contact address

1. THA The Kanyarat Building, 539/1, Soi Chalong Krung 1 Yak 3/Chalong Krung Road, Lad krabang, bangkok, Thailand.
2. IND. Buana Subang Kencana Residence, 048/016, Soklat, Subang, West Java, Indonesia.

Education background

1. D.Eng. (Food and Agricultural Intelligence Engineering) from King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand (2021-2024, GPA: 0.00).
2. M.Sc. (Agricultural and Food Machinery Engineering) from IPB University, Bogor, Indonesia (2013-2015, GPA: 3.84).
3. B.Eng. (Agricultural Engineering) from Universitas Sumatera Utara, Medan, Indonesia (2008-2012, GPA: 3.62).

Scholarships

KMITL Doctoral Scholarships from King Mongkut's Institute of Technology Ladkrabang, Thailand (Grant numbers KDS2020/049).

International selected published papers

1. **Sitorus, A.**, and Lapcharoensuk, R. (2024). Exploring Deep Learning to Predict Coconut Milk Adulteration Using FT-NIR and Micro-NIR Spectroscopy. Sensors, 24(7), 2362. Doi: https://doi.org/10.3390/s24072362.
2. **Sitorus, A.**, Pambudi, S., Boodnon, W., and Lapcharoensuk, R. (2024). Near-Infrared Spectroscopy with Machine Learning for Classifying and Quantifying Nutmeg Adulteration. Analytical Letters, 57(2), 285-306. Doi: https://doi.org/10.1080/00032719.2023.2206665.
3. Jongyingcharoen, J. S., Howimanporn, S., **Sitorus, A.**, Phanomsophon, T., Posom, J., Salubsi, T., A. Kongwaree, C. H. Lim, K. Phetpan , P. Sirisomboon and Tsuchikawa, S. (2024). Classification of the Crosslink Density Level of Para

Rubber Thick Film of Medical Glove by Using Near-Infrared Spectral Data. Polymers, 16(2), 184. Doi: https://doi.org/10.3390/polym16020184.

4. **Sitorus, A.**, and Lapcharoensuk, R. (2023). A rapid method to predict type and adulteration of coconut milk by near-infrared spectroscopy combined with machine learning and chemometric tools. Microchemical Journal, 195, 109461. Doi: https://doi.org/10.1016/j.microc.2023.109461.
5. Lapcharoensuk, R., Fhaykamta, C., Anurak, W., Chadwut, W., and **Sitorus, A**. (2023). Nondestructive detection of pesticide residue (Chlorpyrifos) on bok choi (Brassica rapa subsp. Chinensis) using a portable NIR spectrometer coupled with a machine learning approach. Foods, 12(5), 955. Doi: https://doi.org/10.3390/foods12050955.
6. **Sitorus, A.**, and Lapcharoensuk, R. (2022). A comprehensive overview of near infrared and infrared spectroscopy for detecting the adulteration on food and agroproducts—a critical assessment. INMATEH-Agricultural Engineering, 67(2). Doi: https://doi.org/10.35633/inmateh-67-47.
7. Pramono, E. K., Karim, M. A., Fudholi, A., Bulan, R., Lapcharoensuk, R., and **Sitorus, A.** (2022). Low cost telemonitoring technology of semispherical solar dryer for drying arabica coffee beans. INMATEH-Agricultural Engineering, 66(1). Doi: https://doi.org/10.35633/inmateh-66-34.

Areas of expertise and interest

1. Data science and analytics;
2. Chemometrics using machine learning and deep learning;
3. Design machinery for food and Agricultural.

Reference persons

Asst. Prof. Dr. Ravipat Lapchareonsuk, Department of Agricultural Engineering, Faculty of Engineering, King Mongkut's Institute of Technology Ladkrabang, Thailand, Email: ravipat.la@kmitl.ac.th.

***