DISCRIMINATION OF WEEDY RICE BY USING NEAR-INFRARED SPECTROSCOPY COMBINED WITH CHEMOMETRICS

Miss Sureerat Makmuang

A Dissertation Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy in Chemistry
Department of Chemistry
FACULTY OF SCIENCE
Chulalongkorn University
Academic Year 2021

การคัดแยกข้าววัชพืชโดยใช้สเปกโทรสโกปีอินฟราเรดย่านใกล้ร่วมกับเคโมเมทริกซ์

น.ส.สุรีรัตน์ มากเมือง

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรดุษฎีบัณฑิต
สาขาวิชาเคมี ภาควิชาเคมี
คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2564

Thesis Title	DISCRIMINATION OF WEEDY RICE BY USING NEAR-INFRARED SPECTROSCOPY COMBINED WITH CHEMOMETRICS
By	Miss Sureerat Makmuang
Field of Study	Chemistry
Thesis Advisor	Associate Professor KANET WONGRAVEE, Ph.D.

Accepted by the FACULTY OF SCIENCE, Chulalongkorn University in Partial Fulfillment of the Requirement for the Doctor of Philosophy

..... Dean of the FACULTY OF SCIENCE
(Professor POLKIT SANGVANICH, Ph.D.)

DISSERTATION COMMITTEE

..... Chairman
(Sumaporn Kasemsumran, Ph.D.)

..... Thesis Advisor
(Associate Professor KANET WONGRAVEE, Ph.D.)

..... Examiner
(Professor VITHAYA RUANGPORNVISUTI, Dr.rer.nat.)

..... Examiner
(Associate Professor APICHAT IMYIM, Ph.D.)

..... Examiner
(Associate Professor VIWAT VCHIRAWONGKWIN, Dr.rer.nat.)

..... External Examiner
(Sumaporn Kasemsumran, Ph.D.)

สุรีรัตน์ มากเมือง : การคัดแยกข้าววัชพืชโดยใช้สเปกโทรสโกปีอินฟราเรดย่านใกล้ร่วมกับเคโมเมทริกซ์. (DISCRIMINATION OF WEEDY RICE BY USING NEAR-INFRARED SPECTROSCOPY COMBINED WITH CHEMOMETRICS) อ.ที่ปรึกษาหลัก : รศ. ดร.คเณศ วงษ์ระวี

ข้าววัชพืชเป็นหนึ่งในวัชพืชที่เกิดขึ้นมากในพื้นที่ปลูกข้าวโดยเฉพาะแถบเอเชียตะวันออกเฉียงใต้ ข้าววัชพืชมีลักษณะทางกายภาพภายนอกเหมือนกับข้าวที่ปลูกโดยเฉพาะตอนเป็นข้าวเปลือก ด้วยเหตุนี้จึงเป็นเรื่องยากที่จะสามารถจำแนกข้าววัชพืชออกจากข้าวปกติ งานวิจัยนี้ได้นำเสนอการเทคนิคปรับเปลี่ยนแผนที่โยงก่อร่างตัวเอง (Self-Organizing Maps, SOMs) มาใช้สำหรับการจำแนกข้าววัชพืชจากข้าวที่ปลูก ผ่านการวิเคราะห์ข้อมูลที่ได้จากเทคนิคสเปกโทรสโกปีอินฟราเรดย่านใกล้และเทคนิคการถ่ายภาพเชิงสเปกตรัม ก่อนการตรวจวัดตัวอย่างข้าวถูกปรับสภาพโดยเครื่องดูดฝุ่นแบบไซโคลนเพื่อขจัดอนุภาคที่ปนเปื้อนและสิ่งเจือปนอื่นๆ บนเปลือกข้าว ลักษณะทางกายภาพถูกตรวจสอบด้วยกล้องจุลทรรศน์อิเล็กตรอนแบบส่องกราด การสลายตัวขององค์ประกอบจากความร้อนวิเคราะห์ด้วยการเปลี่ยนแปลงน้ำหนักของสารโดยอาศัยคุณสมบัติทางความร้อน ระบุอัตลักษณ์ทางเคมีของสารระเหยง่ายจากตัวอย่างด้วยเทคนิคแมสสเปกโทรสโกปีความละเอียดสูง ในงานวิจัยนี้เทคนิคสเปกโทรสโกปีอินฟราเรดใกล้พร้อมอุปกรณ์เสริมเพื่อวัดการสะท้อนแสงถูกใช้สำหรับการวิเคราะห์ตัวอย่างโดยตรง สเปกตรัมที่ได้จะถูกปรับสัญญาณให้เรียบโดยใช้พหุนามสาวิทซกี-โกเลย์ หลังจากนั้นสเปกตรัมจะถูกปรับความแปรปรวนให้เป็นมาตรฐานและปรับค่าเฉลี่ยให้อยู่ตรงกลาง ในส่วนสุดท้ายสเปกตรัมจะถูกเพิ่มความละเอียดบริเวณที่มีนัยสำคัญโดยใช้ฟังก์ชันอนุพันธ์อันดับสอง จากนั้นข้อมูลทางสเปกตรัมดังกล่าวจะถูกนำมาสร้างแผนที่ SOMs โดยมีการปรับตัวแปรต่าง ๆ ให้เหมาะสมเพื่อนำไปใช้จำแนกข้าววัชพืชออกจากข้าวปลูกสี่ชนิด พบว่าจากเทคนิคที่พัฒนาขึ้นนั้นสามารถทำนายชนิดของข้าวได้ถูกต้องและแม่นยำในช่วง 88% ถึง 99% และ 91% ถึง 99% ตามลำดับ นอกจากนี้แผนที่ SOMs ที่ได้รับการพัฒนายังถูกนำไปประยุกต์ใช้กับข้อมูลจากภาพถ่ายไฮเปอร์สเปกตรัมเพื่อสร้างแผนที่มาตรฐาน SOMs ซึ่งกลุ่มตัวอย่างจะถูกแทนที่ด้วยสีต่างกันบนแผนที่ จากนั้นแต่ละพิกเซลของรูปถ่ายไฮเปอร์สเปกตรัมจะถูกวิเคราะห์ด้วยแผนที่มาตรฐาน SOMs จากนั้นสีของหน่วยแผนที่ที่ดีที่สุด (BMU) บนแผนที่จะถูกฉายซ้ำลงบนพิกเซลของภาพนั้น ๆ กระบวนการนี้ดำเนินไปจนกระทั่งพิกเซลของภาพทั้งหมดถูกแทนที่ด้วยสีของ BMU จากนั้นอัตราส่วนของสีที่อยู่บนภาพตัวอย่างจะทำให้สามารถทำนายกลุ่มของตัวอย่างได้ จากผลการศึกษาพบว่าเทคนิคดังกล่าวสามารถจำแนกเมล็ดพันธุ์วัชพืชออกจากข้าวปลูกได้โดยมีความแม่นยำถึง 90% ซึ่งแสดงให้เห็นถึงศักยภาพของแบบจำลองแผ่นที่มาตรฐาน SOMs ในการประยุกต์กับข้อมูลทางสเปกโทรสโกปีในการประเมินคุณภาพเมล็ดพันธุ์

สาขาวิชา เคมี ลายมือชื่อนิสิต .....

ปีการศึกษา 2564 ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

6172910723 : MAJOR CHEMISTRY
KEYWORD: weedy rice cultivated rice near-infrared (NIR) hyperspectral NIR camera spectroscopy self-organizing map (SOMs)

Sureerat Makmuang : DISCRIMINATION OF WEEDY RICE BY USING NEAR-INFRARED SPECTROSCOPY COMBINED WITH CHEMOMETRICS. Advisor: Assoc. Prof. KANET WONGRAVEE, Ph.D.

Weedy rice is one of the most notorious weeds occurring in rice-growing areas, especially in South-East Asia. Weedy rice especially in form of paddy seed is difficult to manage and separate as they provide common features (morphological resemblance) to cultivated rice. This work presents a modification of self-organizing map (SOMs) for the classification of weedy rice from cultivated rice via in situ direct sample analysis from paddy seed using near-infrared (NIR) spectroscopy and hyperspectral NIR camera. The sample pretreatment was carried out by a cyclone vacuum machine to remove the contaminated particles and other impurities. The physical characteristics and the thermal behavior of rice samples were investigated by optical microscope and thermogravimetric analysis (TGA), respectively, and the volatile chemical profiles were monitored by using DART-MS. They provide the distinctive patterns between cultivated rice and weed rice. A near-infrared with reflectance accessory was used for direct sample analysis. The acquired NIR spectra were smoothed using Savitzky-Golay polynomial, baseline-aligned using standard normal variate (SNV), mean-centered and the second derivative was calculated to reveal the significant NIR regions. Self-organizing maps was well-optimized and was applied for the classification of weedy samples from four cultivated rice. The results were validated and were achieved very high predictive value in the range of 91% to 99% and 88% to 99% for precision and accuracy, respectively. Furthermore, the developed supervised SOMs was applied on the pair-wise hyperspectral image to generate the supervised global SOM map with different color scales as the representative of each sample class. Each hyperspectral pixel from the sample image was validated with the global map, then, the color of best map unit (BMU) was re-projected on the image pixel. The process was undergone until all image pixels was projected with the color of BMU. The classification was achieved by the ratio of the projected color on the sample image. The classification accuracy for weedy seeds was 90%, demonstrating the potential of a global model for seed quality assessment.

Field of Study:	Chemistry	Student's Signature .....
Academic Year:	2021	Advisor's Signature .....

ACKNOWLEDGEMENTS

On the day you don't believe in yourself, how many people will believe in you? On the day you have no choice, how many people will offer you a choice? I felt extremely lucky in this life once I met that person. Six years had turned one stranger into the person who now plays an important role in my life. My supervisor, Associate Professor Dr. Kanet Wongravee, patiently gave advice, inspiration, encouragement, and endless support to me. Without his support, this dissertation would not have been possible. For me, he is one of the best advisors ever. I would like to express my deepest gratitude to him.

I am particularly grateful for the valuable assistance given by SRU members who always shared the moment and cheerfulness with me and assisted me in completing this work through continuously good recommendations. Their support has meant more to me than they could possibly realize.

I would also like to extend my thanks to the Science Achieve Scholarship of Thailand (SAST) for financial support.

A special thank goes to my boyfriend, Mr.Sornsiri Malangpoo, who is my spiritual anchor. I highly appreciate his emotional support for being so understanding and for being such a great listener.

Finally, I would like to express my deep gratitude to my family, especially my mom and dad, for their constant love and endless support. They are always beside me no matter where I am in any situation. I have been exceedingly fortunate to have them as my family.

Sureerat Makmuang

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATION

EDC	: Euclidean distance to centroid
LDA	: Linear discriminant analysis
QDA	: Quadratic discriminant analysis
PLSDA	: Partial least square discriminant analysis
SOMs	: Self-organizing maps
SNV	: Standard normal variate
BMU	: Best Map Unit
%CC	: Percent Correctly Classified
NIPALS	: Nonlinear iterative partial least squares
KHML105	: Khao Hom Mali 105
RD49	: Kor Khor 49
PTT1	: Pratumtani1
PL2	: Phitsanulok2
TGA	: Thermogravimetric analysis
DART-MS	: Direct analysis in real time mass spectrometry
LVs	: Latent variables
PCs	: Principal components
IQR	: Interquartile range (IQR = Q2)
Q1	: Quartile 1
Q3	: Quartile 2
σ	: Standard derivative
H	: Half width
•	: Input samples
○	: Least-squares output sample
×	: Effective impulse response samples
i	: Original element of the spectrum
j	: Variable
$\bar{x}_i$	: Mean of spectrum i
n	: Number of variables or wavelengths in the spectrum

d_{ig}	: Euclidean distance
S_p	: Variance–covariance matrix
$\boldsymbol{X}$	: Data matrix
$\boldsymbol{T}$	: Score matrix
$\boldsymbol{P}$	: Loading matrix
E	: Residuals
x_s	: Sample vector
w	: Weight vector
N_b	: Neighboring map units
w	: Scaling Value
I	: Corrected image
I_0	: Raw image
I_d	: Dark reference image
I_w	: reference image

CHAPTER I
INTRODUCTION

1.1 Introduction

1.1.1 Weedy rice problem

Rice is an important crop that serves more than half of the world population consumption as a staple food [1]. It is an essential nutritional resource and is globally grown on approximately 153 million hectares of land with 90% of the rice production area worldwide harvested in Asia [2]. Due to continuous growth of economies and population throughout the world, therefore high-yields, high-quality and genuineness of rice are needed [3]. In cooperate farming, the rice seed with authentical family gene is supplied by either the large registered companies or government associations in order to plant in the large-scale agriculture farms and avoid seed mutagenesis. However, especially in developing countries, the small farmer communities with local plant fields contribute on the majority up to 70-95% of the farming population [4, 5]. The traditional image of farm households reflects that they focus on rice farming with various types on the same rural farm area. Moreover, they trend to reduce planting costs by collecting the harvested rice seeds in order to re-plant in the next season. These traditional behaviors including human selection and out crossed hybridization with some wild rice species cause random mutations on the harvested rice product. [6, 7]. This makes the rice seed contains high level of genetic diversity even at a local level and it is impossible to differentiate and identify the quality and authenticity of rice paddy seeds by human visualization platform.

There are generally two species, namely *Oryza sativa* and *O. glaberrima* as the most popular cultivated rice. They are more widely distributed and are produced in Asia, especially in Thailand [7]. Weedy rice (*Oryza sativa f. spontanea*) is one of the aftermaths of the mutagenesis. It is the most notorious weed occurring in rice-growing areas worldwide and the problems are still prevailing. The weedy rice can be called as "red rice" (as shown in Figure 1.1) because of its appearance in the red pericarp [1, 8]. However, the weedy rice phenotypes with white, light red, and light green pericarps could be also found [2]. Weedy rice may be defined explicitly as unwanted rice that infests and competes with rice and alternate crops. They could be fast outbreaks

through the field as they are generally taller, have a higher growth rate, and produce more tillers than cultivated rice [8].

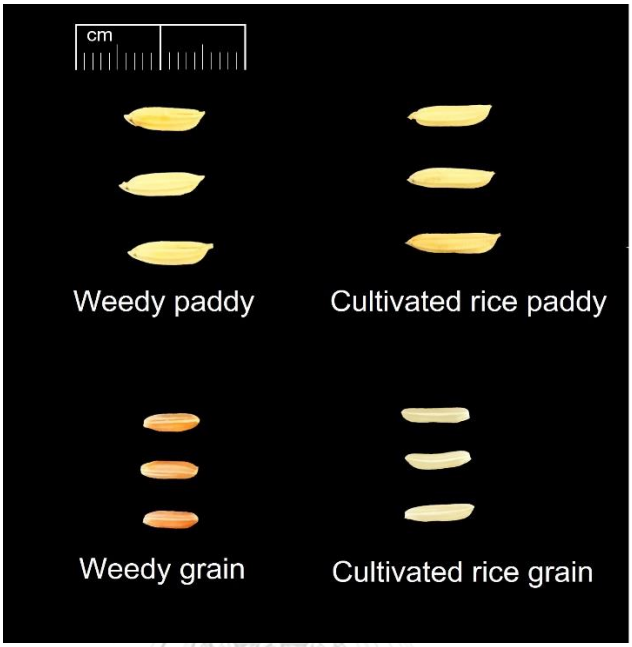


Figure 1.1 Physical appearance of weedy rice, which can be called as "red rice" from paddy seed and grain seed compared with the cultivated rice

Nowadays, the spreading of the weedy rice is now becoming a serious problem found in local rice-planting areas all over the world, especially the areas where the rice seeds are directly transferred to be planted for the next season [9, 10]. This is not only affected to the household planting but also influences to the rice industry for separating them from the regular cultivated rice. The presence of weedy rice in the fields diminishes the farmer's income both quantitatively through reduction of grain yield and qualitatively through lowered commodity value at harvest. As a result, it leads to loss of economic benefits –particularly in terms of the additional costs associated with managing the rice crisis [11]. In Asian countries, weedy rice has been reported to reduce rice yield from 16% to 74% [12].

1.1.2 Strategies for controlling weedy rice

Therefore, in order to avoid the negative consequences of the weedy seed outbreak, it is necessary to conduct and control seed quality using effective control methods [2, 13]. There are several standard methods to elucidate authenticity, contamination, and genetic of rice cultivation, such as the DNA finger-printing method [3], enzyme linked immunosorbent assays [14], high performance liquid chromatography [15, 16]. These methods offer high precision and accuracy; however, they are also burdensome, high cost, complicated operation, and time consuming [14]. They may not be appropriate for testing a large number of samples. On the other hand, there are still have many particular methods which are straightforward, and no requirement of high-class instruments. For example, preventive method including certified seed, cleaning of machinery and field inspections [17]. However, high investment cost leads to the limitation of these procedures. Other cultural methods include soil tillage, burning of stubble and straw, stale seedbed preparation, water seeding, transplanting, and crop rotation [18]. However, all these methods have many deficiencies, such as multi-stage process, extreme supply source and intensive labor. The mechanic method [19] is also a useful approach; unfortunately, it has several drawbacks, such as being time-consuming and ineffective compared to chemical treatment. To get more details, different strategies and methods to control weedy rice were summarized in Table 1.1

Table 1.1 Strategies to control and manage weedy rice in the real plant field

Control Strategy	Control Method	Disadvantages	Ref
Preventive	▪ Certified seeds ▪ Cleaning of machinery ▪ Field Inspections	▪ High production costs	17
Cultural	▪ Soil tillage (minimum tillage) ▪ Fallowing ▪ Burning of stubble and straw ▪ Stale seed bed preparation ▪ Water seeding ▪ Water management ▪ Rice variety ▪ Hand weeding ▪ Transplanting ▪ Crop rotation	▪ Multi-stage process ▪ Extreme supply source ▪ Intensive Labor	18
Mechanical	▪ Before rice planting ▪ After rice planting	▪ More time consuming and significantly less effective than chemical treatment	19
Chemical	▪ Pre-plant application ▪ Post-plant application	▪ Ineffective result because weedy rice and cultivated rice belong to the same biological species	18
Genetic	▪ Biotechnology: Herbicide-resistant rice varieties ▪ DNA finger-printing method	▪ Require highly expert ▪ Time consuming	20

1.1.3 Literature reviews on spectroscopy

To overcome the drawback of traditional strategies mentioned above, an invasive technique to reveal the authenticity of rice seed is necessary. Near infrared

spectroscopy (NIR) is one of alternative methods for agricultural quality assessment. It is a simple, rapid, non-destructive [8, 21-24], and environmentally friendly technique that has been used in several fields especially in quality control of agricultural products [25, 26] and food processing [14, 27].

Specifically, in previous studies for rice seed quality evaluation, NIR spectroscopy has been used for seed purity analysis [28], seed cultivar identification [29], and seed authenticity detection [30]. However, the spectral information obtained by this method is theoretically confined to only a tiny section of a sample where the measuring probe is positioned, without taking into account the spatial information [31]. Therefore, in order to recover a representative spectral fingerprint of the entire sample, the sample being studied should be sufficiently homogeneous. This drawback of traditional spectroscopy can be easily alleviated by adopting near-infrared (NIR) hyperspectral imaging (HSI) techniques to incorporate spectral and spatial information.

In 2011, Fernando *et al*. [32] integrated spectral and image analysis of hyperspectral image for prediction of apple fruit firmness and soluble solid contents. In 2013, Wenqian *et al*. [33] performed hyperspectral imaging with the NIR wavelength range of 1000 to 2500 nm to detect bruises on 'Fuji' apples. Qin *et al.* [34] established a small-feature of hyperspectral reflectance imaging for real-time detection of grape canker, but the captured image only provides a small area of observations from the whole fruit. Although the NIR hyperspectral imaging system is an effective technique, it provides the massive data volume, complex spectral information, and time consuming to acquire and collect the data. Multispectral imaging based on selected critical wavelengths has received great attention. Due to their relative small size of spectral data, low instrument cost and high analytical speed, multispectral imaging systems could be widely used in on-line detection and be applied in manufacturing scale for agricultural products [35]. Huang *et al*. [36] selected three effective wavelengths 750, 820 and 960 nm to detect the bruises of apples. However, few selected and discrete wavelengths would be carefully determined because they may be one of the reasons that causes worse performances on the detection of agricultural product characteristics [37].

There are many applications of NIR spectroscopy as well as multispectral and hyperspectral imaging technologies for measuring quality seed quality assessment.

In 2020, Su *et al*. reviewed that the strategy of utilizing both spatial and spectral information in the discriminating stage has been proposed to improve the existing state of weed detection. Combining NIR spectroscopy and hyperspectral imaging provides richer spatial and spectral data, and it demonstrates a more vital ability to distinguish between crops and weeds [38]. This is consistent with a review by ElMasry and Nakauchi that machine vision has currently been used as a proficient inspection technique for visualizing the inherent physicochemical qualities among many food products during quality and safety assessments; in addition, it also provides geometrical, textural, and aesthetic elements [39]. With the progress of current science and technology, NIR hyperspectral technology has been applied in a variety of industries, including quality assessment, seed quality identification [40, 41], seed authenticity detection [42] and agricultural product quality breeding [43]. NIR hyperspectral technology has been proven to be a fast and non-destructive multi-component analytical approach allowing many determinations simultaneously without extensive sample preparation [43, 44].

Furthermore, hyperspectral imaging evaluation frequently generates a vast amount of high dimension data, which involves the use of effective chemometric techniques to extract and understand critical and insightful information [44, 45]. Furthermore, hyperspectral imaging evaluation frequently generates a vast amount of high dimension data, which involves the use of effective chemometric techniques to extract and understand critical and insightful information [44, 45]. In particular, there are several mathematical approaches such as linear discriminant analysis (LDA) [46], partial least squares discriminant analysis (PLS-DA) [47], the *k*-nearest neighbors (*k*NN) [48], support vector machine (SVM) [49], principal component analysis (PCA) [50], and artificial neural networks (ANNs) [38] to be performed on the complex NIR dataset in order to visualize, estimate, predict and classify product quality as shown in Table 1.2

Table 1.2 Literature reviews of quality assessment on agricultural products by using NIR combined with chemometrics.

Year	System	NIR mode	Chemometrics		Ref
			Method	Number of latent variables (LVs)	
2012	Determine antioxidant activity of bamboo leaf extract	Reflectance	PLS	5	51
2014	Discriminate between red wines of different designation of origin	Transmittance	PCA SVM LDA	2	52
2017	Predict the internal quality index in variety nectarine samples	Reflectance	PLS	7	53
2019	Distinguish almonds cultivated in the Avola area from others presenting a different geographical origin	Reflectance	PLS-DA SIMCA	4	54
2019	Predict quality and maturation stage attributes of wine grapes	Reflectance	PCR PLSR	2 7	55
2019	Discriminate green tea with grades, varieties and geographical origins	Reflectance	PCA SVM	5 3	56

**Note PLS-DA : Partial least squares-discriminant Analysis ; SIMCA : Soft Independent Modelling of class Analogies, PCR : Principal component regression ; PLSR : Partial least square regression, PCA : Principal component analysis ; SVM : Support vector machine, PCA-LDA : Principal component analysis combined with linear discriminant analysis*

Determination of the significant number of latent variables (LVs) in multivariate analysis techniques is one of the important steps to build the statistical model with high efficiency [57]. The number of latent variables should be re-determined whenever the new set of samples is added in the calculation in order to keep the prediction viability. The extra added samples would gather the variance in the dataset; therefore, the calibration model needs to be re-generated. The calibration model from these stated methods is based on the defining model, which is valid at the time, but it needs to be modified as time goes on. These are a kind of impact limitation of the

"progressing time window" research to be applied in a practical way because the data may be ever-increasing, generated in the period of time and the developed model might not be valid when time passes quickly.

A more concrete picture is shown in the Figure 1.2, the reduction methods such as principal component analysis (PCA), which require re-calculating the number of latent variables (LVs) every time a new set of samples is added to maintain prediction viability. Furthermore, the classification strategy involves either linear (EDC, LDA, PLS-DA) or non-linear (QDA, SVM, KNN), each of which has its own set of difficulties to be calculated. The selection of classification method is the pain point when the underlying data is difficult to defined as linear or non-linear, therefore, an inappropriate classifier will lead to the overfitting problem.

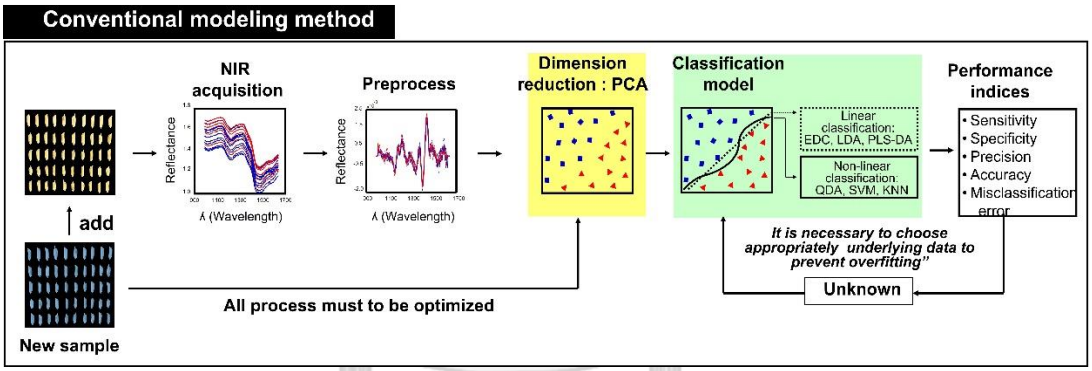


Figure 1.2 Conventional modeling method

SOMs (Self-Organizing Maps) is artificial neural networks (ANNs) that are not dependent on latent variables. Furthermore, the technique can be applied to data that is either linear or non-linear [58-60]. SOMs is extensively performed on complex relationships between the samples (NIR spectra in the case) which can be revealed through by only the single map. It simplifies the analysis allows multivariate exploratory comparisons between sample-to-sample by direct visual inspection [61-63]. Recently, SOMs have recently become a popular and powerful technique for analyzing multicomponent data especially agricultural product assessment, In 2006, Lin and Wang [64] compared SOMs with various hierarchical cluster analysis methods. The results showed that the performance of SOMs to cluster groups of samples is better than other hierarchical clustering methods. In 2008, Siripatrawan [65] showed the

application of electronic nose sensors combined with SOMs as an effective feature extraction method to determine the foodborne pathogens contamination in a packaged fresh vegetable. In the same year, Meunkaewjinda *et al*. [66] presented automatic plant disease diagnosis using multiple artificial intelligent techniques including self-organizing feature map together with a back-propagation neural network. In 2017, Luna *et al*. [67] presented a study of chemometric tools for the classification of Coffea canephora (whole beans) cultivars via in situ direct sample analysis using near-infrared spectroscopy (NIR). The result showed that SOMs are highly effective method which can provide 100% correct identification of testing samples. In 2019, Theanjumpol *et al*. [68] detected and estimated the occurrence of granulation in 'Sai Num Pung' tangerine based on the use of near infrared (NIR) spectroscopy and difference classification models. The results revealed that the classification results from supervised self-organizing map (SSOM) could provide the best predictive granulation level.

To our knowledge, there are no studies on the use of NIR technique combined with SOMs to distinguish between cultivated rice and weedy rice directly from paddy seeds. In this study, a modified algorithm of self-organizing maps (SOMs) network architecture is developed in the form of a global map model. The global map model was calculated using reference samples (weedy rice and cultivated rice paddy seed in the case). The prediction and classification of unknown samples from other sources can be performed using the supervised SOM map model without any regeneration. When the unknown sample is introduced, the map model will start to learn and search the best matching unit (BMU). If BMU of an unknown sample is determined and give the larger value in the region of "weedy rice", unknown sample will be assigned to the class of "weedy rice". On the other hand, If BMU of an unknown sample is determined and give the larger value in the region of "cultivated rice", unknown sample will be assigned to the class of "cultivated rice". The following section of the experiment will be undertaken utilizing the NIR spectroscopic approach combined with the global SOM map after empirical evidence demonstrates that weedy rice can be discriminated from cultivated rice (discussed in Chapter 3). In this second part, weed rice will be distinguished from cultivated rice directly from the object on the image. This process has been simply done by projection the color shades e.g. red,

green, and blue to the image in order to visualize the types of the image object. This proposed approach is more practical to be used in the real situation. It is the first attempt in NIR hyperspectral imaging applications in seed quality monitoring by using actual HSI data for the analysis (discussed in Chapter 4). The overview of the research theme is diagrammatically summarized in Figure 1.3.

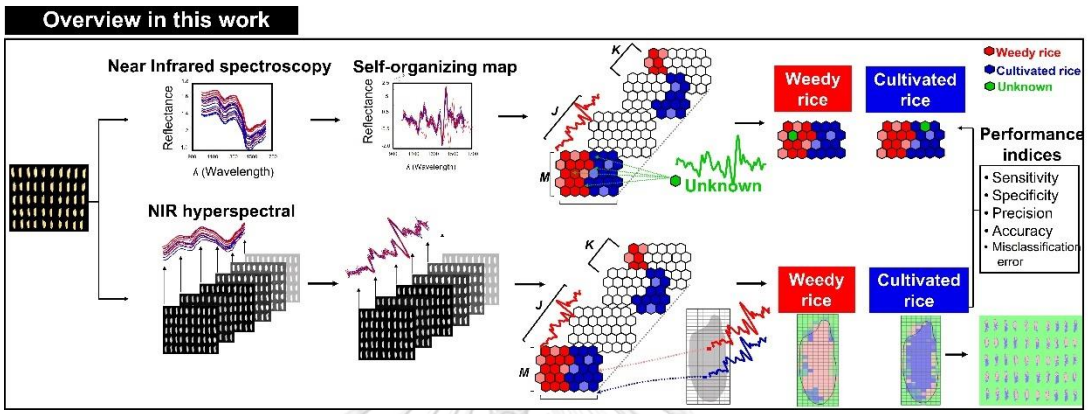


Figure 1.3 Overview of the research theme involving NIR and NIR-hyperspectral imaging combined with modified SOMs as the global map for rice seed classification inspection

1.2 Objective of this work

To modify the classifier model based on supervised SOMs in order to discriminate the weedy rice from cultivated rice directly from paddy seed for seed quality assessment

1.3 Scope of this work

1. Raw rice products from trustworthy organizations will be emphasized on the study.
2. The collected seed samples will be pre-treated by using cyclone vacuum machine to remove the contaminated particles and other impurities attached on the rice seeds.
3. All seed samples were kept in the vacuum boxes at room temperature prior to perform the NIR measurements.
4. Thermo Scientific™ Nicolet™ iS5N FT-NIR spectrometer with extended range indium gallium arsenide (InGaAs) detector, high intensity halogen light source and temperature stabilized solid-state Near-IR diode laser

purchased from thermo fisher scientific will be used to acquire NIR spectra of the seed samples.

5. hyperspectral images of the rice samples were acquired by using one push broom HSI which comprises an imaging spectrograph (Imspector N17E; Specim, Spectral Imaging Ltd., Oulu, Finland), a CCD camera (Xeva 992; Xenics Infrared Solutions, Belgium), two 500 W tungsten–halogen light sources (Lowel Light Inc., NY, USA), and control software (Specim's LUMO Software Suite; Spectral Imaging Ltd., Oulu, Finland).
6. Unsupervised Pattern Recognition, Supervised Pattern Recognition and Variable selection will be performed using SOMs as a statistical method to deal with the complex dataset.
7. All statistical and mathematical calculation will be performed using program MATLAB version R2019b with in-house coding algorithm. Recommended & minimum computer configurations: processor CPU (Intel Core i5, sixth generation or newer or equivalent), operating system (Microsoft windows 10 professional x64), and Memory (16 GB RAM).

1.4 Benefit of this dissertation

A powerful invasive, green and simple techniques which could be performed fast and accurate without using extra chemicals and process required to assess and inspect of rice seed quality.

CHAPTER II
THEORETICAL BACKGROUND

2.1 Near infrared spectroscopy

Near infrared spectroscopy (NIR) is well-established as a rapid and non-destructive analytical technique [69]. As shown in Figure 2.1, the record regions (780–2500 nm) contain 1st-3rd overtones as well as combinations of fundamental vibrations of C-H, O-H, and N-H chemical functional groups, which are the principal constituents of agricultural products. Although their molar absorptivity are low and detection limits are around 0.1% but they are adequate to be used for original tracing, adulteration, authenticity and discrimination detection in agricultural products [70, 71].

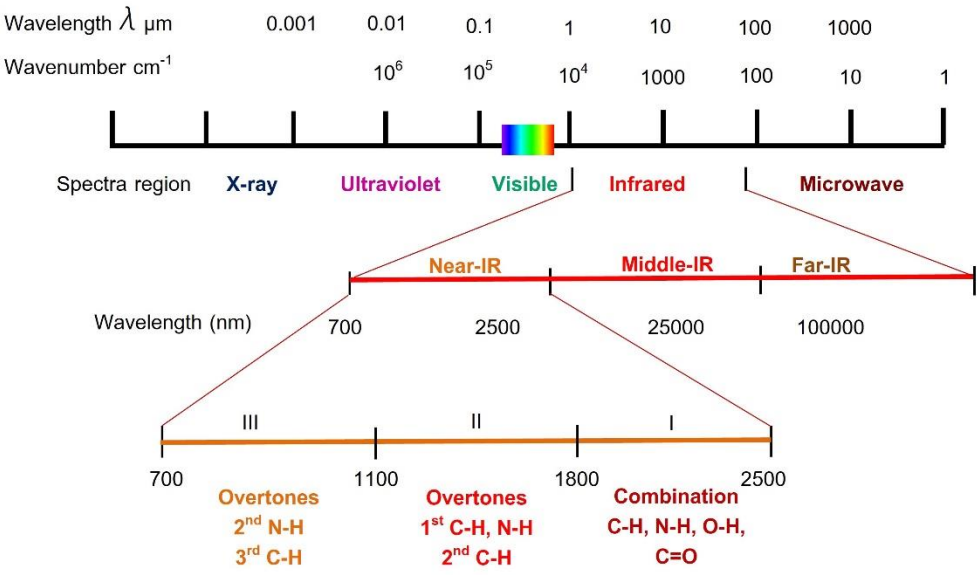


Figure 2.1 the range of electromagnetic radiation in UV (10 nm to 400 nm), visible (400 to 700 nm), infrared (700 nm to 1 mm) and NIR (700nm-2500 nm)

There are three modes of data collection including reflectance, transmittance, and transflectance. Different spectral modes have been assigned for different samples depending on the types, physical properties, and characteristics of the samples [72]. The transmittance mode measures the amount of light transmitting through the sample,

which is usually used for the analysis of liquid samples and certain solid samples such as grains, meat, and dairy products [72]. In case of the reflectance mode, the radiation light is reflected from the sample surface and return to the detector which is situated at the critical angle of the light source to capture the reflected light from the sample. This mode is usually used for solid or granular samples [73]. In case of transflectance mode, it is a combination of reflectance and transmittance. It is in doubling the optical path as the radiation beam passes twice through the sample [71] which is suitable for internal disorder detection such as detecting brown spots in pear [74]. The schematic figure of the three data collection modes is shown in Figure 2.2.

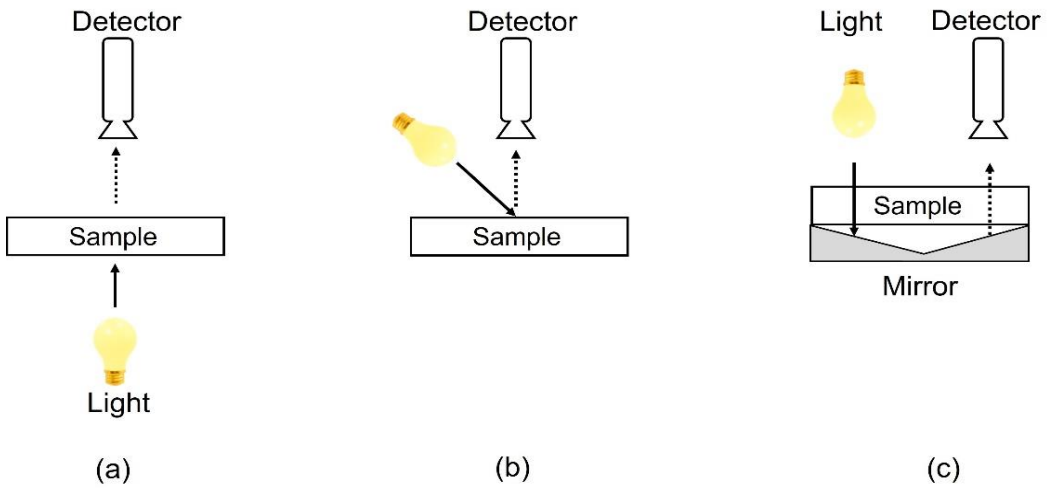


Figure 2.2 Mode of data acquisition using NIR spectroscopy involving (a) transmittance; (b) reflectance and (c) transflectance [72].

NIR spectroscopy has been continuously developed as a powerful technique for assessing internal and external quality attributes of agricultural products [75]. The NIR data was acquired as a single spectrum which represents whole underlying chemical distribution on the sample. In case of non-homogeneous samples, the spatial information of the sample could not be collected. Recently, a potential of NIR spectroscopy has been further developed including multi- and hyperspectral imaging techniques which also provide both spectral, spatial information and time-resolved spectroscopy which allows measurement of absorption and scattering processes

separately [76]. The images can be analyzed and visualized as chemical images, providing identification as well as localization of chemical compounds in non-homogenous samples such as seeds, grain, fruit and meet [77].

2.2 NIR hyperspectral imaging

Regarding of detecting the molecular bonds in the sample, the HSI technique in association with near infrared (NIR) spectroscopy is commonly used to identify or inspect various seed components [43]. The potential of NIR hyperspectral imaging techniques can provide both spectral and spatial information [76], enabling chemical image analysis and visualization, as well as the identification and localization of chemical compounds in nonhomogeneous sample [77]. Before entering the various image operations which can be performed to hyperspectral images, it's vital to understand how these images are constructed and what are the parameters of the system that created them. Based on the relative movement between the sample and the detection unit, a hyperspectral image can be obtained in three different ways (i.e. the camera and spectrograph): point-to-point spectral scanning (whisk-broom imaging); line-by-line spatial scanning (push-broom imaging); and area scanning (staring imaging or wavelength scanning) [78]. Since continuous scanning in one direction has been used, line scanning is therefore exceptionally suitable for food quality monitoring and safety inspection in conveyor belt systems [79].

The setup of a NIR hyperspectral imaging system (1000 – 2500 nm) is shown in Figure 2.3a [80, 81]. It consists of a CCD camera, a spectrograph, a standard C-mount lens, an illumination unit (tungsten halogen lamps), a translation stage and a computer supported with a data acquisition software. The sample is scaned with pixels by pixels. The reflected light of each pixel is recorded as a NIR spectrograph and then the sample image is captured by the CCD array detector [81]. The collected image data is presented in the form of a three-dimensional matrix called a hypercube. This hypercube consists of row and column of pixels that each pixel represents a NIR spectrum as a depth profile. The spectrum of each pixel can be visualized and the image plane at each respective wavelength can be revealed [82]. The obtained hypercube with its spatial and wavelength dimensions contains an NIR spectrum for each pixel is shown in Figure 2.3b.

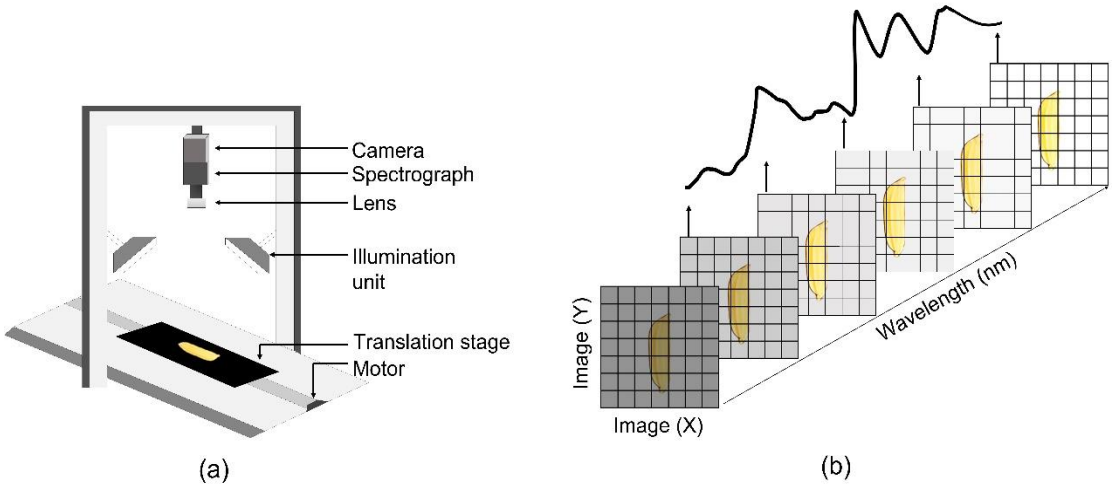


Figure 2.3 (a) Scheme of the main parts used in NIR hyperspectral imaging system. (b) Illustration of an NIR hyperspectral imaging hypercube comprising wavelength (depth profile) and spatial (x-and y- pixels) dimensions.

Hyperspectral imaging technique evaluates the product quality based on integrated computer image processing by combining color and spectral images. The color imaging system will be used to measure morphological characteristics (dimensions, color, shape, texture, etc.), germination ability (radicle elongation, timing of germination, germination speed, vigor, and so on), and seed disorders (infected parts, pest attacks, abiotic stress, and so on). In accompanying with color imaging system, the spectral image processor will be used in conjunction with the color imaging system to offer spectrum information about the investigated seeds in order to provide comprehensive information about the chemical components (protein, lipid, moisture, pigments, and so on) of the seeds [31].

2.3 Chemometrics

NIR hyperspectral technology has been shown to be an efficient technique for seed quality assessment that allows simultaneous measurements without requiring sample preparation [83]. Nonetheless, a single hyperspectral image can consist of up to 200000 spectra, chemometric techniques are therefore required to handle such large data sets [82]. Chemometrics (or multivariate data analysis) is an approach for manipulating and extracting useful information from spectral data using mathematics and statistics. Apart from obtaining relevant information, Mathematical procedures can be used to eliminate unwanted information (such as spectrum noise or particle size impact) without sacrificing critical or required data [71, 82].

2.3.1 Preprocessing

2.3.1.1 Interquartile range (IQR)

The Interquartile Range (IQR), commonly known as the middle 50%, is a percentage ranging from the 25th to the 75th percentile [84]. IQR formular is shown below (Eq. 1):

$$Q3 - Q1 = IQR \quad (1)$$

where,

IQR = Interquartile range (IQR = Q2)

Q1 = (1/4)[(n + 1)]th term)

Q3 = (3/4)[(n + 1)]th term)

n = number of variables

As it is a statistical dispersion measurement, the interquartile range can be effectively used as an indicator to identify outliers. They are observation that occur outside that fall below Q1 - 1.5 IQR or above Q3 + 1.5 IQR as shown in Figure 2.4 [85].

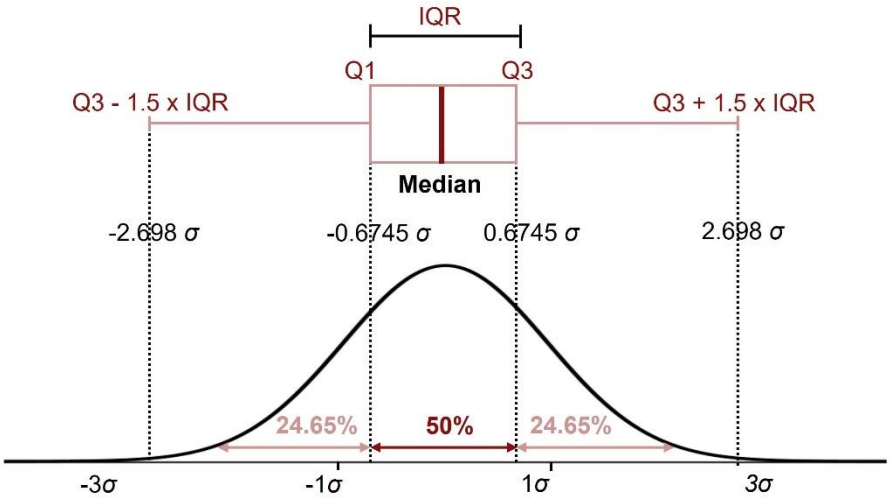


Figure 2.4 Interquartile range (IQR) projection on a normally distributed density. The median of IQR the equivalent to the mean 0 σ. The value IQR = Q3 – Q1 corresponds to 50% of the density distribution and the first quartile corresponds to −0.67 of the population while the third quartile corresponds to +0.67

2.3.1.2 Savitzky-Golay smoothing

Savitzky-Golay filter (SG) is linear and shift-invariant. It performs on a vector of input sample x[n] to generate a smoothed output vector y[n]. Figure 2.5 illustrates the basic concept of least-squares polynomial smoothing by portraying a signal sequence of samples x(n) as solid dots. In consideration of the group of 2M+1 samples centered at n = 0 as a starting point, the coefficients of polynomial is obtained [86].

$$p(n) = \sum_{k=0}^{N} a_k n^k \qquad (2)$$

where $p(n)$ is least-squares polynomial smoothing function, a_k is least-squares principal, n is number of interval shifts.

The deliberation is the same for any other different collection of 2H+1 input samples, where H refer to the “half width” of the approximation interval.

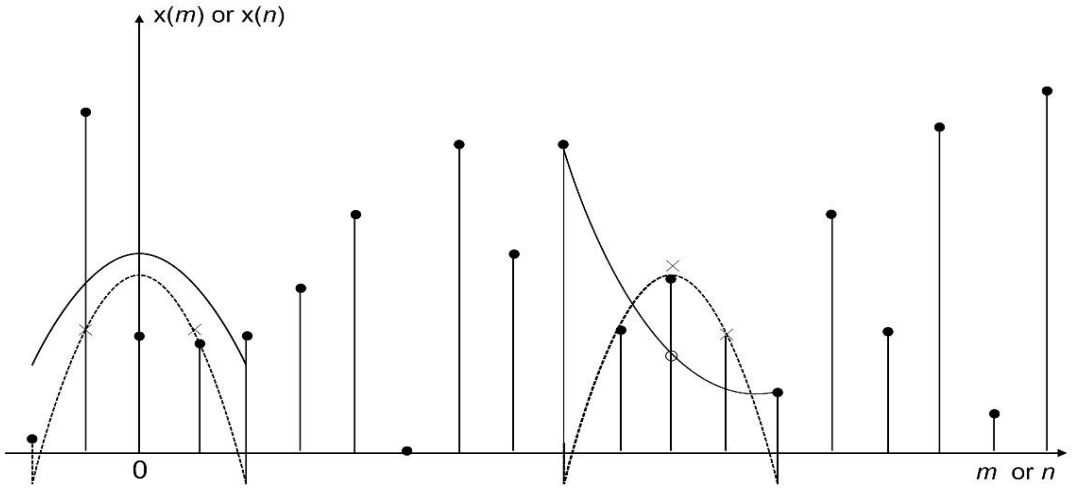


Figure 2.5 Illustration of least-squares smoothing by locally fitting a second-degree polynomial (solid line) to five input samples: • denotes the input samples, ○ denotes the least-squares output sample, and × denotes the effective impulse response samples (weighting constants). (The dotted line denotes the polynomial approximation to centered unit impulse.)

In particular, the approximation interval does not have to be symmetric around the evaluation point resulting in nonlinear phase filters. This may be beneficial for smoothing at the ends of finite-length input sequences. In consonant with output at the next sample, it is retrieved by shifting the analysis interval to the right by one sample, revising the origin to be the position of the middle sample of the new block of 2*M*+1 samples, and repeating the polynomial fitting and evaluation at the central point. This can be accomplished for each input sample, leading to a new polynomial and a new value for the output sequence *y[n]*

$$y(n) = \sum_{m=-M}^{M} h[m]x[n-m] \tag{3}$$

where *y(n)* is new polynomial and a new value for the output sequence, *M* fitting sequence with 2M+1, and *h[m]* is finite impulse response value, which was adopted as a weighted value.

The value mark with × (in Figure 2.5) are the shifted impulse responses $h[0 - m]$ that might be employed to calculate the output samples labeled with ○, thereby substituting the polynomial fitting procedure at each sample with a single assessment of Eq 3.

2.3.1.4 Standard Normal Variate (SNV)

Standard Normal Variate (SNV) is one of well-known preprocessing technique which used to reduce the scattering and multiplication effects of particle sizes as well as disparities in the global intensities of the signals [87]. Each spectrum is scaled by dividing its standard deviation by its center as shown in Eq 4.

$$x_{i,j}^{SNV} = \frac{(x_{i,j} - \bar{x}_i)}{\sqrt{\frac{\sum_{j=1}^{n}(x_{i,j} - \bar{x}_i)^2}{n-1}}} \quad (4)$$

where $x_{i,j}^{SNV}$ is the element of the transformed spectrum and $x_{i,j}$ is the corresponding original element of the spectrum i at variable j, $\bar{x}_i$ is the mean of spectrum i, and n is the number of variables or wavelengths in the spectrum.

SNV is particularly remarkable because the transformation is carried out on individual samples. However, it should be noted that SNV procedure is related to sum of the deviation of absorbance at individual wavelengths, artificial negative correlation can be occurred [88].

2.3.1.5 Second derivative

Derivatives are a technique for addressing with two common issues in NIR spectra: overlapping peaks and excessive baseline variations. According to the increment of linear baseline in NIR spectra, the second derivative tends to be preferred because it contain negative peaks where the original had a peak resulting in simple interpretation. Second derivative is performed to extract hidden information in the spectrum and eliminate baseline effect. Integer derivatives, on the other hand, lack

the sensitivity to gradual alterations in slit and curvature, resulting in noise introduction and information loss [89].

The general method of calculating derivative is shown Eq. 5 [90]. If A is a spectrum defined for evenly spaced wavelength λ_n, n = 0,1,…,N-1, then the first derivative A'_n at point n is defined by:

$$A'_n = A_{n+g} - A_{n-g} \quad (5)$$

where g is an integer called the gap or derivative size and A_n is the NIR value at point *n*. Similarly, the second derivative are defined in Eq 6.

$$A''_n = A_{n+2g} - 2A_n + A_{n-2g} \quad (6)$$

2.3.2 Principal Component Analysis

Principal component analysis (PCA) is one of popular tool from multivariate statistics that help to drastically reduce dimensionality in a large dataset, while that most of the crucial information is preserved [91]. Basically, PCA was used to extract the major component from data matrix base on two main concepts: the number of meaningful PCs which ideally equal to the number of significant component (for example, if there are three components in the mixture, then only three PCs should be expected), and the other one is characterization of each PC by loadings and scores.

NIPALS (Nonlinear Iterative Partial Least Squares) is an ordinary, iterative algorithm frequently used for PCA [62, 92]. In short, it extracts components one at a time, and can be stopped after the desired number of PCs has been obtained. The steps are as follows:

1. Originate a data matrix $\boldsymbol{X}$ which is used for PCA.
2. Take a column of this matrix (often the column with greatest sum of squares) as the first guess of the scores first principal component; called $t_{initial}$
3. The loading vector p^{T}_{unnorm} is calculated: $p^{T}_{unnorm} = \dfrac{(t^{T}_{initial})(\boldsymbol{X})}{(t^{T}_{initial})(t_{initial})}$

4. The loading vector is normalized to unit length: $p_{norm}^{T} = \dfrac{p_{norm}^{T}}{(p_{unnorm}^{T})(p_{norm})}$
5. The new score vector is calculated: $t_{new} = (\boldsymbol{X})(p_{norm}^{T})$
6. Check for convergence by comparing the $t_{initial}\ and\ t_{new}$. The sum of squared differences between all elements of the two consecutive score vectors is calculated. If the value meets the criterion (small enough), this indicates that the PC has been extracted; otherwise, replace $t_{initial}\ with\ t_{new}$ and return to step 2, repeating until convergence is achieve
7. Subtract the effect of the new PC from the data matrix to obtain a residual data matrix:

$$\boldsymbol{E} = \boldsymbol{X} - tp^{T}$$

where $\boldsymbol{E}$ is residual data matrix, $\boldsymbol{X}$ is column with the greatest sum of squares (variance), *t* is score vector, *p* is loading vector, $\boldsymbol{T}$ is transpose.

8. If it desires to compute further PCs, substitute the residual data matrix for $\boldsymbol{X}$ and go to step 2.

The eigenvalue (*x*) for each component is calculated by the sum of the squares of the scores vector of all I samples:

$$ç_a = \sum_{i=1}^{I} t_{ia}^{2}$$

where $ç_a$ is eigenvalues, *I* is sample, *t* is score vector.

$$\sum_{i=1}^{I} ç_a = \sum_{j=1}^{J}\sum_{i=1}^{I} x_{ij}^{2}$$

where $\sum_{i=1}^{I} ç_a$ is the sum of all eigenvalues, $\sum_{j=1}^{J}\sum_{i=1}^{I} x_{ij}^{2}$ is equal to the sum of squares of the data matrix.

The significance of each PC can be determined using the percentage of the total amount of variance calculated by

$$\%ç_a = \frac{ç_a}{\sum_{j=1}^{J}\sum_{i=1}^{I} x_{ij}^2} \text{ x } 100$$

where $\%ç_a$ percentage of the total amount of variance.

2.3.3 Euclidean Distance (ED)

The Euclidean distance to centroids is a straightforward classification [93]. It is employed to measure the length of a line segment between the two points in Euclidean space. Owing to the fundamental concept of this method, the centroid of each class g $(\bar{x}_g)$ in a dataset are created. For each of the variables, the centroids are computed using the mean among all samples in a group. Beside the mean of each group, no further information regarding class distribution is available for this method, and it is presumed that the distribution of samples around the centroid is symmetrical [62]. The Euclidean distance of a sample $i(x_i)$ from class g is calculated as below:

$$d_{ig} = \sqrt{(x_i - \bar{x}_g)(x_i - \bar{x}_g)^T}$$

where d_{ig} is the Euclidean distance between sample i and the centroid of class g.

2.3.4 Linear Discriminant Analysis (LDA)

Linear discriminant analysis (LDA) is one of the most famous supervised method to extract discriminative features and expand to various variables. It involves a pooled variance-covariance matrix ($\boldsymbol{S_p}$) in distance measurement. The distance between samples to the class centroid is weighted based on the overall variance of each variable. Consequently, any correlation between variables (if present) is now properly considered. The Mahalanobis distance is used to calculate the LDA distance to the class centroid g as follows [62, 94]:

$$d_{ig} = \sqrt{(x_i - \bar{x}_g)\boldsymbol{S_p}^{-1}(x_i - \bar{x}_g)^{\mathrm{T}}}$$

where $\boldsymbol{S_p}$ is the pooled covariance matrix, calculated for two classes as follows:

$$\boldsymbol{S_p} = \frac{\sum_{g=1}^{G}(I_g - 1)S_g}{\sum_{g=1}^{G}(I_g - 1)}$$

where I_g is the number of samples in class g and $\boldsymbol{S_g}$ is the variance–covariance matrix for group g. It's imperative to note that the LDA approach uses the Mahalanobis distance relying on a variance–covariance matrix for the entire dataset, rather than for each class individually [62].

2.3.5 Quadratic Discriminant Analysis (QDA)

The correlation between variables is taken into account in the LDA method where the measurements are expected to be regularly distributed. QDA is comparable to LDA, except instead of utilizing overall pooled matrix, it uses the variance–covariance matrix of each class. Consequently, QDA does not infer that the variances of different classes have a similar variance–covariance matrix. The distance can be calculated as follows:

$$d_{ig} = \sqrt{(x_i - \bar{x}_g)S_g^{-1}(x_i - \bar{x}_g)^T}$$

where $\boldsymbol{S_g}$ is the variance–covariance matrix of class g.

It can be observed that the calculation of both LDA and QDA method are required variance–covariance matrix ($\boldsymbol{S}$) which result to a limitation of these methods. On the assumption of the number of variables in a dataset is greater than the number of samples, $\boldsymbol{S}$ will be a singular matrix that cannot be inverted. As state by EDC, LDA, and QDA, the class of a sample is assigned to the class with the minimum distance; moreover, they can be employed to classify multiple classes (two or more classes) [62].

2.3.6 Partial Least Squares Discriminant Analysis (PLSDA)

Partial Least Squares Discriminant Analysis is regression technique that functions by projecting the original data onto latent variable space. Because it intends to determine the best latent variables to represent the data, it is therefore similar to the PCA approach [94, 95]. With exception of PCA, however, PLS is a widely used method for determining the best latent variables describing the relationship between a data

matrix $\boldsymbol{X}$ (usually containing spectra or chromatographic data) and a class membership matrix $\boldsymbol{C}$ (usually containing quantitative values such as class labels or concentrations). The fundamental PLS-DA equations are as follows:

$$\boldsymbol{X} = \boldsymbol{T}\,\boldsymbol{P} + E$$
$$c = \boldsymbol{T}\,q + f$$

where $\boldsymbol{T}$ is common score matrix for this implementation. E and f can be considered residuals. In the following algorithm, the consecutive columns of the score matrix $\boldsymbol{T}$ (PLS components) are orthogonal, while the rows of the $\boldsymbol{X}$ loadings matrix $\boldsymbol{P}$ are not. On the other hand, the models with successive PLS components are additive since the scores are orthogonal [94].

To be noted that PLS-DA is a feature extraction and classification algorithm that perform better than PCA and LDA. One explicit reason is because PCA scores do not always explain differences between samples but rather variances in the spectral data [96].

2.3.7 Self-organizing Maps (SOMs)

SOMs (Self-Organizing Maps) were first introduced by Kohonen 20 years ago and are now extensively used to visualize sample relationships; moreover, it can reveal hidden patterns in the datasets [63, 97, 98]. The SOM is a neural network method that can be applied for both unsupervised and supervised learning. It is an effective alternative to PCA for visualizing data. Comparable to scores plots, a SOM map illustrates the relationship between samples and component planes that can be used to display distinguishing variables. SOM map is created using hexagonal or square units; however, only the hexagonal unit is described in this work. A weight (w) for each variable is contained in each map unit (u) on a SOM map, resulting in a 1 x J weight vector (note: J equals the number of variables in the dataset). A rough depiction is produced if the number of map units is small, meanwhile a better detailed map of the samples is provided if the number of map units is large [99]. For the sake of clarity, the SOM calculation algorithm for unsupervised learning is as follows:

1. An initial output map is established in ***M*** *x* ***N*** = ***K*** unit. A weight vector (*w*) of each unit will be randomly chosen from the maximum and minimum values of variable *j* in the input data. It's worth noting that the sizes of *M* and *N* should be carefully evaluated in order to cover the majority of samples that will be matched in the following step.
2. Sample vectors (x_s) in the dataset are then compared with the weight vector of each unit (w_k) on the initial SOM map from step 1. The Euclidean distance between x_s and w_k for each map unit *k* is calculate:

$$d_{sk} = \sqrt{(x_s - w_k)(x_s - w_k)^{\mathrm{T}}}$$

   Considering the rows of x_s as vectors, compute the distance matrix between each pair of vectors. This process will be repeated until the distance of ***K*** units on the map is calculated.
3. The map unit that gives the smallest distance will be declared as the best matching unit (BMU) of the chosen sample (x_s): $BMU = \min_k\{d_{sk}\}$
4. The BMU and the neighboring map units (N_b) within the length from the BMU are updated to become more similar to the sample vector x_s. The learning rate which is used to determine the amount that a map unit can learn to represent a sample in each iteration is calculated:

$$w_k = \begin{cases} w_k + w\alpha(x_s - w_k) & k \in N_b \\ w_k & k \in N_b \end{cases}$$

   where α is the learning rate and *w* is the neighborhood learning weight. The amount of learning decreases with each iteration of the algorithm, as does the neighborhood learning rate with distance from the BMU.

The learning of the entire process is repeated until the map regions are stable (for approximately 10,000 times) [100]. The overall of calculation protocol of SOM is illustrated graphically in Figure 2.6.

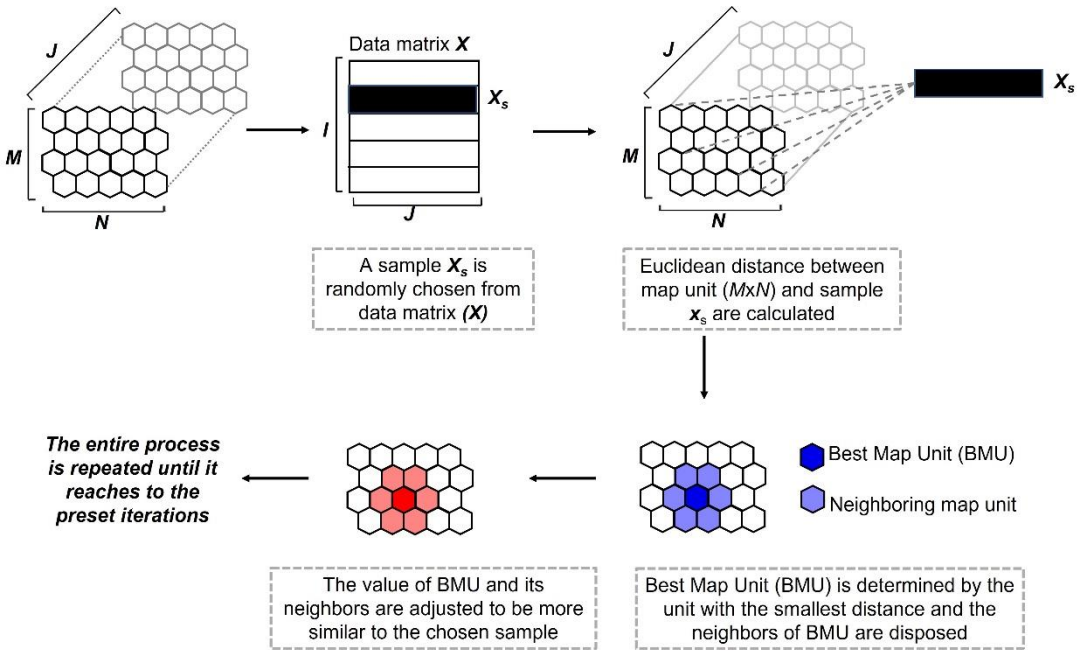


Figure 2.6 Calculation protocol of SOM for unsupervised learning

After the training process, the samples containing similar underlying information are closely mapped together. Samples originated from the same groups are assigned into analogous regions on the SOM map, while samples from different groups are laid on the different regions. For visualization, the color map is created to reveal the clusters of samples. The shading of the color map units is updated in each iteration which directly related to the updated SOM map. The color map will help in interpretation and so it is possible to monitor it in real time during the training process. The intensive details of SOM algorithm including the BMU, adjusted learning rate, neighborhood widths, etc. during the training process was already explained in our previous study elsewhere [62, 99, 101].

CHAPTER III
DISCRIMINATION OF WEEDY RICE USING NEAR INFRARED SPECTROSCOPY AND MODIFIED SELF-ORGANIZING MAPS (SOMS)

In the study, we aim to modify the classifier model based on supervised SOMs in order to discriminate the weedy rice from cultivated rice directly from paddy seed for seed quality assessment. The rice seeds involving weedy (red, ellipsoid and long tail) and four cultivated rice including Khao Hom Mali 105 (KHML105), Kor Khor 49 (RD49), Pratumtani1 (PTT1) and Phitsanulok2 (PL2) which were collected from trusted distributors certificated by Rice Department Ministry of Agriculture and Cooperatives, Thailand. Physical characteristic and thermal behavior of the rice samples were observed by optical microscope and thermogravimetric analysis (TGA), respectively. To access chemical information, the NIR spectra were acquired and examined by reflection NIR spectrometer. To discriminate class of rice sample, SOM classifier was generated with well optimization in order to prevent the overfitting problem. The performance of SOM classifier was validated with 100 different training and test sets to obtain the robust prediction. The classification performance was monitored by several indices including sensitivity, specificity, precision, accuracy, and misclassification error. It should be noted that the detection on the mixed proportion of weedy rice seems more significant than the classification in real application. However, the prediction of the mixed proportion usually could not be discovered until the classification of the target object (weedy rice) is completely achieved especially for the unknown system. In our study, to classify the weedy rice directly from paddy seed by using NIR technique has not been reported elsewhere so far. Therefore, to prove the capability of NIR technique combined with our modified SOM method in order to discriminate the weedy rice (the target object) from the cultivated rice is the first priority. This work could be further expanded to develop a multi-spectral camera instead of expensive standard laboratory instruments in order to reach a broad user community for seed quality assessments and inspections in the future. Further detail was shown in Appendices (Figure A2).

3.1 Experimental Setup

3.1.1 Sample collection and preparation

Three types of weedy rice (paddy seed) involving red weedy, ellipsoid weedy and long tail weedy were collected from the local fields at Phrom Phiram district, Phitsanulok province, Thailand. Moreover, the four types of standard cultivated rice seeds including Khao Hom Mali 105 (KHML105), Kor Khor 49 (RD49), Pratumtani1 (PTT1) and Phitsanulok (PL2) were collected from Lifestyle and Spirit of Thai Farmers-Nahai Chai Learning Center at Supanburi province, Thailand. They are certificated by Rice Department Ministry of Agriculture and Cooperatives, Thailand. All details of these samples are shown in Table 3.1. The collected seed samples were pre-treated by using cyclone vacuum machine to remove the contaminated particles and other impurities attached on the rice husk. After that, the seed samples were safely kept in the vacuum boxes at room temperature prior to acquire the NIR measurements. The cyclone vacuum machine was shown in appendices Figure A1.

Table 3.1 Information of collected weedy seeds and rice seeds from the certificated rice seed distributors in Thailand

Type of rice	Common name	Species	Plant Origin	Harvest date	Collecting date	NIR acquisition date	Number of Detection (Spectra)
Weedy	Red rice	*Oryza sativa f. spontanea*	Phitsanulok	1 Nov 2019	8 Nov 2019	22 Jan 2020	500
Weedy	Ellipsoid	*Oryza sativa f. spontanea*	Phitsanulok	1 Nov 2019	8 Nov 2019	24 Jan 2020	500
Weedy	Long tail	*Oryza sativa f. spontanea*	Phitsanulok	1 Nov 2019	8 Nov 2019	26 Jan 2020	500
KHML105	Hommali105	*Oryza sativa L.*	Phitsanulok	1 Nov 2019	8 Nov 2019	16 Feb 2020	500
RD49	Kor Khor 49	Oryza sativa L.	Suphanburi	17 Dec 2019	24 Dec2019	21 Feb 2020	500
PTT1	Phatumtani1	Oryza sativa L.	Suphanburi	17 Dec 2019	24 Dec 2019	24 Feb 2020	500
PL2	Phitsanulok2	Oryza sativa L.	Suphanburi	17 Dec 2019	24 Dec 2019	26 Feb 2020	500

3.1.2 NIR Spectral acquisition

Thermo Scientific™ Nicolet™ iS5N FT-NIR spectrometer with extended range indium gallium arsenide (InGaAs) detector, high intensity halogen light source and temperature stabilized solid-state Near-IR diode laser purchased from Thermo

Fisher Scientific was used to acquire NIR spectra of the seed samples. Regarding with acquisition process, we try to include all variations during the detection including variations from instrument, light scattering variation and sample variation. Owing to the experiment part, the samples were prepared in 5 different batches (variation from sample). In each batch, the rice sample was randomly rolled gently to obtain the uniform mixing. After rolling, the NIR spectra were continuously acquired for 10 repeated times without moving the sample container (instrument variation). The mixing process was repeated for 10 times for each batch (light scattering variation). Therefore, the total NIR spectra of each sample type was up to 500 spectra from 5 different batches $\times$ 10 mixing $\times$ 10 repeated detections. The background was re-measured in every 10 spectra. The scheme of the data acquisition with the brief explanation was added in Figure 3.1. Prior to data acquisition, the samples were prepared with the identical height and the surface of the samples were flattened to avoid the interfered scattering effects. Furthermore, the black box was used to cover the sample holder to avoid the error from external incident lights while the spectrum was acquired. The NIR spectra of the samples were collected using reflection mode in the range of 1000 nm–2400 nm with 16 averaged scans. Throughout the experiment, a room temperature was controlled at 27–29 ºC.

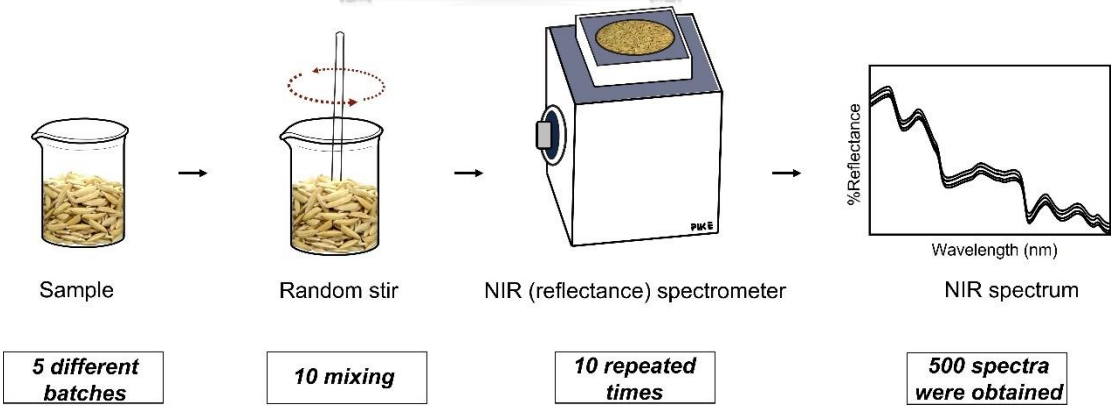


Figure 3.1 The scheme of the NIR acquisition procedure

3.1.3 Thermogravimetric analysis (TGA)

The thermogravimetric experiments were conducted by using Perkin Elmer Pyris 1 TGA Thermogravimetric Analyzer to reveal the thermal behavior of a sample. The system was carried out under the inert condition with a steady nitrogen flow of 20 mL/min. All samples were prepared in the range of 3–15 mg prior to be pyrolyzed. To remove the adsorbed water and moisture on the sample, the sample was firstly isothermal heated at 35ºC for 1 minute. After isothermal scan, the samples were continuously heated with rate of 20ºC/min from 50ºC to 800ºC.

3.2 Data analysis

3.2.1 Preprocessing method

In the first step of data analysis, the interquartile range (IQR), which is the difference between the 75th percentile and the 25th percentile, is used to detect outliers. The average NIR spectrum of each sample class was calculated as a centroid of the class. Euclidean distance of the NIR spectrum of in-class samples was then computed. Samples provide Euclidean distance outside 1.5 × interquartile range (IQR) from the average in-class NIR spectrum are identified as outliers and subsequently eliminated. [84, 102] After that, the spectra will be performed by the Savitsky-Golay smoothing coupled with standard normal variate (SNV) in order to effectively remove the signal variation from light scattering in the heterogenous samples [103].

3.2.2 Self-organizing maps (SOMs)

A self-organizing map (SOMs) is one of the most well-known artificial neural networks (ANNs) [98, 104]. The main principal of SOMs is its ability to not only transform multi-dimensional data into visually decipherable clusters in a low-dimensional grids (2D grids) form, but also maintain relative distances between existing data units in a multidimensional space form [61]. Basically, SOMs involves two processes including vector quantization and vector projection. At the beginning, SOMs was used as unsupervised model where only the predictive data was used for constructing the model [63, 99] To generate SOM map, it starts from initializing the map as represented by the two-dimensional hexagonal. Each map unit consists of the weight vector which was randomly generated from a uniform distribution between the

maximum and minimum intensities in the dataset. In particular, the number of map layers is set to be equal number of variables J (wavelengths in the case) in the dataset A map can be generated by $\boldsymbol{M} \times \boldsymbol{N}$ units which was normally set to approximately 2.5 times compared with number of samples in the dataset. During each training step, a sample vector $\boldsymbol{x}_s$ from data matrix $\boldsymbol{X}$ is randomly chosen and projected on each map. Euclidean distance between a sample vector and weight vector on each map unit is calculated. For each sample, the unit with the smallest distance to the chosen input sample is selected as a Best Matching Unit (BMU). The BMU and its neighboring units are updated to become more similar to the sample vector. The learning rate could be adjusted to determine the amount that a map unit can learn to represent a sample vector in each iteration. The iterations are repeated for approximately 10,000 times until the map regions are stable [100]. After the training process, the samples containing similar underlying information are closely mapped together. Samples originated from the same groups are assigned into analogous regions on the SOM map, while samples from different groups are laid on the different regions. For visualization, the color map is created to reveal the clusters of samples. The shading of the color map units is updated in each iteration which directly related to the updated SOM map. The color map will help in interpretation and so it is possible to monitor it in real time during the training process. The details of SOM algorithm including the BMU, adjusted learning rate, neighborhood widths, etc. during the training process was already explained in our previous study elsewhere [62, 99, 101] Most of SOM algorithm are applied in an unsupervised learning aspect similar to clustering, visualization, and dimensionality reduction. However, unsupervised SOMs cannot be used as a classifier in order to predict class of the unknown samples. Therefore, the supervised SOMs was developed in our previous study [59] in order to improve the capability of SOMs to be used for visualization and also classification. In supervised SOMs, the class weight vector (K) including information of class membership is added to the initial map. The dimension of the class weight vector is depending on the number of classes in the data, for example, if the data contains three classes involving A, B and C, then the class weight vector will be assigned as [w 0 0], [0 w 0] and [0 0 w], respectively. The class weight vector will be also trained during the iteration similar to the color and SOM maps. The separation between different groups of

samples in supervised method are strongly influenced by the Optimal Scaling Value (*w*). The samples are forced into predefined groups on the map when the large value of *w* is used. On the other hand, the class membership has little influence on learning process, and classes may not always be clearly separated when the small value of *w* is used. Therefore, the Optimal Scaling Value (*w*) is the important parameter which is needed to be optimized in the supervised method as it strongly affects to performance of the classifier. To predict class of an unknown sample, the BMU of the unknown sample is searched by assigned to the SOM unit with the smallest Euclidean distance. The class of the sample is assigned by the class with the largest value of class weight vector, for example, if class weight vector of BMU is [2.5 3.7 1.2], the class of sample will be assigned to class B as it provided the largest value. The scheme of supervised SOM algorithm is expressed in Figure 3.2 and the details of supervised SOM algorithm including the BMU, adjusted learning rate, neighborhood widths, optimized optimal scaling value etc. during the training process was already explained in our previous study elsewhere [59, 101]

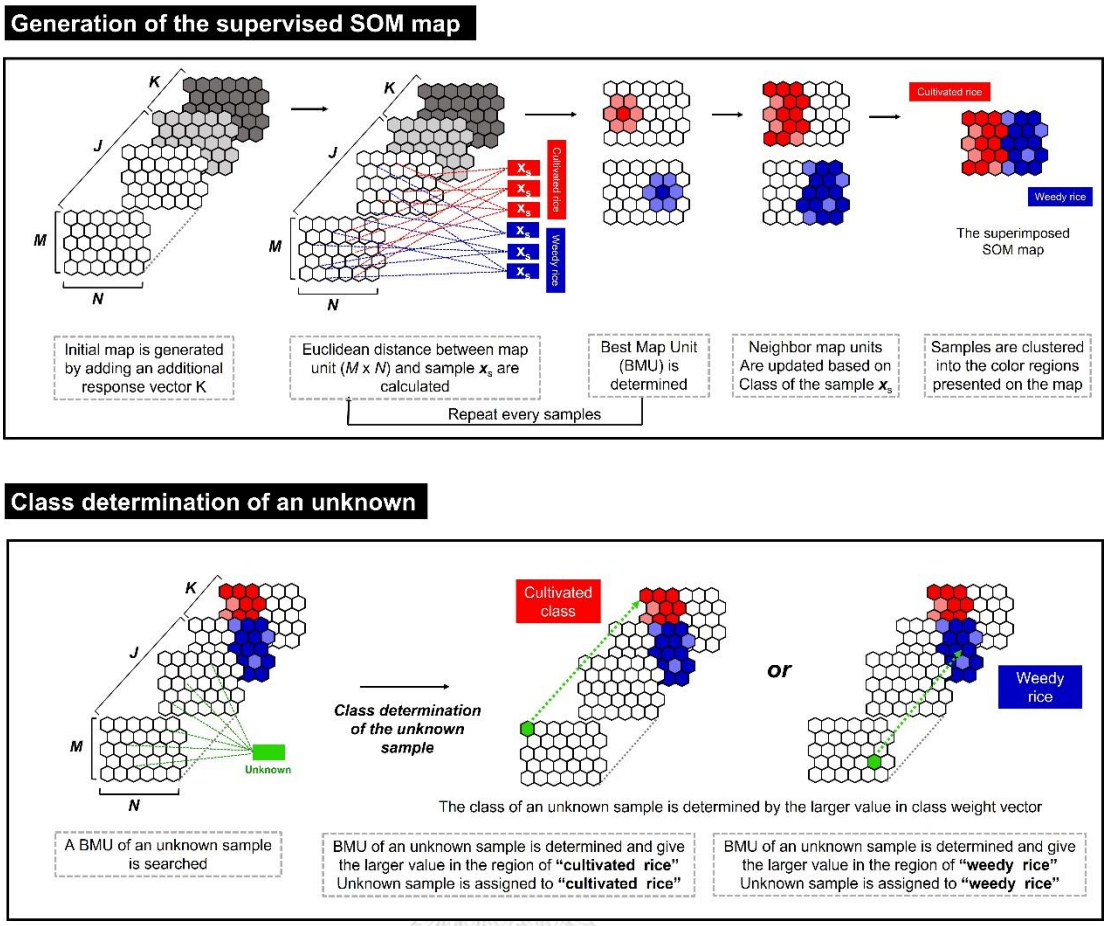


Figure 3.2 Schematic diagram for sample visualization and classification of weedy rice and cultivated rice using the modified supervised SOMs for *K* classes and *J* variables with the SOM map in dimension of *M* × *N*. The modified SOMs can be operated in two modes involving the training process of supervised SOM map to be used as reference map for the classification purpose and the class determination of an unknown sample by mapping the unknown to the reference SOM map.

In this work, validation was used to estimate the performance of the classifier by dividing the data into training set and test set. The supervised SOM map was constructed from the training set to predict the test set. Two-thirds of the dataset was randomly split into a training set and the remaining samples was assigned as test set. The training set samples were used to generate the supervised SOM map to be used as a classifier to predict the class of either training set or test set samples. This procedure

was repeated for 100 times using different random splits of the data, each time constructing different classifier model to reveal the robustness of the supervised SOM model [105].

3.3 Result and discussion

3.3.1 Rice seed characteristics

In order to visualize the features of rice seed, each type of rice seed including weedy rice and cultivated rice was photographed by digital microscope camera as shown in Figure 3.3. In case of weedy seed, they consist of red weedy, ellipsoid weedy and long tail weedy that their external appearances are slightly different. When they were compared with the cultivated varieties (KHML105, RD49, PTT1 and PL2), the weedy ecotypes were slightly shorter, rather dark-red pericarp and longer tail. However, it is undeniable that most weedy rice ecotypes are so phonologically and morphologically similar to cultivated rice varieties in term of shape and color. It causes difficulty in order to discriminate weedy rice from cultivated varieties directly from paddy seed by human visualization as shown in Figure 3.3(a)–3.3(g). In the experiment, a developed cyclone was used to remove contaminated particles on the rice husk surface. The rice husk with magnification of 100× captured by optical microscope of paddy rice seed before and after incubating in the cyclone was investigated as shown in Figure 3.3(h)–3.3(i), respectively. There are no significant differences in physical characteristics between with/without cyclone. It might be suggested that cyclone vacuum machine is appropriate to remove the external substance contaminated on rice husk and can keep their chemical characteristics, resulting the acquired spectra come from their intrinsic factors.

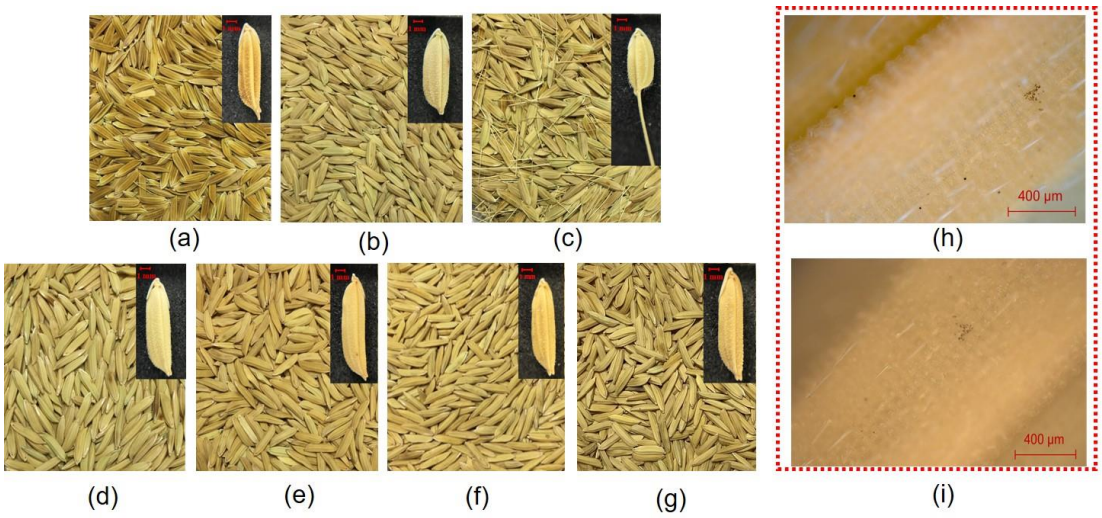


Figure 3.3 Morphological features of rice seeds including weedy rice seeds: (a) red weedy, (b) ellipsoid weedy, (c) long tail weedy and cultivated rice seeds: (d) KHML105, (e) RD49, (f) PTT1 and (g) PL2. The magnified optical images. On the right-hand side showed the optical microscope images (100×) of the rice (h) without cyclone (i) with cyclone.

The thermal stabilities and decomposition of the rice husk from weedy seed and cultivated seed were investigated using TGA which scans from 50ºC to 800ºC with temperature rate of 20ºC/min under N_2 flow in dynamic heating conditions. The thermal profiles could provide about physical and chemical phenomena including chemical compositions of the samples. Generally, the pyrolysis of any biomass can be divided into three phases including drying and evaporation of light components (phase 1), devolatilization of hemicellulose and cellulose (phase 2) and decomposition of lignin (phase 3) [106]

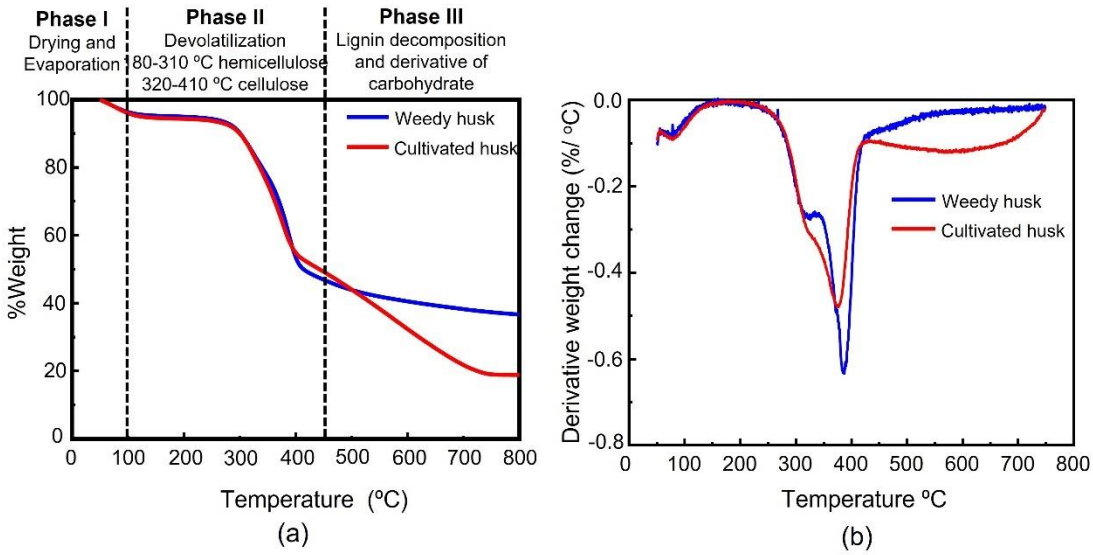


Figure 3.4 (a) TGA and (b) DTG curve of weedy (blue line) and cultivated rice (red line) with heating rate 20°C/min under nitrogen flow

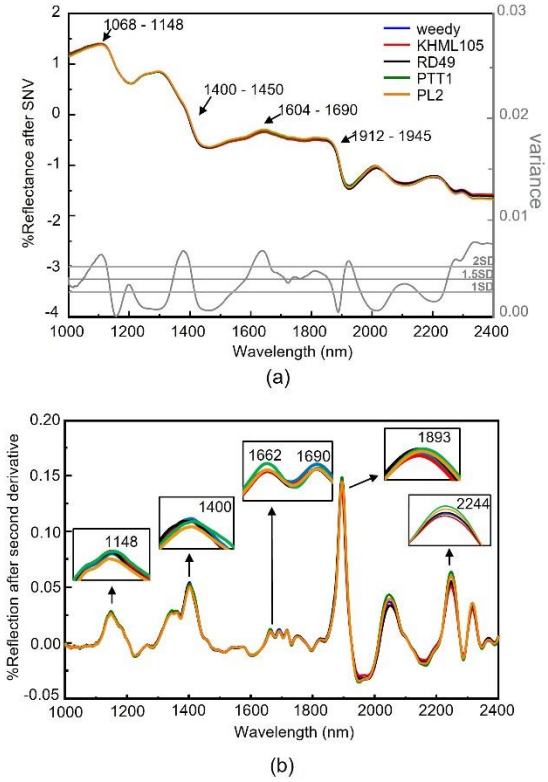
According to TGA/DTG thermograms in Figure 3.4, the weight loss takes place with three distinct steps corresponding to the water evaporation and pyrolysis stages. The first loss of weight at 80–100ºC is due to the evaporation of adsorbed water and dehydration of rice husk [107]. Both of weedy and cultivated rice paddy seed are hydrophilic (with enormous of –OH group) which easily adsorbed by water in air [108]. However, this involves only 3.5% weight loss. Typically, rice husk contains the main component including cellulose (25 to 35%), hemicellulose (18 to 21%), lignin(26 to 31%), silica (15 to 17%), soluble (2 to 5%), and moisture ca (7.5%) [109, 110]. The onset of the peaks from DTG thermograms at 332.95ºC and 301.7ºC for the weedy rice husk and the cultivated rice, respectively. These onsets of pyrolysis stage represent the starting point of decomposition of rice husk due to the degradation of hemicelluloses and cellulose. Hemicellulose composes amorphous structure with degradation thought the peak at 300-310ºC, while the sharp decrease in weight might relate to the splitting of cellulose macromolecules (at > 320ºC). [111]. At temperature above 400ºC, degradation of lignin starts, and the residue consists primarily of charcoal from lignin decomposition. Interestingly, the weedy rice husk provides the remaining 37.68% weight residue, whereas only 20.64% weight residue for cultivated rice (Figure 3a). The charcoal of lignin decomposition in the step shows significantly different. It might be assumed that lignin contents and derivative of carbohydrate

compositions in weedy rice husk and cultivated rice husk is undoubtedly different that shows high possibility to be monitored by spectroscopic technique discussed in the next section.

3.3.2 NIR spectra of rice

Figure 3.5 shows average NIR spectra of paddy seed of the cultivated rice and weedy rice samples in the wavelength regions of 1000 nm–2400 nm. It can be seen that different types of rice samples generate different NIR pattern. Predominant differences in the NIR spectra may originate from the inequality amount of chemical compositions on the rice husks. This is in good agreement with the thermal decompositions observed from the TGA thermograms. However, it is not easy to identify the overtone which distinct type of rice samples directly from the average NIR spectra. The variance of the average NIR spectra was calculated and plotted in Figure 3.5a (bottom line). Any overtone regions which provide a high variance with two time of standard deviation (2SD) indicates the possible features to discriminate type of rice samples. These characteristic reflection bands are similar to those of paddy rice seed reported by others [112]. The band of 1068 nm presents first overtone of O–H stretching mode while band of 1148 nm corresponds to second overtone of C–H stretching. The reflection bands at 1068 nm–1148 nm might be assigned to be part of either glucose [113] or lignin [114, 115]. The reflection bands at 1400 nm–1450 nm mainly represent first overtone of O–H stretching of amorphous / free OH groups / weakly hydrogen bond of polysaccharides [112, 116]. The reflection bands at 1604 nm–1690 nm are second overtone of C-H stretching of aromatic [115] and phenolic hydroxyl group [116] of lignin. Moreover, the reflection bands at 1932 nm–1945 nm are attributed to polysaccharide [117] arising from the vibration of O–H stretching, O–H bending, and O–H combination [118]. Eventually, the reflection band at 2244 nm is combination band of O–H stretching/ C–O stretching which corresponds to cellulose [119]. These assigned bands are in good agreement with the variation observed in the second derivative superimposed spectra presented in Figure 4b. According to the reflection intensity of superimposed peaks at 1148 and 1690 nm corresponding to lignin, it can be seen that the reflection intensity of weedy rice is higher than the cultivated rice. This might be

suggested that the two types of the paddy seeds majorly consist of the different amount of lignin content which is consistent with the TGA result



Wavelength (nm)	Band assignment	Structure	Ref
1068	O-H stretch 1st overtones	Glucose	López, García-González, & Franco-Robles, 2017
1148	C-H stretch 2nd overtones	Lignin	Naik, Goud, Rout, Jacobson, & Dalai, 2010
1400 -1450	O-H stretch 1st overtone	amorphous / free OH groups / weakly hydrogen bond of polysaccharides	Krongtaew, Messner, Ters, & Fackler, 2010
1604 -1690	C-H stretch of 2nd overtone	Aromatic or phenolic of lignin	Rambo, Schmidt, & Ferreira, 2015; Krongtaew, Messner, Ters, & Fackler, 2010
1932 - 1945	O-H stretching, O-H bending, and O-H combination	Poly-Saccharide	Zhang, Yin, & Ma, 2019; Abe et al., 2013
2244	O−H stretch/C−O stretch combination	Cellulose	Nakawajana, Posom, & Paeoui, 2018

Figure 3.5 NIR spectra of weedy (blue), KHML105 (red), RD49 (black), PTT1 (green) and PL2 (orange) after performing (a) standard normal variate (SNV) with the variance plot on the bottom (b) second derivative and (c) the band assignment of significant NIR regions for rice discrimination. The inset figures demonstrate the variation of 2nd derivative spectra chosen by NIR region with high variance.

3.3.3 Classification of rice by SOMs

It can be clearly seen that thermal and chemical properties of the weedy rice is strongly different from the cultivated rice as they provide the different patterns of thermogram and NIR spectra. In the section, we try to perform the statistical methods to differentiate the NIR spectra of the weedy rice from the cultivated rice. There are five different cases to be investigated that involves (I) weedy vs KHML105, (II)

weedy vs RD49, (III) weedy vs PTT1 and (IV) weedy vs PL2, respectively and (V) weedy vs combination of all types of the cultivated rice. In the study, the modified Self Organizing Maps (SOMs) was used here in order to visualize the underlying relationship and to classify group of the rice samples. Typically, supervised SOMs operates in two modes including (i) model construction and (ii) classified mapping. The map was trained by using the input samples which are training set in the case. Whereas the constructed map was automatically used to classify group of test set samples. To reveal the robustness of the generated SOM model, the dataset was divided into training set and test set for several times (100 iterations in the case) [105]. A modified algorithm of self-organizing map network architecture has been used to differentiate the weedy and the cultivated rice in the form of two-dimensional mapping. Herein, the SOM map with the size of units 20 × 30 (600 in total) was used in the study. The scaling value (*w*) was carefully optimized to avoid the overfitting problem. If value of *w* is too small, classifier model will not adequately influence the generated map to classify unknown samples. On the other hand, if large value of *w* is used, classifier model will overfit resulting in poor performance of classification especially for test set samples [59]. The satisfied scaling value for SOM classifier in each case was chosen by considering the maximum point of the classification rate from both training and test set. The %Correctly Classified (%CC) is used as a classification rate index to determine the promising value of *w* in each case. The %CC is basically calculated from the frequency of corrected prediction. A higher %CC refer to the greater model, resulting in greater accuracy and precision of unknown classification. The overall %CC of the training and test sets using the different scaling values (*w*) is shown in Figure 3.6. In all cases, the %CC of both training set and test set is monitored when scaling value was changed to build the supervised SOM model. Initially, the %CC increases when *w* is raised up until the classification model is either stabilize or slightly decrease when higher value of *w* is used. The optimal scaling value for each case is directly determined at a certain point where the rate of %CC is either flatted off or slightly decreased. From this, it can be determined that the optical scaling values equal to $w = 1$ for case I–IV, while the optimal $w = 4$ for case V, respectively.

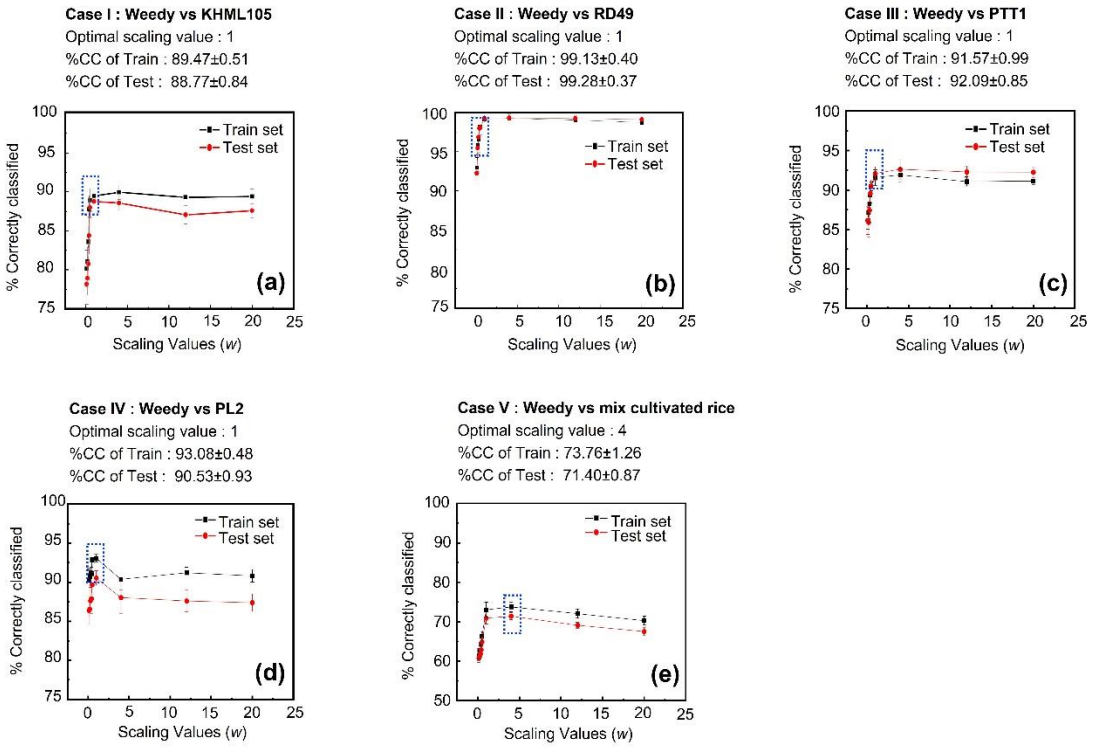


Figure 3.6 Percent Correct Classified (%CC) of the training set and test set (average from 100 iterations) with the different scaling value (*w*) used to build the supervised SOM model for (a) case I : weedy vs KHML105, (b) case II : weedy vs RD49, (c) case III : weedy vs PTT1, , (d) case IV : weedy vs PL2, (e) case V : weedy vs mix cultivated rice with the selected optimal scaling value for each case including %CC of training and test set.

For unsupervised pattern recognition, our modified SOMs are neural networks which offer some advantages over the orthogonal linear transformation method e.g., Principal Component Analysis (PCA) because SOMs could work well in either linear- or non-linear underlying dataset and it provides more options for graphical representation. In the case, the score plots of the top 3 largest principal component (PC1–PC3) of the dataset case I–V are displayed in the Figure 3.7a. The cluster separation of samples from the score plots are compared with the unsupervised and supervised SOM map as shown in Figure 3.7b and 6c, respectively. It could be seen that in PC score plots, the groups of samples are considerably overlapped, and the symbols becomes crowded and hard to distinguish, while the sample groups are

reasonably well spread out from the SOM maps. The separation of sample cluster is dramatically improved when our developed SOM are used. It can be seen that there are very distinct regions especially in case (II), case (III) and case (IV) suggesting that the characteristic patterns of NIR spectra of RD49, PTT1, and PL2 are strongly distinguished from the weedy. On the other hand, there is not such good separation for case (I) and (V) which suggests that the variability from the cultivated rice has relatively small influence on the overall map. However, this is not surprised as it is expected that most of chemical compositions are similar in the cultivated rice. This cannot easily recognize the differences when using only unsupervised method. Supervised SOMs using the optimal scaling values were performed to the same dataset for all cases. These supervised maps are shown in Figure 3.7c. It can be seen that there is dramatically improved the separation between groups of rice sample on these maps for all cases especially for case (V) to compare the types of cultivated rice. To be fair, the main application of the supervised SOM map is not for visualization of the data because it will tend to give bias interpretation according to the addition of class variables in the initial map before training. However, these map appearances are similar to the classifier used for prediction the class of unknown samples.

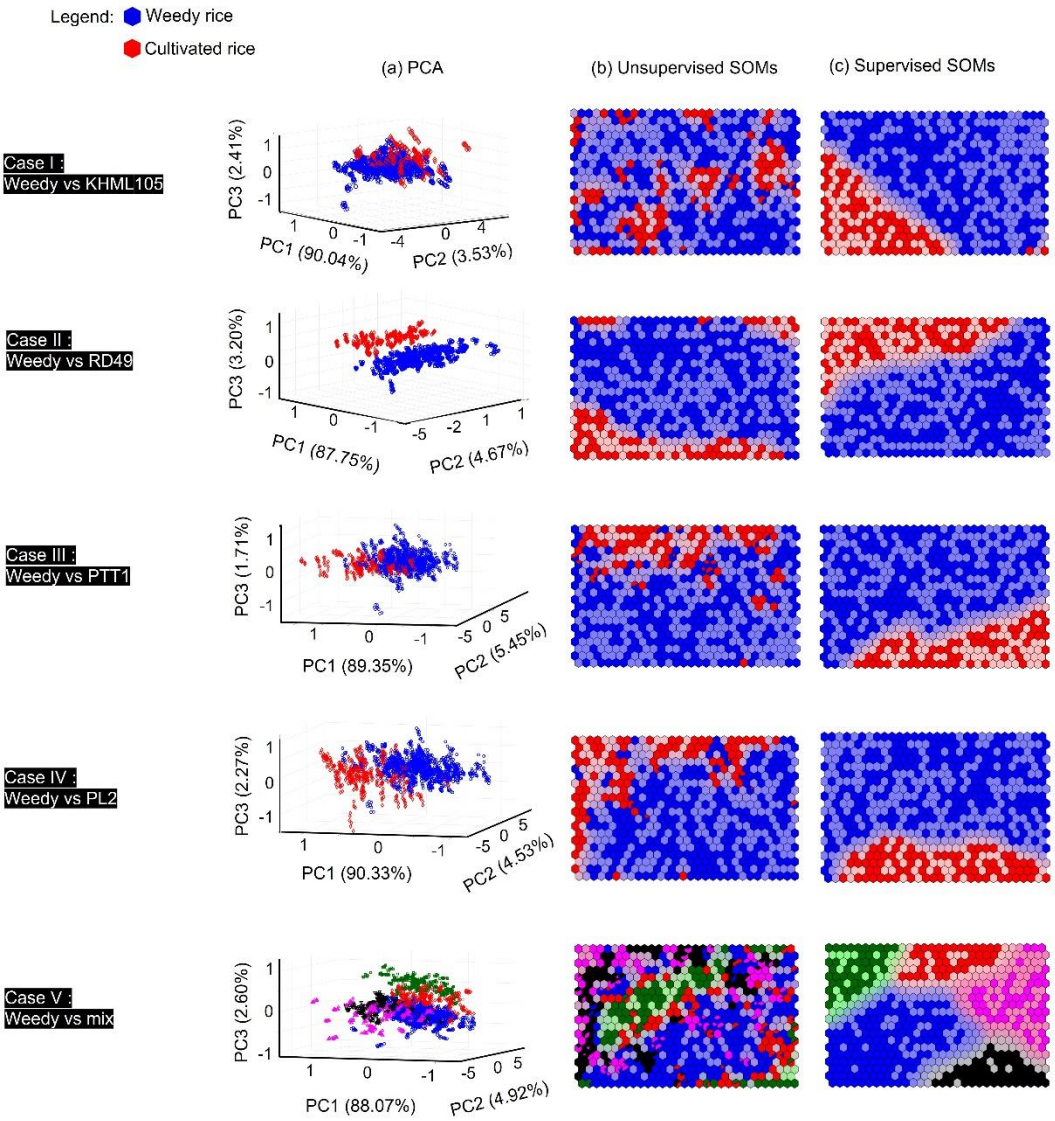


Figure 3.7 (a) PCA score plots (PC1-PC3), (b) Unsupervised SOMs and (c) Supervised SOMs of Case (I)–Case (V) using the optimal scaling values (*w*)

To evaluate the classification method, the performance indices of the developed method were calculated based on the results from the contingency table. For case I–IV, we define the positive case as the class of weedy rice, while the negative case corresponds to the cultivated rice. On the other hand, for case V, the approach of one vs all was employed when positive case refers to weedy rice and negative case represents mix cultivated rice (KHML105 + RD49 + PTT1 +PL2). From the contingency table, four indicators could be calculated where TP is the

number of correctly classified for positive case, FP is the number of negative cases that were classified as positive, TN is the number of correctly classified negative cases, and FN is the number of positive cases that were classified as negative. From these assigned indices, the classification performances including sensitivity, specificity, precision, accuracy and misclassification error can be computed as follows:

Sensitivity = TP/(TP+FN) **Specificity** = TN/(TN+FP)

Precision = TP/(TP+FP) **Accuracy** = (TP + TN) / (TP + FP + TN + FN)

Misclassification error = (FP + FN) / (TP + FP + TN + FN)

These performance indices are commonly used as metric for evaluation of the developed classifier model. The sensitivity refers to the ability of the classifier to correctly identify those samples with positive class. A high sensitivity is clearly imperative where the classifier is used to identify the correct positive class. On the other hand, specificity is inversely proportional to sensitivity where it has the ability of the classifier to correctly identify the samples with negative class. The accuracy and precision play significant roles to reveal the prediction rate where the classifier model can be very precise but inaccurate. The higher the value of those indices, the better classification is made. In case of misclassification error, it is directly related to the accuracy as the summation of accuracy and misclassification error should be equal to 1. [120, 121]. Table 3.2 summarizes the classification results which demonstrate the correctness of a model classifies the dataset in each class for case I–case V.

Table 3.2 Performance indices including sensitivity, specificity, precision, accuracy and misclassification error (ME) averaged from 100 iterations of training and test set using supervised SOM classifies with optimal scaling values

		Sensitivity	Specificity	Precision	Accuracy	M.E
Case I Positive class: Weedy Negative class: KHML105	**Train**	0.94±0.01	0.79±0.04	0.93±0.01	0.90±0.01	0.10±0.01
	Test	0.94±0.02	0.73±0.06	0.91±0.02	0.88±0.01	0.12±0.01
Case II Positive class: Weedy Negative class: RD49	**Train**	1.00±0.00	0.96±0.02	0.99±0.01	0.99±0.01	0.01±0.01
	Test	1.00±0.00	0.96±0.02	0.99±0.01	0.99±0.01	0.01±0.01
Case III Positive class: Weedy Negative class: PTT1	**Train**	0.95±0.01	0.81±0.03	0.94±0.01	0.91±0.01	0.09±0.01
	Test	0.94±0.02	0.84±0.04	0.95±0.01	0.92±0.01	0.08±0.01
Case IV Positive class: Weedy Negative class: PL2	**Train**	0.95±0.01	0.87±0.02	0.96±0.01	0.93±0.00	0.07±0.00
	Test	0.94±0.01	0.78±0.03	0.92±0.01	0.90±0.01	0.10±0.01
Case V Positive class: Weedy Negative class: mixed cultivated rice	**Train**	0.86±0.03	0.85±0.02	0.80±0.02	0.85±0.01	0.15±0.01
	Test	0.82±0.02	0.85±0.02	0.80±0.02	0.84±0.01	0.16±0.01

From Table 3.2, the standard deviation of the indices is small suggesting that several splits of training and test set with 100 iterations are required to obtain a stable estimation. The prediction in training set is an indicator of how well the model is optimized, while the prediction in test set has meaning as to show how well the classifier model can be used to predict the unknown data. In all cases, the prediction from training set and test set are relatively small different that reveals the classifier model was well optimized and not overfitting. For case I–IV, it could be seen that the sensitivity and specificity are not good balance. The sensitivity (proportion of weedy rice that is correctly identified as weedy) are approximately 5%–15% higher than the specificity. This suggests that the proposed methods could be preferably used to predict class of weedy rice rather than to predict class of cultivated rice because there are only a few false negatives occurred. Sensitivity of ≥ 0.94 is observed in cases I–IV. Even it is not good balance prediction for positive (weedy) and negative classes (cultivated rice) but the developed model is appropriated to our study for discriminate

weedy rice from the cultivated rice. The classifier model generated for case I–IV give very high value of precision (0.91–0.99) and accuracy (0.88–0.99). This reveals that the classifier model does not give bias prediction due to unequal class size. From performance indices obtained from case I–IV, it could be concluded that the developed classifier using supervised SOMs can be used to classify and discriminate the weedy rice from the cultivated rice with high precision and accuracy. For case V, we further observed the discrimination power of the model in order to classify weedy rice from the mixed cultivated rice. Interestingly, it can be seen there is a good balance between the sensitivity (0.82) and specificity (0.85) which suggests that the classifier of the model in case V is not biased toward either group. Due to the big variations from the mixed cultivated rice, the precision and accuracy for case V is reduced to 0.80 and 0.84, respectively. However, the prediction accuracy and precision are still in acceptable range (≥ 0.80). In consideration of the performance and validation using different chemometric methodologies involving Euclidean distance to centroid (EDC), Linear discriminant analysis (LDA), Quadratic discriminant analysis (QDA) and Partial least square discriminant analysis (PLSDA), they are compared with the developed SOMs using the merit performance indices as shown in Table 3.2. Here in, it can be observed that the performance of SOMs is optimal for all parameters and all cases, demonstrating the suitability of the method to discriminate weedy rice out of cultivated rice

Table 3.3 Table of merit for the discrimination of weedy rice out of cultivated rice using different chemometric methodologies involving Euclidean distance to centroids (EDC), Linear discriminant analysis (LDA), Quadratic discriminant analysis (QDA), Partial least-squares discriminant analysis (PLSDA) and our developed SOMs.

			Sensitivity	Specificity	Precision	Accuracy	M.E.
Case I : Weedy vs KHML105	SOMs	train	0.94±0.01	0.79±0.04	0.93±0.01	0.90±0.01	0.10±0.01
		test	0.94±0.02	0.73±0.06	0.91±0.02	0.88±0.01	0.12±0.01
	EDC	train	0.59±0.01	0.57±0.01	0.80±0.01	0.58±0.01	0.42±0.01
		test	0.59±0.02	0.56±0.03	0.79±0.01	0.58±0.02	0.41±0.02
	LDA (4 PCs)	train	0.83±0.01	0.84±0.01	0.92±0.01	0.83±0.01	0.17±0.01
		test	0.82±0.02	0.84±0.02	0.92±0.01	0.83±0.01	0.17±0.01
	QDA	train	0.82±0.01	0.86±0.01	0.94±0.01	0.83±0.01	0.17±0.01
		test	0.82±0.01	0.86±0.02	0.94±0.01	0.83±0.01	0.17±0.02
	PLSDA (6 PCs)	train	0.76±0.01	0.91±0.01	0.92±0.01	0.81±0.01	0.19±0.01
		test	0.75±0.02	0.90±0.02	0.92±0.01	0.80±0.02	0.20±0.02

			Sensitivity	Specificity	Precision	Accuracy	M.E.
Case II : Weedy vs RD49	SOMs	train	1.00±0.00	0.96±0.02	0.99±0.01	0.99±0.01	0.01±0.01
		test	1.00±0.00	0.96±0.02	0.99±0.01	0.99±0.01	0.01±0.01
	EDC	train	0.92±0.04	0.92±0.05	0.92±0.02	0.91±0.04	0.09±0.04
		test	0.91±0.04	0.92±0.06	0.91±0.02	0.90±0.05	0.10±0.05
	LDA (2 PCs)	train	0.89±0.01	0.82±0.01	0.93±0.01	0.87±0.01	0.13±0.01
		test	0.88±0.02	0.81±0.03	0.93±0.01	0.87±0.01	0.13±0.01
	QDA	train	0.88±0.01	0.83±0.01	0.94±0.01	0.86±0.01	0.14±0.01
		test	0.88±0.02	0.83±0.02	0.94±0.01	0.86±0.01	0.14±0.01
	PLSDA (4 PCs)	train	0.98±0.01	1.00±0.00	0.99±0.01	0.98±0.01	0.02±0.01
		test	0.97±0.01	1.00±0.00	0.99±0.01	0.98±0.01	0.02±0.01

			Sensitivity	Specificity	Precision	Accuracy	M.E.
Case III : Weedy vs PTT1	SOMs	train	0.95±0.01	0.81±0.03	0.94±0.01	0.91±0.01	0.09±0.01
		test	0.94±0.02	0.84±0.04	0.95±0.01	0.92±0.01	0.08±0.01
	EDC	train	0.68±0.03	0.64±0.02	0.85±0.01	0.67±0.03	0.33±0.03
		test	0.68±0.04	0.64±0.05	0.85±0.02	0.67±0.04	0.33±0.04
	LDA (2 PCs)	train	0.81±0.01	0.75±0.01	0.90±0.01	0.80±0.01	0.20±0.01
		test	0.81±0.02	0.75±0.03	0.90±0.01	0.79±0.01	0.21±0.01
	QDA	train	0.77±0.01	0.79±0.01	0.91±0.01	0.78±0.01	0.22±0.01
		test	0.77±0.02	0.78±0.03	0.91±0.01	0.78±0.01	0.22±0.01
	PLSDA (5 PCs)	train	0.84±0.01	0.91±0.01	0.92±0.01	0.88±0.01	0.12±0.01
		test	0.83±0.02	0.91±0.03	0.91±0.02	0.88±0.01	0.12±0.01

			Sensitivity	Specificity	Precision	Accuracy	M.E.
Case IV : Weedy vs PL2	SOMs	train	0.95±0.01	0.87±0.02	0.96±0.01	0.93±0.00	0.07±0.00
		test	0.94±0.01	0.78±0.03	0.92±0.01	0.90±0.01	0.10±0.01
	EDC	train	0.63±0.01	0.63±0.01	0.83±0.01	0.63±0.01	0.37±0.01
		test	0.63±0.02	0.62±0.04	0.83±0.01	0.63±0.02	0.37±0.02
	LDA (2 PCs)	train	0.84±0.01	0.78±0.01	0.92±0.01	0.82±0.01	0.18±0.01
		test	0.84±0.02	0.78±0.03	0.92±0.01	0.82±0.01	0.18±0.02
	QDA	train	0.81±0.02	0.81±0.01	0.93±0.01	0.81±0.01	0.19±0.01
		test	0.80±0.02	0.82±0.04	0.93±0.01	0.81±0.01	0.19±0.01
	PLSDA (5 comp)	train	0.85±0.01	0.91±0.01	0.94±0.01	0.89±0.01	0.11±0.01
		test	0.85±0.02	0.91±0.01	0.93±0.01	0.89±0.01	0.11±0.01

Next, the capability of the modified SOMs for multi-classification is revealed. The classification of case V is extended to 5 different classes involve weedy, KHML105, RD49, PTT1, and PL2. The average of percent correct classified of each class is summarized in Table 3.4. From case V, it could reveal that the supervised SOMs can be used to discriminate the weedy rice from the mixed cultivated rice directly from paddy seed with satisfactory accuracy and precision (for both dichotomous and multi-classification).

Table 3.4 Percent correctly classified over 100 iterations of training and test sets using the multi-classification on case V which involve 5 different classes (Weedy, KHML105, RD49, PTT1, PL2)

Class	Average % Correct Classified		
	Training set	**Test set**	**Random***
Weedy	**83.93 ± 2.40**	**80.22 ± 2.60**	100
KHML105	50.64 ± 7.70	47.58 ± 8.20	
RD49	83.00 ± 5.17	82.55 ± 5.93	
PTT1	57.13 ± 6.81	59.80 ± 7.23	
PL2	71.52 ± 7.67	67.14 ± 6.85	

The prediction of case V (in the main manuscript) is extended to involve 5 different classes which are weedy, KHML105, RD49, PTT1, and PL2. For unbiases interpretation, the prediction accuracy should be compared with the background prediction. The value of background prediction depends only on the number of classes by considering the classification when samples are randomly assigned to each class. In the case, there are 5 different classes, therefore, the background prediction should equal to $100 \times (1/5) = 20\%$. From the table, it is clear that the % correctly classified of the training set and test set for weedy rice is still in high predictive accuracy (83% for training set and 80% for test set) which is similar to dichotomous classification in Table 3.2 in the main manuscript (86% for training set and 82% for test set). This suggests that our modified SOMs provides the well-organized trained maps and successfully classified the weedy rice from the cultivated rice. The prediction accuracy of weedy rice can be still kept in the high level even using in the process of multi-classification.

Furthermore, the performance of our developed SOM method was compared to the discrimination and classification results of the previous research methods as shown in Table 3.5. Please note that the classification results could not be directly compared as different sample types and different techniques on data acquisition have been used in each work. However, it shows that our developed SOMs could give the

satisfied results with the classification precision in the range of 91%–99% depending on the study cases. The modified SOM algorithm has potential to be further performed on the data obtained from some alternative techniques e.g., electronic nose, hyperspectral camera.

Table 3.5 Comparison of other research methods on the quality control of rice.

Year	Sample	Data	Chemometric methods	Accuracy	Ref
2008	Paddy seeds (storage period)	Vis/NIR spectroscopy	WT, PCA, ANN	97.5%	122
2016	Weedy rice grain seed	RiSe-IViS prototype	DFA	95.8%–96.0%	123
2017	Rice mutants	Hyperspectral camera	SVM-PCA	90%–93%	124
2018	Maize seed varieties	NIR (reflectance)	PCA, LDA	90%	125
2019	Group of weedy rice	RiSe-IViS prototype	DFA	98%	126
2019	Organic rice	NIR	PLS, PCA	87.5%	127
2019	Grade of rice and geographical origin	Hand-held NIR	MSC-PCA-KNN	90.6%–91.8%	14
-	Weedy rice (paddy seed)	NIR (reflectance)	SOM	91%–99%	This work

**Note:Wavelet transform (WT), Principal component analysis (PCA), Artificial neural networks (ANN), Linear discriminant analysis (LDA), Discriminant function analysis (DFA), Support vector machine (SVM), Partial Least Squares regression (PLS), Multiplicative scatter correction (MSC), k-Nearest Neighbor (kNN), Self-Organizing Map (SOM)*

3.4 CONCLUSIONS

In our study, we developed the unsupervised and supervised SOMs as a classification method to potentially discriminate the weedy rice from the cultivated rice directly from paddy seed. Their physical properties and thermal behaviors of the weedy and cultivated rice paddy samples were investigated. The results displayed that there is no significantly different in their physical appearances due to the similarity of their morphological features of rice husk. The thermal behaviors and thermal decomposition of the rice samples at temperature above 400ºC revealed the different amounts of lignin contents and derivative of carbohydrate between the weedy rice husk and cultivated rice husk. According to NIR measurement, there were five important overtone regions selected using the variance including 1148, 1400 nm, 1690, 1893 and 2244 nm. The tendency of the reflection NIR spectra of the rice samples is consistent with the TGA results. For the classification part, supervised SOMs were used to discriminate the weedy rice from the mix cultivated directly from NIR spectra of their paddy seeds. The optimal scaling value (*w)* of the develop SOM model is well optimized to prevent the overfitting problem. The performance of SOM classifier was validated with 100 different training and test sets to obtain the robust prediction. In order to evaluate the developed classification model, the performance indices including sensitivity, specificity, precision, accuracy and misclassification error were used to access the classification performance. The classifier model gives very high value of precision (0.91–0.99) and accuracy (0.88–0.99) for the data contain weedy and a cultivated rice, where they slightly reduced to 0.80 and 0.84 for precision and accuracy, respectively, for the data of weedy against the mixed cultivated rice. This suggests that the supervised SOMs can be used to discriminate the weedy rice from the mixed cultivated rice. In the future, near infrared spectroscopy combined with supervised SOMs might become a powerful invasive, green and simple techniques which could be performed fast and accurate without using extra chemicals and process required to assess and inspect of rice seed quality.

CHAPTER IV
PROJECTED PIXELS ON HYPERSPECTRAL NIR IMAGE BY SUPERVISED SELF-ORGANIZING MAP TO CLASSIFY WEEDY RICE SEED

In general, the NIR spectrum was extracted from hyperspectral image (HSI) with two approaches which are pixel-wise and mean spectrum [128]. The pixel-wise spectral analysis uses the spectra of each pixel in region of interest. Even it contains more details information but could give misleading results due to differences within the sample. Therefore, the mean spectrum of all pixel-wise spectra is preferably calculated as the representative data for the sample. The non-uniformity of pixel space due to various factors (e.g., lens distortion, sensor movement, rugged terrains) in the image have not been concerned by the methods. After obtaining the spectral data, there are large number of linear and non-linear classification methods such as linear discriminant analysis (LDA) [46], partial least squares discriminant analysis (PLS-DA) [47], the *k*-nearest neighbors (*k*NN) [48], support vector machine (SVM) [49], principal component analysis (PCA) [50], and artificial neural networks (ANNs) [38] which can be utilized to quantify and visualize the chemical variation within the heterogeneous samples, such as plant seeds. Moreover, most techniques use reduction methods such as principal component analysis (PCA), which require re-calculating the number of latent variables (LVs) every time a new set of samples is added to maintain prediction viability, resulting in a critical limitation when the extra set of samples have been added. In the study, the new classification approach for the HSI image is proposed by using Self-Organizing Maps [58-60]. The developed supervised SOMs was applied on the pair-wise HSI to generate the supervised global SOM map which visualize the unit of each class. All parameters involving scaling value and map size were systematically optimized. The pair-wise pixels of an unknown sample were projected to the global SOM map in order to determine the class of each pixel. The class of each pixel will be then projected to the image by simple display using color scale (e.g., Red, Green and Blue). Then the class of each sample image is determined by the ratio of the projected color on the image. This approach is more appropriate to

real implementations of using NIR hyperspectral imaging systems for seed quality as it classifies based on individual seed image using all pair-wise HSI instead of just using only the mean-spectrum. Moreover, the global SOM map could be updated anytime when there is more sample information available without the requirement of re-optimized parameters. The proposed classification approach using SOMs compared with the classical approach on HSI analysis was illustrated in Table 4.1.

Table 4. 1 A survey on current publication

Year	Sample	Extracted Features	Sensing modality	Classifiers	Accuracy	Ref
2005	Identification of rice seed varieties	Color, Morphology	RGB	Neural Network	84.33%	129
2013	Identification of rice seed cultivar	Spectral	HSI	PLS-DA, K-NN, SIMCA, SVM, and RF models	80-100	29
2016	Classification of four varieties of bulk rice grain	Color, Texture, Wavelet	RGB, HSI	BPNN	96-100%	130
2017	Identifying Paddy Seed Varieties	Shape-base	RGB	BPNN	95.53%	131
2017	Identification of rice origin from four different regions	Spectral, morphological and texture features	HSI	SVM	91.67%	132

Table 4. 1 A survey on current publication (continued)

Year	Sample	Extracted Features	Sensing modality	Classifiers	Accuracy	Ref
2018	Inspecting rice seed species purity	Morphological, color, and textural traits	RGB	DT, RF, Adaboost	≥80%	133
2019	Determination of rice seed vitality of different years	Spectral	HSI	PCA	93.67%	43
2020	Detection of rice kernels infected with rice false smut	Spectral	HSI	PCA	99.27%	134
2021	Prediction of Anthocyanins Content in Black Rice Seeds	Spectral	HSI	PLSR	85-95%	135

**Note: Partial least squares-discriminant analysis (PLS-DA), K-Nearest Neighbors (K-NN), Soft independent modelling by class analogy (SIMCA), Support Vector Machine (SVM), Random Forest (RF), Back-Propagation Neural Network (BPNN), Decision Tree (DT), Principal component analysis (PCA), Partial least squares regression (PLSR)*

To the best of our knowledge, no reports on the approach using supervised SOMs are available. The approach was used to discriminate the weedy rice from cultivated rice directly from paddy seed for seed quality assessment as the case study. This concept method of this article is the first attempt in NIR hyperspectral imaging applications in seed quality monitoring by using actual HSI data for the analysis.

4.1 Experimental setup

4.1.1 Sample collection and preparation

Weedy was collected from the local fields at Phrom Phiram district, Phitsanulok province, Thailand. The two types of standard cultivated rice seeds, including Phitsanulok2 (PL2) and Kor Khor 49 (RD49), were collected from the Lifestyle and Spirit of Thai Farmers-Nahai Chai Learning Center Suphanburi province Thailand. They are certificated by the Rice Department Ministry of Agriculture and Cooperatives, Thailand. The collected seed samples were pre-treated in order to remove the contaminated particles and other impurities attached to the rice husk by using a cyclone vacuum machine. After that, the seed samples were safely kept in the vacuum boxes at room temperature before acquiring the NIR measurements.

4.1.2 NIR-Hyperspectral acquisition

Before the data collection, the power supply was turned on to warm up the HSI system for 30 minutes to eliminate the baseline drift and other errors caused by the system. Then, hyperspectral images of the rice samples were acquired by using one push broom HSI as shown in Figure 4.1. It comprises an imaging spectrograph (Imspector N17E; Specim, Spectral Imaging Ltd., Oulu, Finland), a CCD camera (Xeva 992; Xenics Infrared Solutions, Belgium), two 500 W tungsten–halogen light sources (Lowel Light Inc., NY, USA), and control software (Specim's LUMO Software Suite; Spectral Imaging Ltd., Oulu, Finland). The HSI system was constructed to cover near-infrared (NIR) wavelengths for reflectance measurements. The information of the system in detail was described by Kim et al. [136]

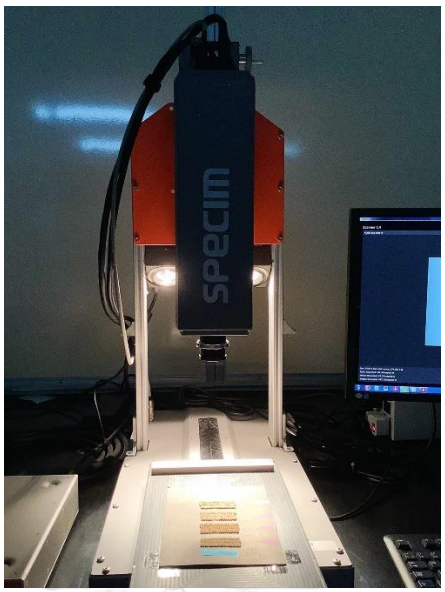


Figure 4.1 Sample presentation for NIR hyperspectral imaging in wavelength region: 900–1700 nm

Image acquisition was carried out at room temperature. In order to facilitate the segmentation of rice seeds from the background, rice seeds were plated on a black plate with very low reflectivity. At each time, fifty seeds, including twenty weedy seeds (40%w/w of the total) and thirty seeds cultivated rice, were randomly plated without overlapping each other in five rows, and each row was divided into ten seeds (Figure 4.2). All images were collected by obtaining spectral/spatial data line-by-line as the translation table moved the sample plate under the instantaneous field of view (IFOV) of the HSI system. In order to obtain clear images without deformation, the height between the camera lens and the samples was set at 30 cm, and the camera's exposure time was set at 9 ms. The system scanned the samples line by line along the Y-axis, which used the two-dimensional image sensor in a spectral range of 1000–2350 nm, and the samples were moved along the X-axis at a constant speed of 10 mm s^{-1}. The size of the hyperspectral image determined by the camera was 267 × 320 with 256 active bands.

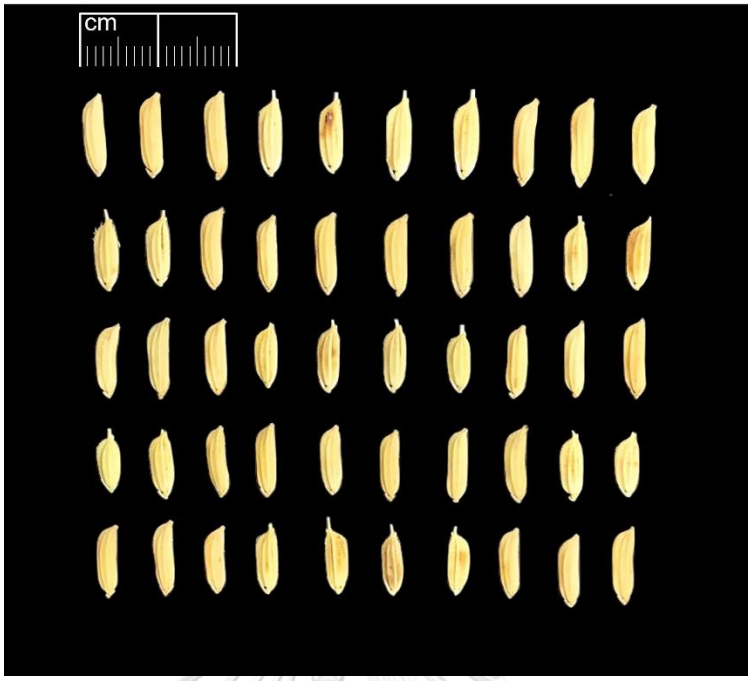


Figure 4.2 Diagonal rice arrangement on seed plate.

The raw hyperspectral images of the samples were corrected using two reference standards, including the white reference image and the dark reference image. They are obtained under the same condition as sample image acquisition. The white reference image was obtained using a white Teflon bar of nearly 100% reflectance, and the dark reference image was acquired by turning the light source off and completely covering the lens with its opaque cap. Then the corrected image was calculated by the following equation

$$I = \frac{I_0 - I_d}{I_w - I_d}$$

Where I is the corrected image, I_0 is the raw image, I_d is the dark reference image, and I_w is the white reference image

It's well known that hyperspectral imaging (HSI) system integrates digital imaging and spectroscopic techniques into one system [40]. The information recorded in HSI represents three-dimensional data which contains the spatial information of the image and the spectral data and is called a "hypercube." as shown in Figure 4.2. Hyperspectral image of each sample with dimensions of x × y × n where x and y are

the spatial dimensions and n is the number of wavelengths was captured. In this study, the resolutions ratio of hyperspectral image is 256 × 320 × 256, which having active band on the line in spatial range of 1000–2350 nm. Furthermore, spectral information (X-matrix) of the imaged sample that represents its physicochemical properties could be extracted either directly from the segmented objects in the image as the main region of interest [39]. It should be noted the segments usually contain pixels from the shadow's boundary regions rather than the seeds' pure spectra when the seeds are segmented only using HSI data. As a consequence, any morphological feature measurements relying only on spectral image segmentation may be imprecise, and spectra from non-rice-seed pixels will be included in the assessment [13].

Herein, the spectra arising from one rice seed are obtained in the area of approximately 30 × 10 pixel. The effective spectra for global model construction were picked up from relatively small areas of 5 pixels × 3 pixels square almost center of the rice (as shown in Figure 4.3) to ensure that the spatial and spectral features are appropriately included in the data analysis.

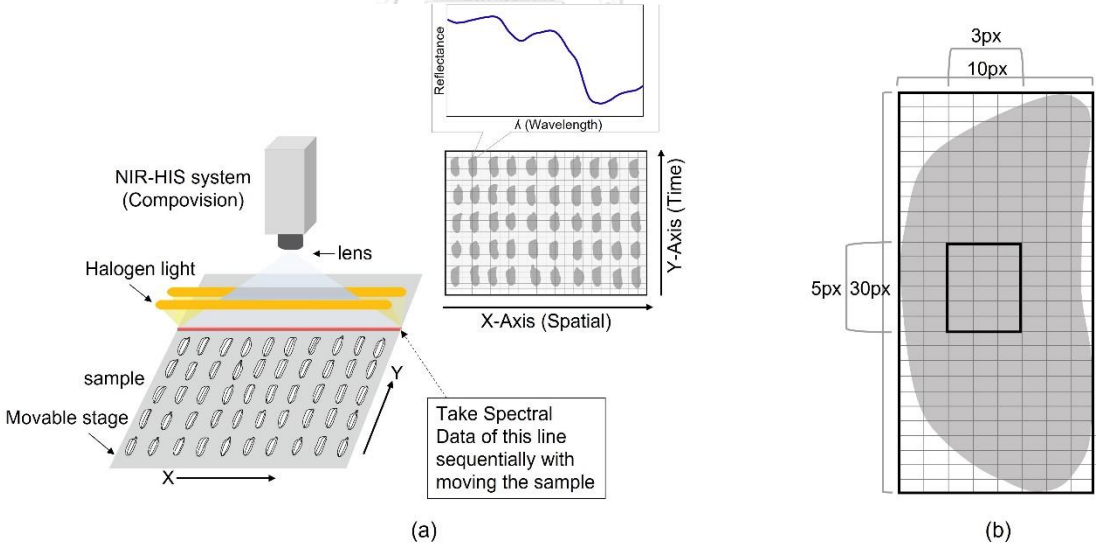


Figure 4.3 (a) Components of a hyperspectral imaging system (b) Image of the pixel number in a rice seed

4.1.3 Scanning electron microscope (SEM)

The morphology of the rice seed samples was investigated by scanning electron microscopy (SEM) technique. Samples were fixed on a carbon tape and attached on an aluminum stub. The SEM sample was vacuum dried for 1 h before imaging. SEM micrographs of samples were performed using a scanning electron microscope (SEM, JEOL JSM-6510) operated at 2–15 kV under high vacuum mode.

4.1.4 Direct analysis in real time mass spectrometry (DART-MS)

The rice samples were analyzed by DART (SVP 100; IonSense, Inc., Saugus, MA, USA) (JEOL USA. Inc.) coupled with AccuTOF LC-plus (JMS-T100LP) mass spectrometer (JEOL Ltd., Tokyo, Japan). DART was operated in positive ion mode using helium (Ultra high purity: 99.999%, and 300 °C) as the discharge gas with a flow rate of 2.25 L/min. The mass spectra were recorded with m/z from 100 to 1000 and processed using MS Tune Manager Software Version 1.3.0.0.

4.2 Data analysis

The global map was created from 3 sample maps which randomly contain weedy and cultivated rice seeds (total 50 seeds) for each map. Therefore, a total of 150 sample seeds were obtained. After that, the random pixel from each seed corresponding to the relatively small areas mentioned above was selected. All reflectance constituents were from the effective wavelength ranged from 990 to 1640 nm with 3 nm resolutions. The reflectance data of all samples were then arranged in a matrix (1800 samples × 209 wavelengths). The hyperspectral data of all sample seeds were further analyzed using a supervised self-organizing map. The main procedure for analyzing hyperspectral imaging data is depicted in Figure 4.4. They were divided into 4 parts: (a) data acquisition, (b) sample segmentation, (c) global SOM map, and (d) determination of unknow seeds. The explanation in detail was shown in the following section.

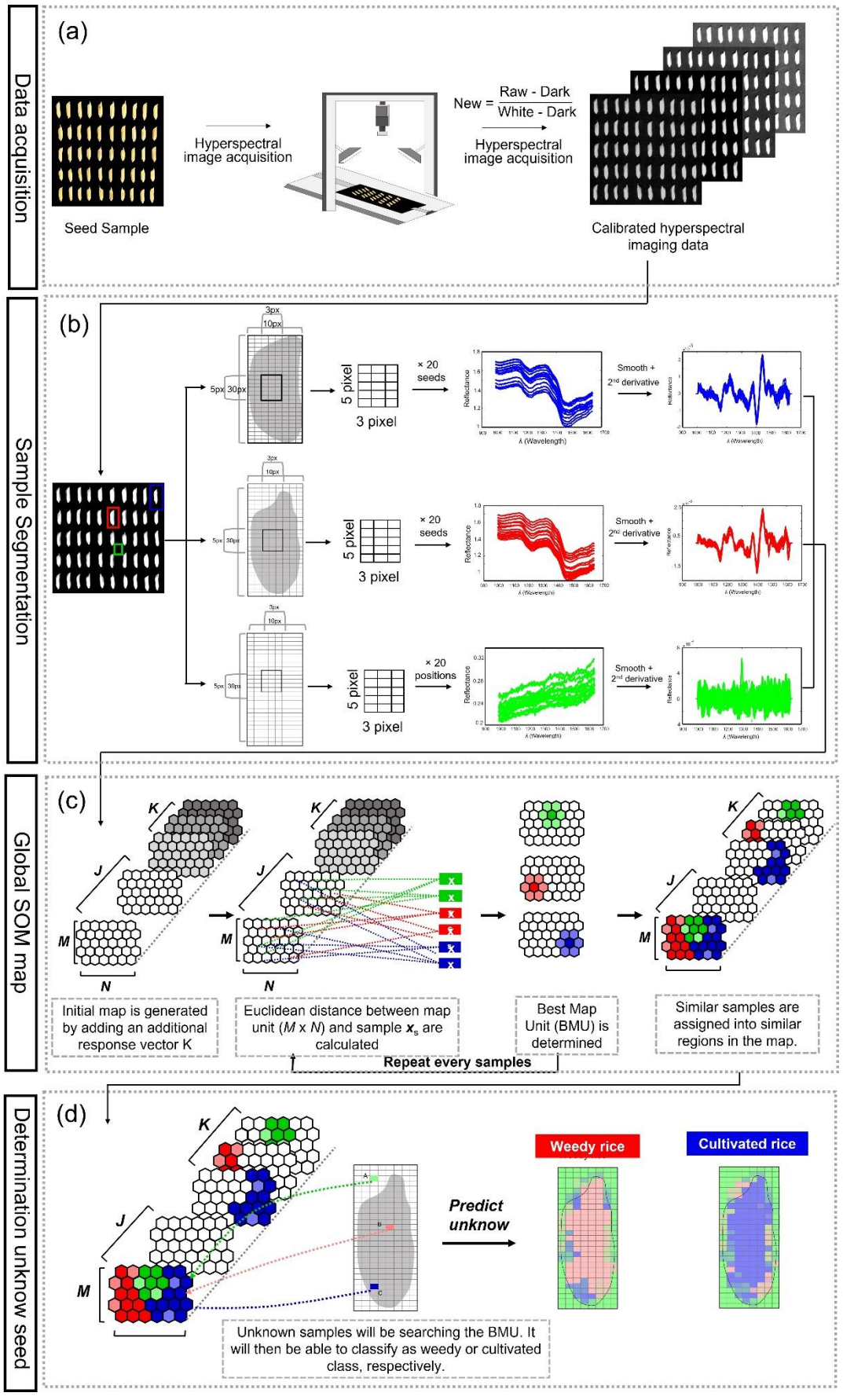

Data acquisition
(a)
Seed Sample
Hyperspectral image acquisition
New = Raw - Dark / White - Dark
Hyperspectral image acquisition
Calibrated hyperspectral imaging data
Sample Segmentation
(b)
3px
10px
5px 30px
5 pixel
3 pixel
× 20 seeds
× 20 positions
Smooth + 2nd derivative
Reflectance
λ (Wavelength)
Global SOM map
(c)
K
J
M
N
Initial map is generated by adding an additional response vector K
Euclidean distance between map unit (M x N) and sample x_s are calculated
Best Map Unit (BMU) is determined
Similar samples are assigned into similar regions in the map.
Repeat every samples
Determination unknow seed
(d)
Predict unknow
Weedy rice
Cultivated rice
Unknown samples will be searching the BMU. It will then be able to classify as weedy or cultivated class, respectively.

Figure 4.4 Schematic diagram for sample visualization and classification of weedy rice and cultivated rice from HSI spectra using supervised SOM for *K* classes and *J* variables with the SOM map in the dimension of *M* × *N*. There are 4 sub-steps in total: (a) data acquisition, (b) sample segmentation, (c) global SOM map, and (d) determination of unknow seeds.

4.2.1 Preprocessing method

The data collected by the hyperspectral system is necessary to perform a series of processing on the original images to finally perform the suitably spectral data. The first step begins with the interquartile range (IQR), commonly used to detect outliers. It measures statistical dispersion, being equal to the difference between the 75th percentile and the 25th percentile. The average NIR-HSI spectrum of each sample class was calculated as a centroid of the class. Euclidean distance of the NIR spectrum of in-class samples was then computed. Samples that provide Euclidean distance outside the 1.5 × interquartile range (IQR) from the average in-class NIR spectrum are identified as outliers and subsequently eliminated [137]. After that, the obtained spectra will be performed spectral pretreatment using Savitzky-Golay smoothing (SGS) coupled with standard normal variate (SNV) in order to effectively remove the signal variation from light scattering in the heterogenous samples [103]. The other preprocessing is Savitzky-Golay smoothing coupled with 2nd derivatives (2D), which mainly used to resolve peak overlap (or enhance resolution) and eliminate constant and linear baseline drift between samples, resulting in improve spectral resolution, identify overlapping spectral peaks, and highlight spectral peaks with significant differences [138, 139].

4.2.2 Development of self-organizing map

A self-organizing map (SOMs) is one of the most well-known artificial neural networks (ANNs) [98, 104]. It can be constructed without assuming any mathematical functions. In other words, it is a non-linear method. The main principal of SOMs is its ability to not only transform multi-dimensional data into visually decipherable clusters in a low-dimensional grids (2D grids) form, but also maintain relative

distances between existing data units in a multidimensional space form [99, 140]. SOMs basically involves two processes, including vector quantization and vector projection. In this work, supervised SOM model was applied using the algorithm presented in detail in the literatures [99, 141]. Therefore, only essential steps are described here. The first step is initialization. A trained map consisting of a grid as represented by the two-dimensional hexagonal of units is generated in this step. The shape of the units is not specific, although squares and hexagons are particularly favorable because they have neighbors that have the same distance apart in numerous directions [142]. Each map unit consists of the weight vector randomly generated from a uniform distribution between the maximum and minimum intensities in the dataset. In particular, the number of map layers is set to be equal number of variables $\boldsymbol{J}$ (wavelengths in the case) in the dataset. A map can be generated by $\boldsymbol{M} \times \boldsymbol{N}$ units, where $\boldsymbol{M}$ and $\boldsymbol{N}$ are the numbers of rows and columns of the map, respectively. It was normally set to approximately 2.5 times compared with number of samples in the dataset [101]. The next step is the training process. During each training step, a sample vector $\boldsymbol{x}_s$ from data matrix $\boldsymbol{X}$, which is randomly chosen from the training sample and newly generated for each iteration, is compared to each of the map unit weight vectors. The dissimilarity between the sample vector ($\boldsymbol{x}_s$) and weight vector ($\boldsymbol{w}_k$) on each map unit, namely Euclidean distance, is calculated by

$$s(\boldsymbol{x}_s, \boldsymbol{w}_k) = \sqrt{\sum_{j=1}^{J} (\boldsymbol{x}_s - \boldsymbol{w}_{kj})^2}$$

Herein, the map with the most similar weight (having the lowest dissimilarity) vector is declared the ‘winner’ or the best matching unit (BMU). The BMU becomes the center of learning for that iteration, and its neighboring units are updated to become more similar to the sample vector. The entire process is repeated for 10,000 iterations until the map regions are stable. As the learning proceeds, the samples containing similar underlying information are gradually moved towards a map region and mapped onto the SOM units that are close together in the map space. Samples originated from the same groups are assigned into analogous regions on the SOM map, while samples from different groups are laid on the other regions. For

visualization, it is difficult to directly visualize the updated SOM map, therefore, the color map has been created to reveal the clusters of samples. The shading of the color map units is updated in each iteration of the which directly related to the updated SOM map. The color map will help interpretation, and so it is possible to watch in real-time as the training progresses.

The strategy of SOMs can be adapted for supervised learning with an additional set of variables representing the class membership appended to the input variables for training. In supervised SOMs, the class weight vector (***K***), including class membership information, is added to the initial map. The dimension of the class weight vector depends on the number of classes in the data. For example, if the sample contains three classes involving A, B and C, then the class weight vector will be assigned as [ω 0 0], [0 ω 0] and [0 0 ω], respectively, where ω is used to indicating that the sample is in that class and 0 if not. The class weight vector will also be trained during the iteration similar to the color and SOMs maps. The ability to separate between different groups of samples in supervised method is strongly influenced by the Optimal Scaling Value (ω). If the value ω is too high, it may overfit the data. However, if it is too low, it could render the map unsupervised [59], resulting in moderately performing or even wrong predictive models [143]. Therefore, the Optimal Scaling Value (ω) is the critical parameter needed to be optimized in the supervised method. The other parameter which is strongly affected the performance of the classifier is the size of the map. Different sizes of maps trained at a specific number of iterations will have different resolutions. If the map size is too small, it might not explain the essential differences between samples that should be detected. Conversely, if the map size is too big, the differences are too small to be observed [144]. Therefore, in order to ensure that our global map is completely perfect, map size selection is needed. The map size is usually predefined in SOMs; an appropriate map size can only be decided after training the samples on different sizes of map [143].

The details of supervised SOM algorithm, including the BMU, adjusted learning rate, neighborhood widths, optimized optimal scaling value, etc. during the training process was already explained in our previous study elsewhere [59, 101].

After optimal value ω and size map are defined, they were used to create the global map for further study in the next part. All calculation steps, image processing,

and the spectral value of all pixels were performed on MATLAB R2018b with an in-house coding algorithm.

However, there is one major disadvantage of SOMs, it requires necessary and sufficient data to develop meaningful clusters. The weight vectors must be based on data that can successfully group and distinguish inputs [145]. Lack of data or extraneous data in the weight vectors will add randomness to the groupings, resulting in the limitation of SOMs features to classify correctly. Therefore, in the present study, the developed classification model (supervised SOMs) established following the step above as a global map was used to apply to other hyperspectral images to form classification maps, thereby allowing the rice seeds to be simply classified based on the intensity of the pixels.

Similar to the human eye, this kind of pixel was represented by traditional color imaging, known as RGB imaging. It used three broadband color channels (Red, Blue, green) to produce a signal color value for each pixel in the image. Herein, Red, Blue, and Green were referred to as cultivated rice, weedy, and background, respectively. The essential underlying information from a global map was used to predict samples in other maps using Receiver Operating Characteristic (ROC) as an index to distinguish sample class. The prediction result was represented in an RGB image pixel as shown in Figure 4.4d.

4.3 Results and discussions

4.3.1 Rice seed characteristics

In order to visualize the features of rice seed, each type of rice seed including cultivated rice and weedy rice was photographed by digital microscope camera as shown in Figure 6. Because weedy rice may instead originate from cultivated rice through de-domestication with adaptive mutations, so their external appearance is similar to cultivated rice varieties in term of shape and color [2]. According to this result, it causes difficulty in order to discriminate weedy rice from cultivated varieties directly from paddy seed by human visualization, as shown in Figures 4.5 (a) and 4.5(c).

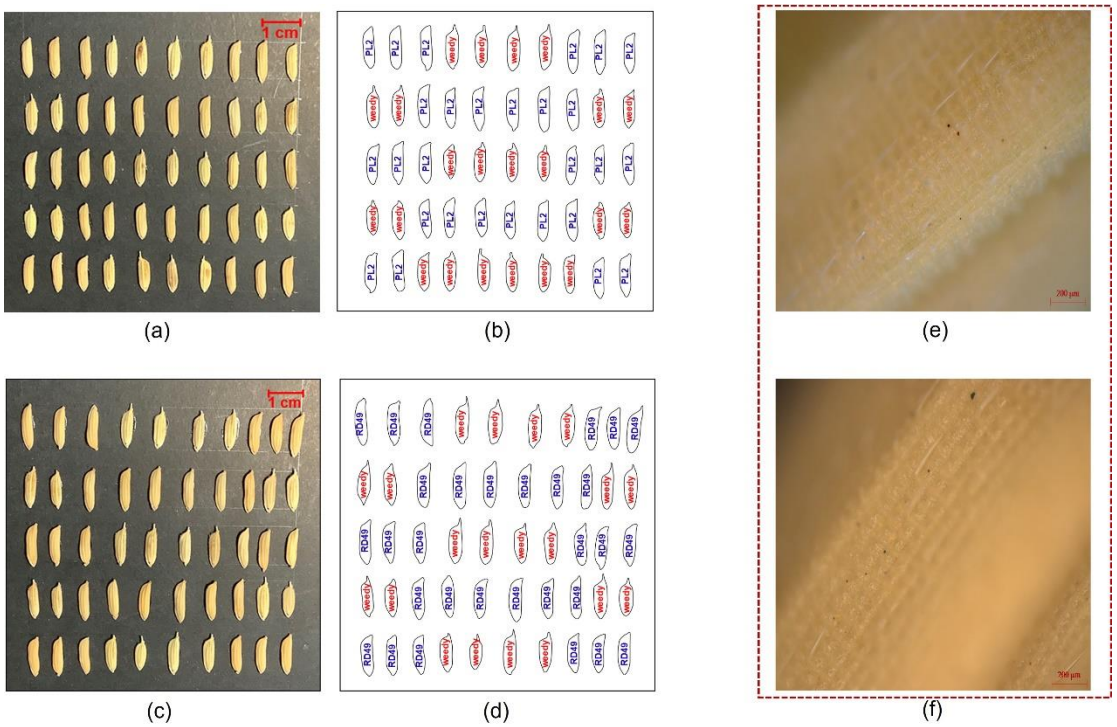
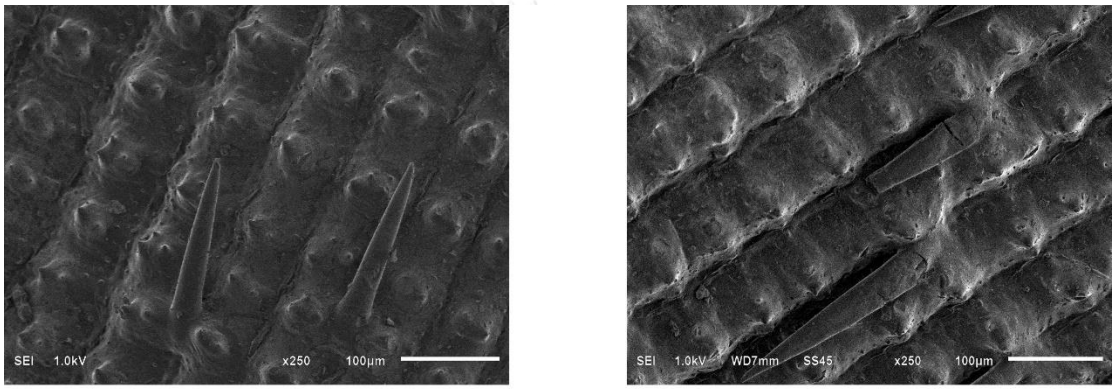


Figure 4.5 Morphological features of rice seeds. The samples are presented on the acquisition stage, belonging to the black paper stage (background). Case I : PL2 and weedy (a) 3D digital image (b) 2D sample image with label visualization. Case II : RD49 and weedy (c) 3D digital image (b) 2D sample image with label visualization. The magnified optical images. On the right-hand side showed the optical microscope images (100×) of the rice (e) without cyclone (f) with cyclone.

Herein, a developed cyclone was used to remove contaminated particles on the rice husk surface. The morphology of rice husk with the magnification of 100× captured by optical microscope of paddy rice seed before and after incubating in the cyclone was investigated as shown in Figure 4.5(e)–4.5(f), respectively. Observation of these samples revealed their native morphology have no significant differences in physical characteristics between with/without cyclone. In other words, it may be suggested that a cyclone vacuum machine is suitable for removing external contaminants on the rice husk and being able to maintain their chemical properties, which render to acquired spectra come from their intrinsic rice.

4.3.2 Scanning electron microscope (SEM) analysis

To better understand the morphology of rice samples, weedy and cultivated rice seeds were subjected to SEM. The surface topographical features are presented in Figure 4.6, which revealed that the husk surface of both types of rice is equipped with a smooth inner surface and a systematically undulating outer surface. Although the morphology of the weedy samples (Figure 4.6a) was more spike observed than the cultivated rice (Figure 4.6b), there was no significant difference in native morphology between these two of rice under SEM.



(a) (b)

Figure 4.6 SEM images of native rice seed (a) Weedy seed (b) Cultivated rice seed

4.3.3 Reflectance spectral characteristic

The average NIR-HSI reflectance spectral data based on the cultivated rice seed (PL2) and weedy sample obtained from the hyperspectral images shown in Figure 4.7. The wavelength range of 1000 nm–1600 nm was chosen as the research focus. They showed a similar pattern which was characterized by several broad peaks in the same region. It is difficult to identify the overtone which is a distinct type of rice samples directly from the average NIR spectra. In order to overcome this problem, the variance of the average NIR spectra was calculated and plotted in Figure 4.7d. Any overtone regions which provide a high variance with two times of standard deviation (2SD) indicates the possible labels to discriminate type of rice samples. The band range from 1000–1200 nm was assigned to C–H second overtone, which may come from aromatic or aliphatic compounds [135]. The band of 1068 nm presents first

overtone of O–H stretching mode, while the band of 1148 nm corresponds to second overtone of C–H stretching. The reflection bands at 1068 nm–1148 nm might be assigned to be part of either glucose [113] or lignin [114, 115]. The reflection bands at 1400 nm–1450 nm mainly represent first overtone of O–H stretching of amorphous / free O–H groups / weakly hydrogen bond of polysaccharides [116]. The next area (1604 nm–1690 nm) is assigned to second overtone of C–H stretching of aromatic [115] and phenolic hydroxyl group [116] of lignin. This similar characteristic also shows in case II (RD49 and weedy) in Figure 4.8.

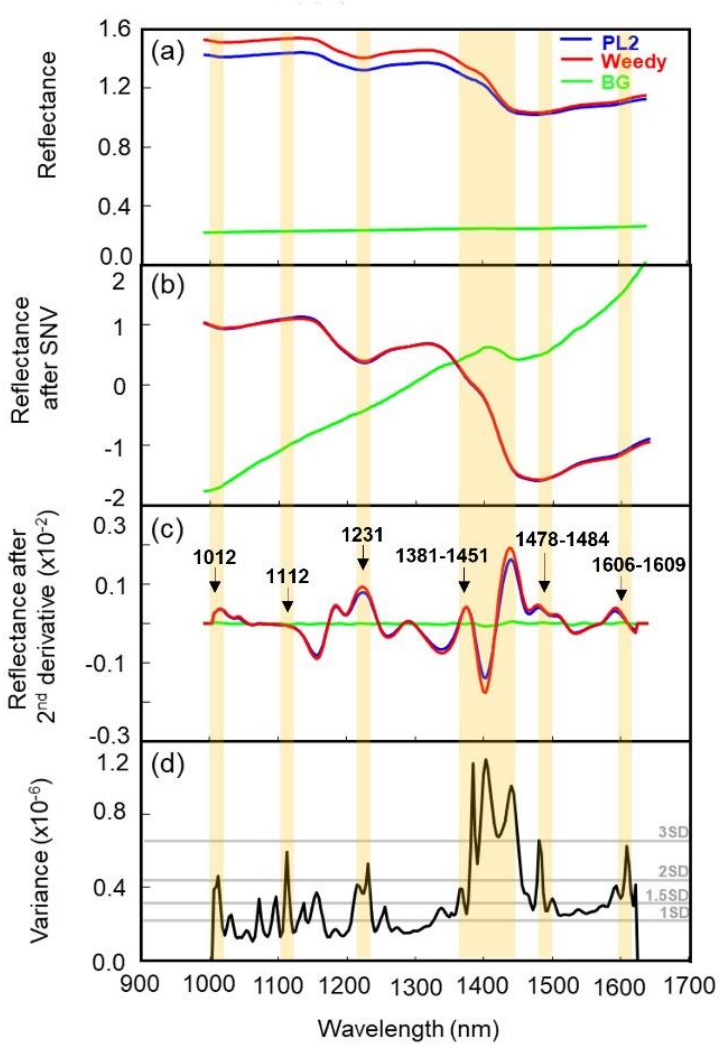


Figure 4.7 Characteristic reflectance spectra (a) raw data and after performing different pretreated method (b) S–G smoothing (C) SNV (D) 2nd Derivative of PL2 and weedy

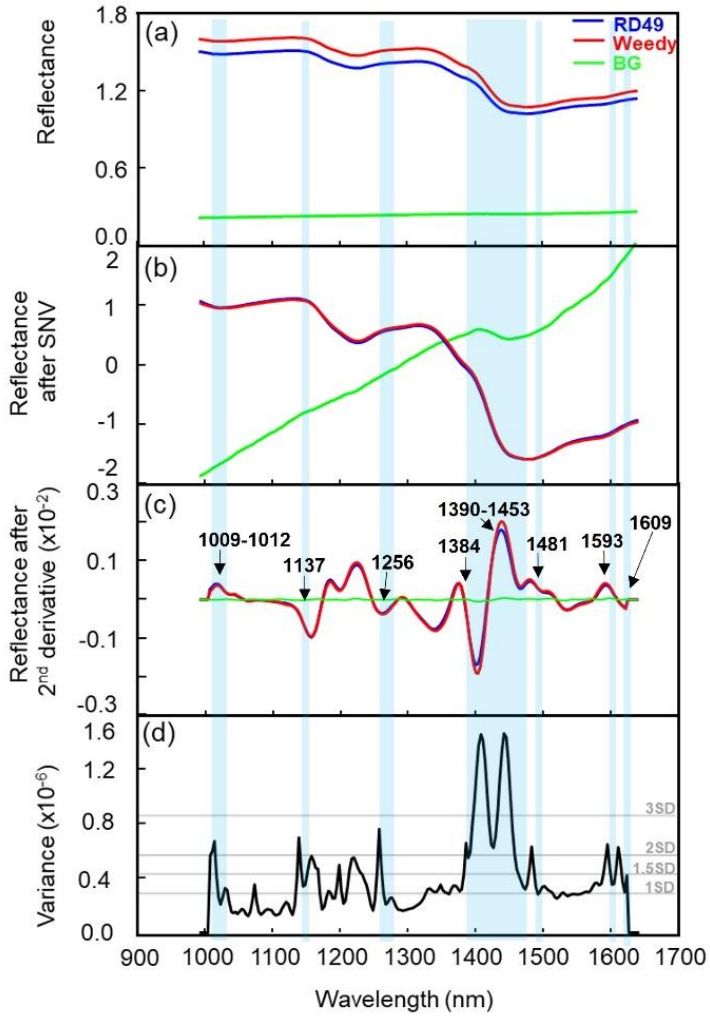


Figure 4.8 Characteristic reflectance spectra (a) raw data and after performing different pretreated method (b) S–G smoothing (C) SNV (D) 2nd Derivative of RD49 and weedy

The spectra were processed with different methods, including Savitzky-Golay smoothing (SGS) coupled with standard normal variate (SNV) and Savitzky-Golay smoothing coupled with 2nd derivatives (2D). Compared with the raw spectra, after preprocessing revealed that relative baseline translation between spectra were corrected and scattering effect were eliminated, resulting to improve spectral resolution, identify overlapping spectral peaks and enhance the useful spectral absorption information [146].

4.3.4 Classification of rice by SOMs

Self-Organizing Maps (SOMs) is an unsupervised learning method. The principal goal of an SOMs is to transform an underlying signal pattern of arbitrary dimension into two-dimensional grid of connected neurons which are multi-dimensional vectors. [143]. In other words, SOM provides effective results which are easily visualized and interpreted from the generated component planes (CPs). Maps samples in the same unit will show more similarities and be represented closer on the map. In contrast, samples with different patterns are located far away from each other [147]. The strategy of SOMs can be adapted for supervised learning with an additional set of variables representing the class membership appended to the input variables for training [59]. In the present study, supervised SOMs were used here to visualize the underlying relationship and classify the group of rice samples. Typically, the learning process in SOMs involves two main steps: selecting the best matching unit and self-organization of the map. The map was trained by using the input samples, which are training set in the case. Whereas the constructed map was automatically used to classify group of test set samples. Supervised SOMs are frequently used to classify an unknown sample into a group by using the trained map as a classifier.

The most straightforward measurement to determine the performance of the classifier is Percent Correctly Classified (%CC) [105]. For the overall %CC, the dataset was divided into training and test sets several times (100 iterations in the case). The %CC results are dependent on the chosen scaling value (ω). Therefore, it is essential to carefully optimize scaling value (ω) to perform a good prediction and avoid the overfitting problem. If the value is too high, this may overfit the data. However, if it is too low, it could render the map unsupervised, resulting in moderately performing or even wrong predictive models [143]. The overall %CC of the training and test sets using the different scaling values (ω) is shown in Figure 4.9. In all cases, the %CC of both training set and test set is monitored when scaling value was changed to build the supervised SOMs model. The result shows that %CC close to 100% in train set, implying that the global model is correctly classified into the appropriate group.

Moreover, %CC in train set is also similar to test sets, it can be suggested that the global model is not overfitted. Considering to lower scaling value, the lower %CC is provided. On the other hand, when ω is raised until the classification model is

either stabilized or slightly decreased, the optimal scaling value for each case is directly determined. Herein, the optimal scaling value is equal to 0.002 for both case I and case II.

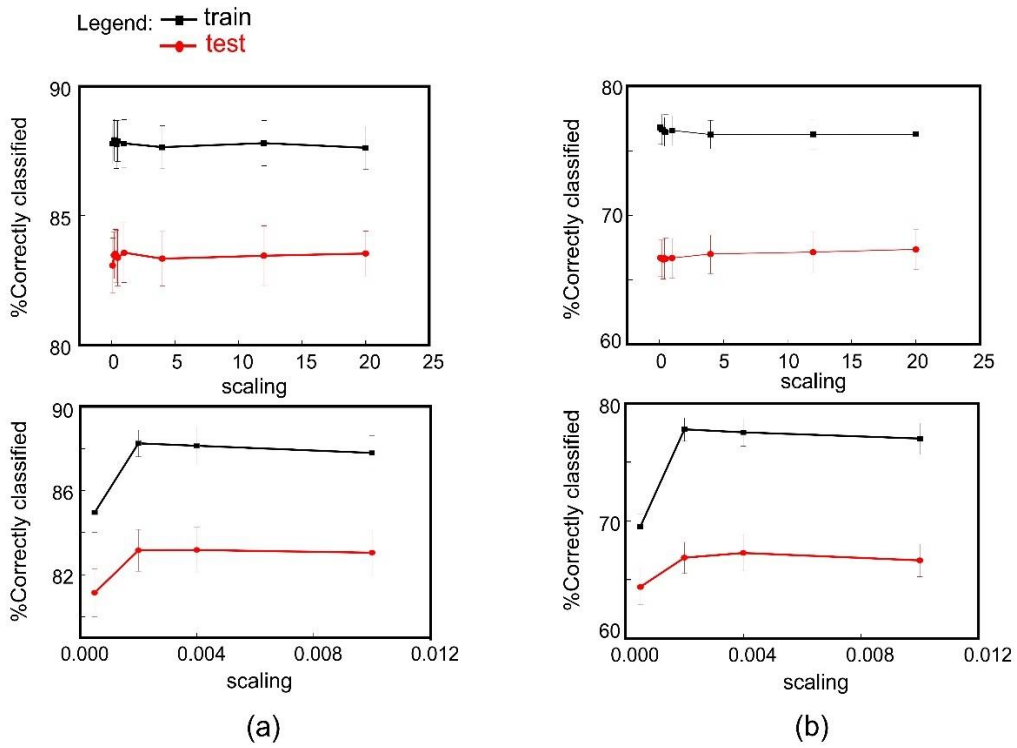


Figure 4.9 Percent Correct Classified (%CC) of the training set and test set (average from 100 iterations) with the different scaling value (*ω*) used to build the supervised SOM model for (a) case I: PL2 vs weedy, (b) case II: RD49 vs weedy.

Besides the scaling value that affects the model performance, size map is a crucial factor that needs to optimize. If the number of samples is smaller than the number of map units render to overfitting problem [148]. On the other hand, if the number of samples is more extensive than the number of map units, detailed information might be lost in the process [147].

Therefore, an appropriate size map will provide better knowledge on the clustering qualities of SOMs [143]. However, there is no theoretical principle to indicate the suitable size map. Quantitative indicators such as quantization error (QE), topographic error (TE) and eigenvalues have proven to be relevant tools to facilitate the selection of the map size [144, 147].

In this study, we will examine the effect of size map on the extended SOMs. Comparisons at the various size of the sample will be made using the percent Correctly Classified (%CC) as shown in Figure 4.10. The small map size of test set (256 and 625) in both case I and case II produce a slight fluctuation of %CC. The map sizes ranging from 1296 onwards, extended SOM produced a stable of %CC. From the result, it might be suggested that a large map unit size probably is highly efficient than a small map size. Although dimensions of the data increases become more critical to visualize and classify samples, unfortunately, time to compute them also increases. Herein, the sample size = 1296 was selected to be the optimal size map.

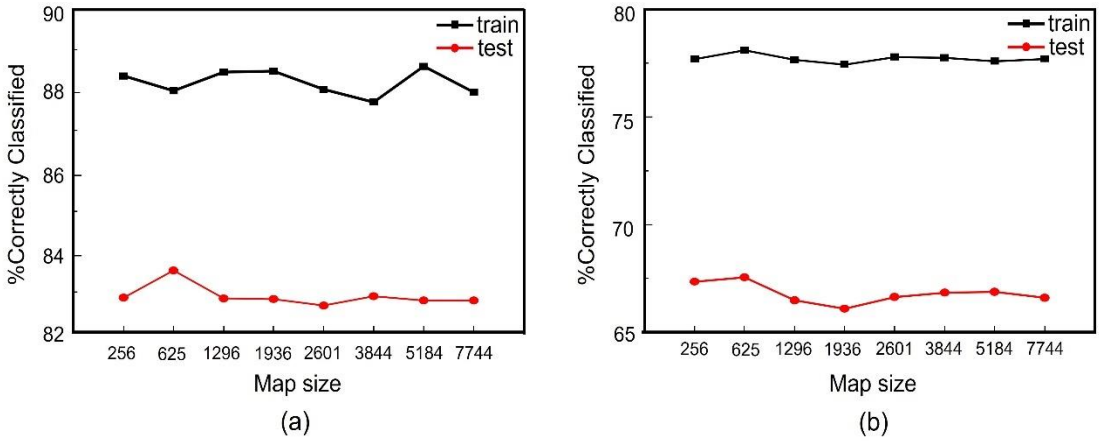


Figure 4.10 Percent Correct Classified (%CC) of the training set and test set (average from 100 iterations) with the different map size used to build the supervised SOM model for (a) case I : PL2 vs weedy, (b) case II : RD49 vs weedy

In order to prove our hypothesis, optimal scaling value and size map are then used to establish supervised SOMs map. The SOM component planes of the input variables for the samples are shown in Figure 4.11. Each hexagonal unit on the map represents a particular location on the different component planes with the exact location on the unit map. The values of the various components are represented using different colors. From Figure 4.11, it can be seen that there is a complete separation between groups of rice samples on all of these maps. As a result, it emphasizes the importance of optimizing the scaling value and map size that affect the SOM model.

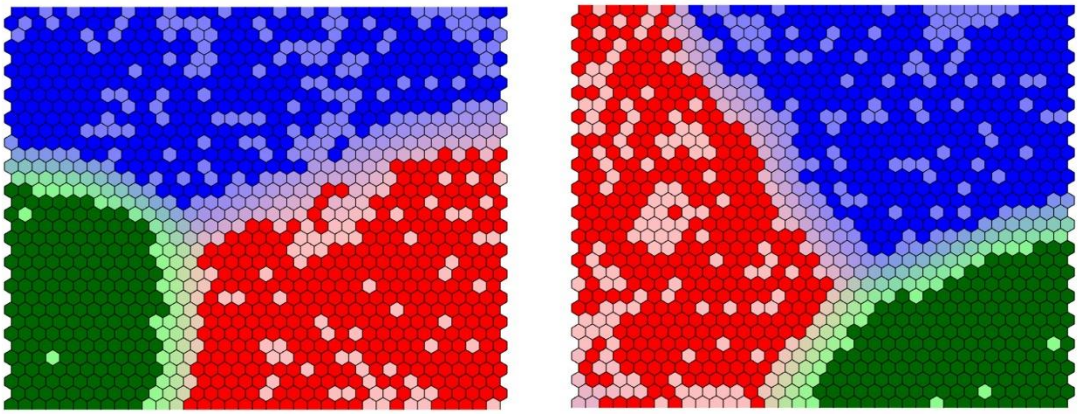


Figure 4.11 supervised SOMs using the optimal scaling values (ω = 0.002) and optimal size map (size = 1296). (a) Case (I): PL2 vs weedy (b) Case (II): RD49 vs weedy

4.3.5 Image Based Classification

Spectral information (X-matrix) of the imaged sample representing its physicochemical properties is extracted directly from the image segment as the main region of interest. Then, all of this spectral information was used to build a global map using supervised SOMs based on optimal scaling value and size map. The global map was mapped to other hyperspectral images to perform classification maps of weedy seed and cultivated seed by simple display being imaged using RGB (red, green, and blue). In other word, the color on any pixel was generated by matching with supervised SOM map. Different image pixel colors provide different information about seeds' morphological features (e.g., color, size, shape, and surface structure) [31]. Therefore, the same class of samples will provide the same color tone. Herein, weedy seed, cultivated seed and background were determined as red (Rpixel), blue (Bpixel) and green (Gpixel), respectively. In the case, green pixel represent the background on the image (irrelevant pixel). Based on the global models, the final color images for the classification are shown in Figure 4.12. The images clearly show a few seeds in both the weedy and cultivated rice seeds that were misclassified. From the result, it might be suggested that developed supervised SOMs have the potential to classify.

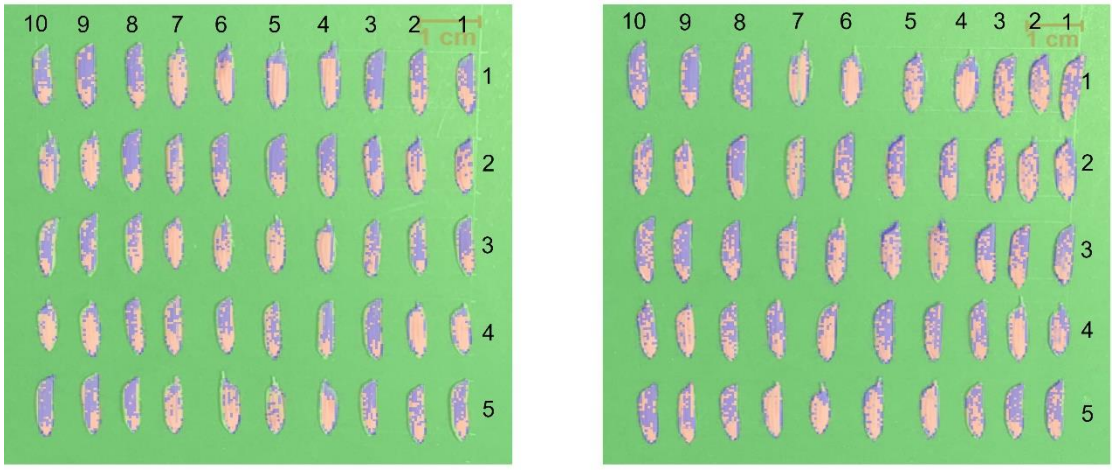


Figure 4.12. Classification map of rice seed based on spectra information of HSI imaging created using the global model of system: (a) Case I : Weedy vs PL2 and (b) Weedy vs RD49

4.3.6 Receiver operating characteristic (ROC) curve

Moreover, the limitations in terms of accuracy as a measure of decision performance require introduction concepts of "sensitivity" and "specificity" tests [149]. The receiver operation characteristic (ROC) curve has been widely assessed the effectiveness of target detection, which represents the varying relationship between the true positive rate (TPR) and the false positive rate (FPR) [150]. Qualitatively, the closer to the upper left corner of the plot in the ROC curves, the better the performance. For the area under the ROC curve (AUC), it measures the accuracy of a prediction test. The area under the ROC curve can assume any value between 0.0 and 1.0. A test with an area of 1.0 is perfectly accurate, whereas a test with an area of 0.0 is perfectly inaccurate [151]. In other words, an enormous value of AUROC indicates a better outcome [152]. Herein, ROC curve and area under the ROC curve are applied to evaluate the model performance. The plot of TPR versus FPR by varying the threshold T_f is shown in Figure 4.13. The ROC curve of is shown in black solid line while the AUC is shown in bold italic letters. From Figure 13a, both the ROC curve

and AUC perform well. According to different thresholds (T_f), ROC curve is monitored. From the result, threshold (T_f) equal 0.5068 and 0.4925 of case I and case II, respectively, are the best threshold (T_f) that make the best performance of the model. Therefore, these thresholds were further used as the index to classify the weedy rice and cultivated rice that were calculated from the global model.

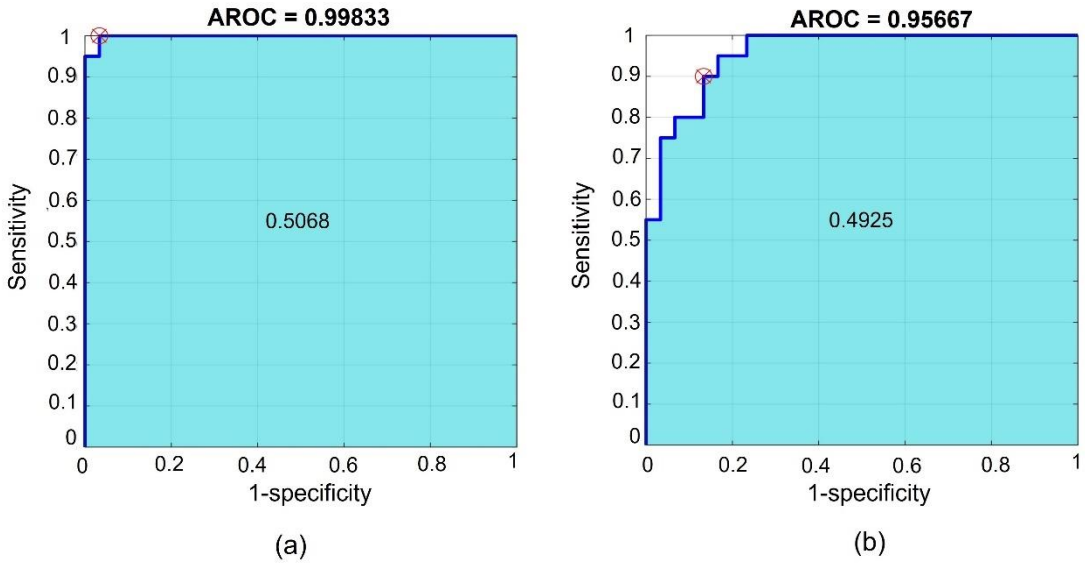


Figure 4.13 ROC curve (a) case I : Weedy vs PL2 (b) case II : Weedy vs RD49

4.3.7 Classification of weedy rice by using the number of pixels (R, G, and B) of an image

According to Figure 4.12, the color on any pixel was generated by matching with supervised SOM map. It can be seen that every single seed consists of Rpixel and Bpixel. In order to classify weedy and cultivated rice, the number of pixels (R–Red, G–Green, and B–Blue) of each seed image was calculated, as shown in Figure 4.14. Pixel values ratio (R/R+B) of samples that are higher than or equal to the threshold values 0.5068 (from ROC curve) were classified as weedy rice, and they were represented bar, whereas the ratio that is less than 0.5068 were classified as cultivated rice (blue bar). The ratio was shown in Figure 4.14b. Only one misclassified seed was obtained (seed number 44, S_{44}). In other words, 49 out of 50 seeds were able to accurately classify, resulting in a 98 percent accuracy rate. As a

result, it can be assumed that pixel classification by using global model matching is an effective alternative way to use the high performance of the HSI technique

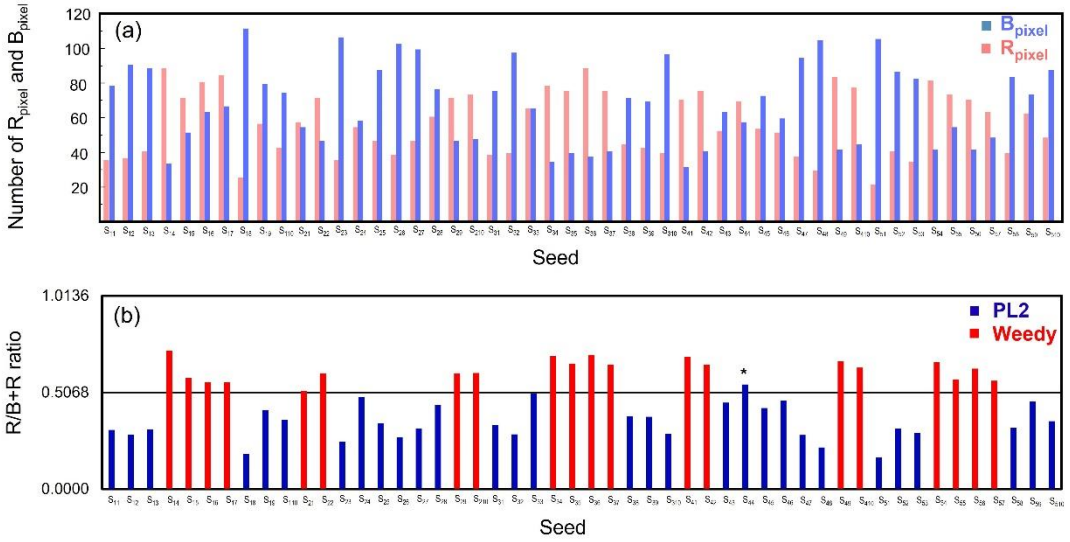


Figure 4.14 Predictive result (case I : Weedy vs. PL2) after using global map (supervised SOM map) matching with color on any pixel image (a) Number of R_{pixel} and B_{pixel} (b) $R_{pixel}/B_{pixel}+R_{pixel}$ ratio, where * is a symbol indicating that the seed was misclassified

To further prove the performance of the global model, it was applied in case II: Weedy vs. RD49. The predictive results were shown in Figure 4.15. six misclassified seeds were obtained (seed number: 18; S_{18}, 21; S_{18}, 24; S_{24}, 28; S_{18}, 210; S_{210}, 45; S_{45}), including two seeds and four seeds of weedy class and cultivated class, respectively. To be specific, 44 out of 50 seeds were able to accurately classify, resulting in the 88 percent accuracy rate. As a result, it can be seen that the percent accuracy rate is slightly less than the case I. One of the possible reasons is the similar variation between weedy and cultivated rice seed, resulting in the global model established from both variations of weedy and cultivated rice providing minor errors. Therefore, it could be proved that the global model coupled with hyperspectral imaging technique can be a potential tool for fast and accurate classification of seeds.

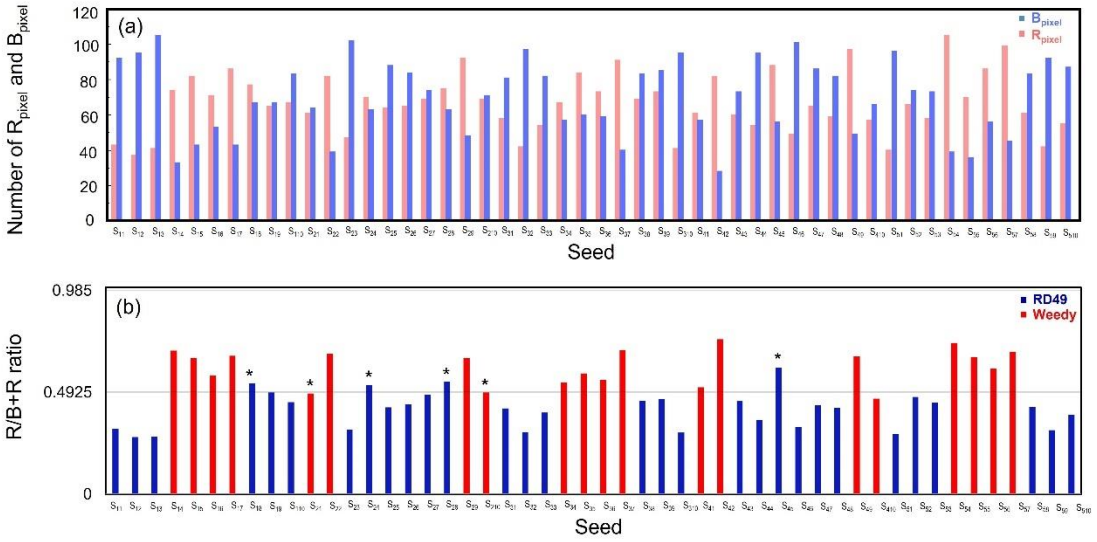


Figure 4.15 Predictive result (case II : Weedy vs RD49) after using global map (supervised SOM map) matching with color on any pixel image (a) Number of R_{pixel} and B_{pixel} (b) $R_{pixel}/B_{pixel}+R_{pixel}$ ratio, where * is a symbol indicating that the seed was misclassified

4.3.8 The evaluation study of bias and overfitting precision in the global model concept

Concurrently, the biased testing was investigated by randomizing the class vector of the dataset for the entire data. Therefore, some data is not assigned to the correct class and some data is still assigned to the class. If the model is not overfitting, the prediction should be closed to background prediction (100 / N; N = number of classes), which equates to 50% in the case. Therefore, based on this basic concept idea, the virtual global map was constructed from two underlying data system: PL2 and weedy rice. Regarding to this virtual global map, it was used to predict seed samples by matching color on any pixel image which the result is shown in Figure 4.16 and 4.17. The finding showed that they produced lower percentage accuracy predictions. According to Figure 4.16, the virtual global map can correctly classified seed 13 of 50 seed, which represents a percentage accuracy of predictions at 26% with the same probability of misclassification occurring in weedy and cultivated rice. The result is consistent with Figure 4.17; the accuracy prediction percentage is 66%, which is still lower compared to our global map as shown in Figure 4.14. As a result

of this finding, it is reasonable to deduce that our global map is unlikely to be biased or overfitting. Furthermore, there are possibly conclusive that the genuine underlying chemical component of the seed sample used to construct our global map plays a critical part in our global map's ability to anticipate accurately.

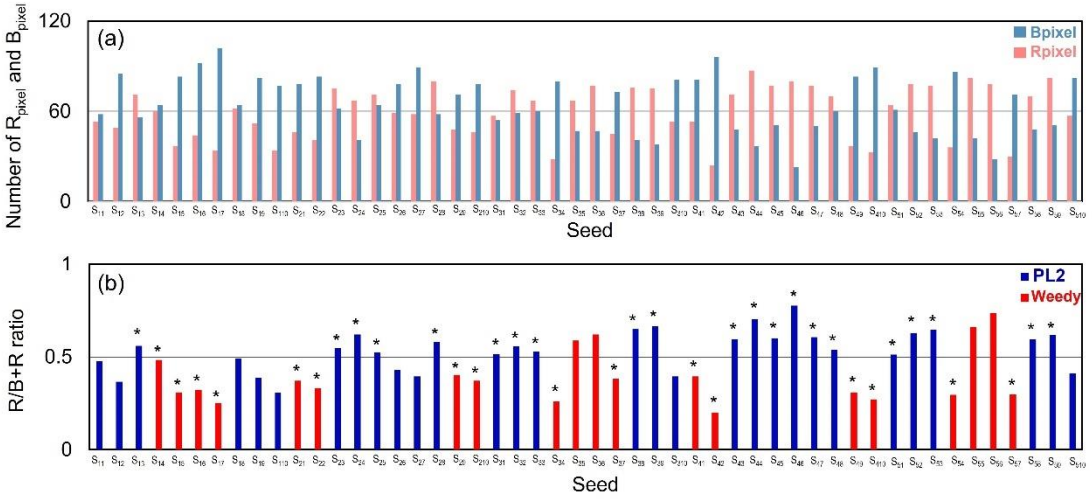


Figure 4.16 Predictive result after using global map which was constructed from PL2 matching with color on any pixel image (a) Number of R_{pixel} and B_{pixel} (b) $R_{pixel}/B_{pixel}+R_{pixel}$ ratio, where * is a symbol indicating that the seed was misclassified

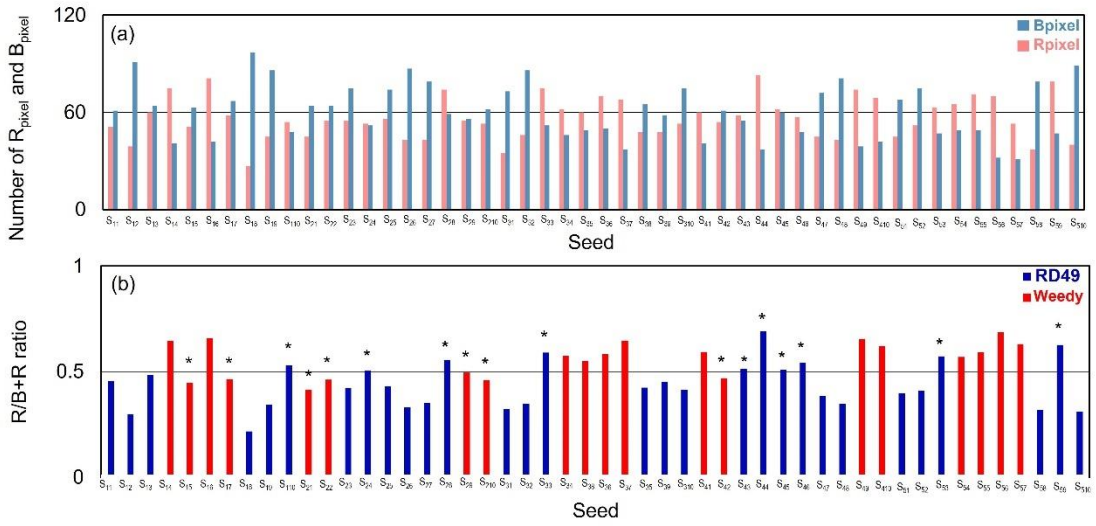


Figure 4.17 Predictive result after using global map which was constructed from Weedy matching with color on any pixel image (a) Number of R_{pixel} and B_{pixel} (b) $R_{pixel}/B_{pixel}+R_{pixel}$ ratio, where * is a symbol indicating that the seed was misclassified

4.3.9 Application of direct analysis in real time mass spectrometry (DART-MS) for rice determination

Weedy and cultivated rice (PL2) were tested directly via DART-MS without any sample preparations. The relatively simple mass spectra typical of the DART ionization method display a single $[M + H]^+$ peak representative of the individual component exclusive to each formulation as sown in Figure 4.18. The peaks at m/z 121, 159, 163, 239, and 245 were found in both rice seed samples, but with significant variances in the absolute intensities (Figure 4.18c). To easily identify m/z which distinct type of rice samples, the variance of the average mass spectra was calculated and plotted in Figure 4.18d–4.18f. Any overtone regions which provide a high variance indicate the possible features to discriminate the type of rice samples. It can be seen that the peak at m/z 245 and 263, which correspond to species of linolenic acid and pentadecenoic acid, respectively, appeared sorely in weedy rice (Figure 4.18a). This result is consistent with the variance of weedy and PL2 that shows a prominent peak at m/z 245 and 263. The characteristic MS/MS fragments of

protonated linolenic acid (m/z 245) and pentadecenoic acid (m/z 263) including other fragments are shown in Table 4.2 which were confirmed by listed in the reported literature [153]. From the results, it can be assumed that it is likely to identify compounds that distinguish between weed rice and cultivated rice by DART-MS.

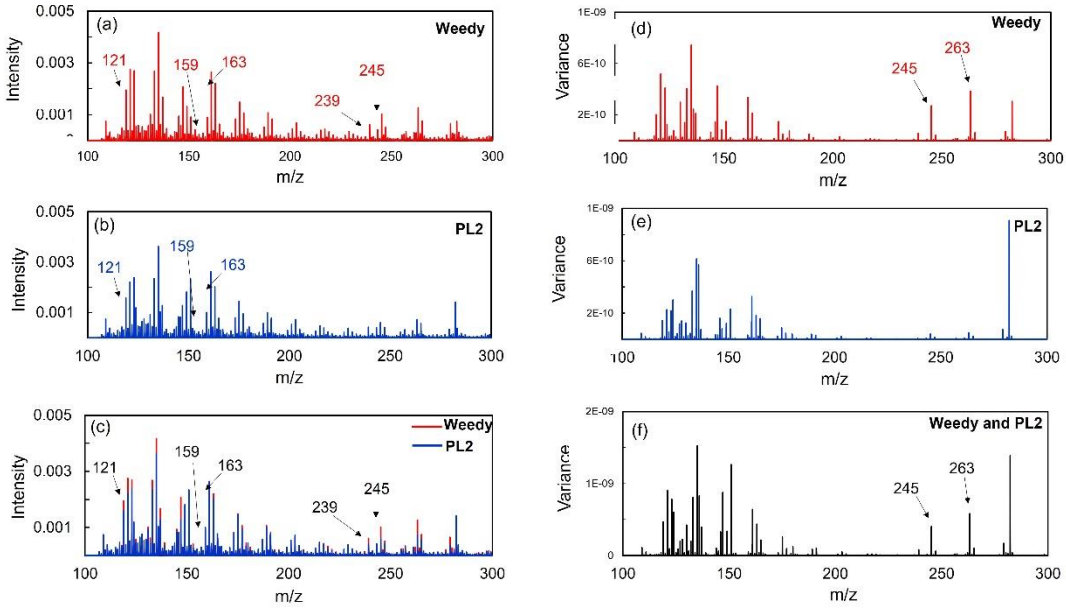


Figure 4.18 Chemical fingerprint of rice paddy sample corresponding to mass spectrum acquisition in positive ion detection mode by DART-MS. (a) Weedy rice, (b) PL2, and (c) overlap peaks between Weedy rice and PL2. On the right-hand side, it showed a variance of DART mass spectrum of rice seed (d) Weedy rice (e) Cultivated rice (PL2), and (f) Weedy and Cultivated rice

Table 4.2 The single grain electrospray ionization (SG-ESI)-MS/MS results of rice samples

m/z	Charge form	Molecular formula	Compound
121	$[M+H]^+$	$C_9H_{18}O_2$	Nonanoic acid
159	$[M+H]^+$	$C_9H_{18}O_2$	Nonanoic acid
163	$[M+H]^+$	$C_{10}H_{14}N_2$	Nicotine (Internal standard)
239	$[M+H]^+$	$C_{16}H_{30}O_2$	Palmitoleic acid
		$C_{18}H_{32}O_2$	Linoleic acid
245	$[M+Na]^+$	$C_{18}H_{30}O_2$	Linolenic acid
245	$[M+Na]^+$		Linolenic acid
263	$[M+Na]^+$	$C_{15}H_{28}O_2$	Pentadecenoic acid

4.4 Conclusion

In the present study, the new classification approach for the HSI image is proposed by using supervised SOMs. This developed supervised SOMs as a global map was applied on the HSI image to classify weedy from cultivated rice directly from paddy seed. The weedy and cultivated rice paddy samples' physical features were explored. Due to the similarity of their morphological characteristics of rice husk, the results revealed no significant differences in their physical appearances. It is consistent with SEM analysis; there was no significant difference in native morphology between these two of rice under SEM. According to NIR hyperspectral measurement, four important overtone regions were selected using the variance, including 1068, 1148, 1400, and 1690 nm. The developed supervised SOMs (global map) was applied on the pair-wise HSI to generate the supervised global SOM map that visualize the unit of each class. Two parameters, including scaling value (*w)* and map size of the global map were optimized. The optimal scaling value (*w)* of the

develop SOM model is well optimized to prevent the overfitting problem. To achieve the reliable prediction, the efficiency of the SOM classifier was validated using 100 different training and test sets. In order to access the developed classification model, %CC was used as the performance indices to evaluate the classification performance. The result showed that the global model provides a high value of %CC at 88.45% and 77.67% for the case I (weedy vs.PL2) and case II (weedy vs. RD49), respectively. This suggests that the global map has the potential to discriminate the weedy rice from cultivated rice seeds.

Furthermore, global map can use to classified rice seed sample based on image classification. According to ROC curve, the result showed that the threshold (T_f) equal 0.5068 and 0.4925 of case I and case II, respectively, are the best threshold (T_f) that make the best performance of the model. In the future, a worldwide model based on NIR hyperspectral imaging applications may become a practical approach that can be carried out quickly and accurately without the need for additional chemicals or processes to evaluate and inspect rice seed quality.

CHAPTER V
CONCLUSION

Weedy rice is one of the most notorious weeds occurring in rice-growing areas, especially in South-East Asia. Weedy rice especially in form of paddy seed is difficult to manage and separate as they provide common features (morphological resemblance) to cultivated rice. Therefore, the quality assessment method for evaluation the rice paddy seed is required to prevent the wide-spreading of the weedy rice. This work presents a modification of self-organizing map (SOMs) for the classification of weedy rice from cultivated rice *via in situ* direct sample analysis from paddy seed using near-infrared (NIR) spectroscopy and hyperspectral NIR camera.

In the study, cultivated rice were collected from Lifestyle and Spirit of Thai Farmers-Nahai Chai Learning Center at Supanburi province, Thailand. They are certificated by Rice Department Ministry of Agriculture and Cooperatives, Thailand. Moreover, weedy rice were collected from the local fields at Phrom Phiram district, Phitsanulok province, Thailand. After sample collection, to eliminate contaminated particles and other impurities, the rice samples were pretreated with a cyclone vacuum machine. Optical microscopy and thermogravimetric analysis (TGA) were used to evaluate rice physical features and thermal behavior, while DART-MS was used to monitor the volatile chemical profiles on the rice husk. Regarding direct sample analysis, a near-infrared with reflectance accessory was employed to acquire NIR spectra. The NIR spectra were then preprocessed using various methods, including the Savitzky-Golay polynomial, standard normal variate (SNV), mean-centered, and second derivative pretreatment, to convert the raw data into a meaningful and efficient form. Self-organizing maps were well-optimized and used to classify weedy samples from four different types of cultivated rice. The results were confirmed and accomplished with the remarkable predictive value of 91% to 99% for precision and 88% to 99% for accuracy, respectively. By comparing with the basic statistical classifier, the modified SOMs demonstrates the powerful method to discriminate and classify the rice types directly from NIR spectra. It is reasonable to conclude that the

modified SOM algorithm will be able to further behave on data provided via specialized techniques such as electronic noses and hyperspectral cameras.

Additionally, the experiment was further undertaken by utilizing the modified SOMs applied on the pair-wise hyperspectral images to generate the supervised global SOM map. Each hyperspectral pixel from the sample image was verified with the global map, then the color of the best map unit (BMU) was re-projected on the image pixel. The steps were repeated until all image pixels were presented in BMU color. Then, the image pixels were replaced by the color shades which represent each class sample. The classification criterion was achieved by considering the ratio of the projected color on the sample image. The suitable threshold in order to be used for classifying the object class was optimized using Receiver operating characteristic (ROC) curve. The accuracy of the weedy seed classification was 90%, suggesting the usefulness of a global model for seed quality evaluation.

This empirical evidence may lead to assumption that NIR hyperspectral imaging has been successfully applied to seed quality monitoring using either actual HSI data or NIR spectra for the analysis. Furthermore, this present work is likely to be extended to quantitative approach for the determination of weedy seed proportion in the cultivated rice seed based on NIR hyperspectral imaging technique. A developed global model based on NIR hyperspectral imaging systems may become a feasible strategy for evaluating and inspecting agricultural plant seed quality that can be conducted rapidly and precisely without the use of extra chemicals or operations.

APPENDICES

1. Cyclone vacuum machine

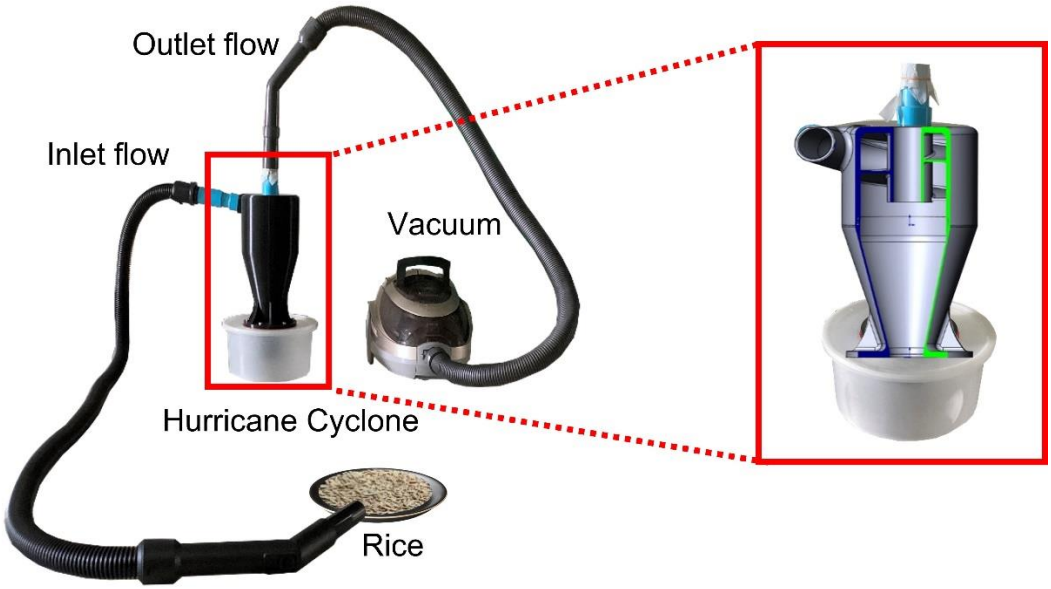


Figure A1 cyclone vacuum machine

2. Prediction of the mixed proportion

It should be noted that the detection on the mixed proportion of weedy rice seems more significant than the classification in real application. However, the prediction of the mixed proportion usually could not be discovered until the classification of the target object (weedy rice) is completely achieved especially for the unknown system. In our study, to classify the weedy rice directly from paddy seed by using NIR technique has not been reported elsewhere so far. Therefore, to prove the capability of NIR technique combined with our modified SOM method in order to discriminate the weedy rice (the target object) from the cultivated rice is the first priority. However, it is worth to try now at least to reveal the possibility for our developed SOMs to classify the mixed proportion of weedy rice sample. The samples were prepared with the mixed proportion of weedy rice in the cultivated rice (RD49) at different %w/w (20%, 40%, 60%, 80% and 100%).

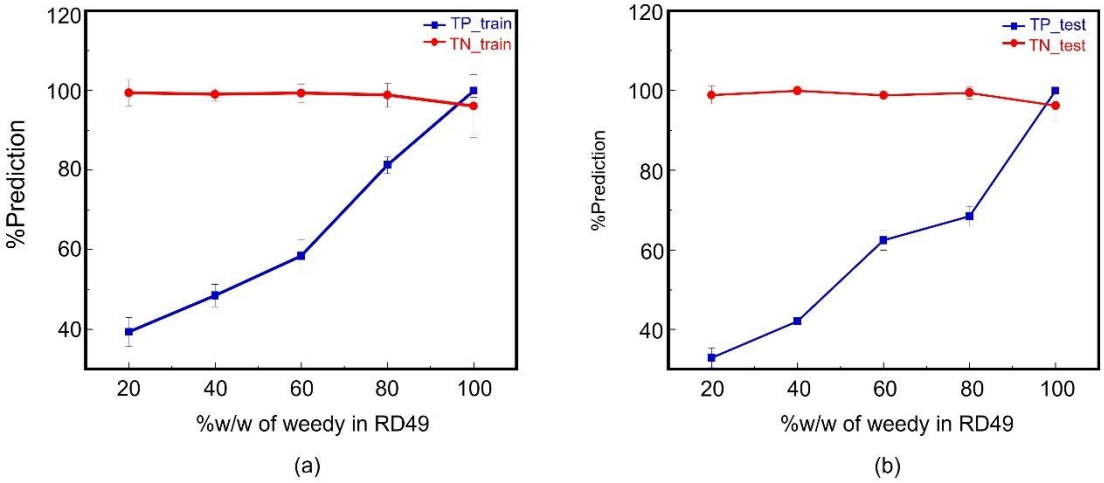


Figure A2 (a) and (b) show %prediction of the training set and test set (average from 100 iterations), respectively, which are estimated using the reference SOM map from the case II (in the manuscript). TP is the number of weedy correctly classified, and TN is the number of cultivated rice correctly classified. In the case of TN (red line), the rate of %CC is stable closed to 100% while the % prediction of TP (blue line) is directly related to the percent of weedy rice in the mixed sample. The higher proportion of weedy rice, the higher predictive rate occurred.

In the present work, the modified SOMs was initially not designed to be used for quantitative analysis, however, the percent prediction of weedy rice (TP) are surprisingly related to the %w/w of weedy rice in the mixed sample. From the results, this suggests that it is highly possible to improve our developed SOMs for further use in quantitative analysis.

REFERENCES

1. Delouche, J. C.; Burgos, N. R.; Labrada, R.; Gealy, D. R., Weedy rices: origin, biology, ecology and control. *Food & Agriculture Org*. **2007**, *188*.
2. Nadir, S.; Xiong, H.-B.; Zhu, Q.; Zhang, X.-L.; Xu, H.-Y.; Li, J.; Dongchen, W.; Henry, D.; Guo, X.-Q.; Khan, S.; Suh, H.-S.; Lee, D. S.; Chen, L.-J., Weedy rice in sustainable rice production. A review. *Agronomy for Sustainable Development* **2017,** *37* (5).
3. Fu, F.-F.; Ye, R.; Xu, S.-P.; Xue, H.-W., Studies on rice seed quality through analysis of a large-scale T-DNA insertion population. *Cell Research* **2009,** *19* (3), 380-391.
4. Almekinders, C. J.; Louwaars, N. P.; De Bruijn, G. H., Local seed systems and their importance for an improved seed supply in developing countries. *Euphytica* **1994,** *78* (3), 207-216.
5. Muthayya, S.; Sugimoto, J. D.; Montgomery, S.; Maberly, G. F., An overview of global rice production, supply, trade, and consumption. *Annals of the new york Academy of Sciences* **2014,** *1324* (1), 7-14.
6. Vaughan, D. A.; Morishima, H.; Kadowaki, K., Diversity in the Oryza genus. *Current Opinion in Plant Biology* **2003,** *6* (2), 139-146.
7. Marambe, B., Weedy rice—evolution, threats and management. *Trop Agriculturist* **2009,** *157*, 43-64.
8. Zhang, Y.; Gao, J.; Cen, H.; Lu, Y.; Yu, X.; He, Y.; Pieters, J. G., Automated spectral feature extraction from hyperspectral images to differentiate weedy rice and barnyard grass from a rice crop. *Computers and Electronics in Agriculture* **2019,** *159*, 42-49.
9. Prathepha, P., Seed morphological traits and genotypic diversity of weedy rice (Oryza sativa f. spontanea) populations found in the Thai Hom Mali rice fields of north-eastern Thailand. *Weed Biology and Management* **2009,** *9* (1), 1-9.
10. Grimm, A.; Sahi, V. P.; Amann, M.; Vidotto, F.; Fogliatto, S.; Devos, K. M.; Ferrero, A.; Nick, P., Italian weedy rice—A case of de-domestication? *Ecology and evolution* **2020,** *10* (15), 8449-8464.
11. Yu, G.-q.; Bao, Y.; Shi, C.-h.; Dong, C.-q.; Ge, S., Genetic diversity and

population differentiation of Liaoning weedy rice detected by RAPD and SSR markers. *Biochemical Genetics* **2005,** *43* (5-6), 261-270.

12. Chauhan, B. S.; Abeysekera, A. S. K.; Wickramarathe, M. S.; Kulatunga, S. D.; Wickrama, U. B., Effect of rice establishment methods on weedy rice (Oryza sativa L.) infestation and grain yield of cultivated rice (O. sativa L.) in Sri Lanka. *Crop Protection* **2014,** *55*, 42-49.
13. Fabiyi, S. D.; Vu, H.; Tachtatzis, C.; Murray, P.; Harle, D.; Dao, T. K.; Andonovic, I.; Ren, J.; Marshall, S., Varietal classification of rice seeds using RGB and hyperspectral images. *IEEE Access* **2020,** *8*, 22493-22505.
14. Teye, E.; Amuah, C. L.; McGrath, T.; Elliott, C., Innovative and rapid analysis for rice authenticity using hand-held NIR spectrometry and chemometrics. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **2019,** *217*, 147-154.
15. Huebner, F. R.; Bietz, J. A.; Webb, B. D.; Juliano, B. O., Rice cultivar identification by high-performance liquid chromatography of endosperm proteins. *Cereal Chemistry* **1990,** *67* (2), 129.
16. Peng, Z.; Yuan, X.; Huang, Y.; Mo, J.; Tan, J.; Zhou, H.; Wang, L., Application of denaturing high-performance liquid chromatography for rice variety identification and seed purity assessment. *Molecular Breeding* **2016,** *36* (2), 19.
17. Chauhan, B. S., Strategies to manage weedy rice in Asia. *Crop Protection* **2013,** *48*, 51–56.
18. Abraham, C. T.; Jose, N., Weedy rice invasion in rice fields of India and management options. *Journal of Crop and Weed*, **2014,** *10* (2), 365-374.
19. Ferrero, A.; Vidotto, F.; Balsari, P.; Airoldi, G., Mechanical and chemical control of red rice (Oryza sativa L. var. sylvatica) in rice (Oryza sativa L.) pre-planting. *Crop Protection* **1999,** *18* (4), 245-251.
20. Shanthi, P.; Jebaraj, S.; Geetha, S.; Aananthi, N., DNA finger printing of salt tolerant and susceptible genotypes using microsatellite markers in rice (Oryza sativa L.). *International Journal of Plant Breeding and Genetics* **2012,** *6* (4), 206-216.

21. Ozaki, Y.; McClure, W. F.; Christy, A. A., *Near-infrared spectroscopy in food science and technology*. John Wiley & Sons **2006**.

22. Siesler, H. W.; Ozaki, Y.; Kawata, S.; Heise, H. M., *Near-infrared spectroscopy: principles, instruments, applications*. John Wiley & Sons **2008**.

23. Ozaki, Y., *Near-Infrared Spectroscopy: Theory, Spectral Analysis, Instrumentation, and Applications*. Springer Nature **2021**.

24. Ozaki, Y., Near-infrared spectroscopy—its versatility in analytical chemistry. *Analytical sciences* **2012,** *28* (6), 545-563.

25. Cen, H.; He, Y.; Huang, M., Combination and comparison of multivariate analysis for the identification of orange varieties using visible and near infrared reflectance spectroscopy. *European Food Research and Technology* **2007,** *225* (5-6), 699-705.

26. Christy, A. A.; Kasemsumran, S.; Du, Y.; Ozaki, Y., The detection and quantification of adulteration in olive oil by near-infrared spectroscopy and chemometrics. *Analytical Sciences* **2004,** *20* (6), 935-940.

27. Porep, J. U.; Kammerer, D. R.; Carle, R., On-line application of near infrared (NIR) spectroscopy in food production. *Trends in Food Science & Technology* **2015,** *46* (2, Part A), 211-230.

28. Zhang, J.; Li, M.; Pan, T.; Yao, L.; Chen, J., Purity analysis of multi-grain rice seeds with non-destructive visible and near-infrared spectroscopy. *Computers and Electronics in Agriculture* **2019,** *164*, 104882.

29. Kong, W.; Zhang, C.; Liu, F.; Nie, P.; He, Y., Rice seed cultivar identification using near-infrared hyperspectral imaging and multivariate data analysis. *Sensors* **2013,** *13* (7), 8916-8927.

30. Chen, H.; Tan, C.; Lin, Z., Authenticity detection of black rice by near-infrared spectroscopy and support vector data description. *International journal of analytical chemistry* **2018,** *2018*, 8032831.

31. ElMasry, G.; Mandour, N.; Al-Rejaie, S.; Belin, E.; Rousseau, D., Recent applications of multispectral imaging in seed phenotyping and quality monitoring—An overview. **2019,** *19* (5), 1090.

32. Mendoza, F.; Lu, R.; Ariana, D.; Cen, H.; Bailey, B., Integrated spectral and

image analysis of hyperspectral scattering data for prediction of apple fruit firmness and soluble solids content. *Postharvest Biology and Technology* **2011,** *62* (2), 149-160.

33. Huang, W.; Zhang, B.; Li, J.; Zhang, C. In *Early detection of bruises on apples using near-infrared hyperspectral image*, Piageng 2013: Image Processing and Photonics for Agricultural Engineering. *International Society for Optics and Photonics* **2013**, *8761*, 87610.
34. Qin, J.; Burks, T. F.; Zhao, X.; Niphadkar, N.; Ritenour, M. A., Development of a two-band spectral imaging system for real-time citrus canker detection. *Journal of Food Engineering* **2012,** *108* (1), 87-93.
35. Lleó, L.; Barreiro, P.; Ruiz-Altisent, M.; Herrero, A., Multispectral images of peach related to firmness and maturity at harvest. *Journal of Food Engineering* **2009,** *93* (2), 229-235.
36. Huang, W.; Li, J.; Wang, Q.; Chen, L., Development of a multispectral imaging system for online detection of bruises on apples. *Journal of Food Engineering* **2015,** *146*, 62-71.
37. Lu, R.; Peng, Y., Hyperspectral scattering for assessing peach fruit firmness. *Biosystems engineering* **2006,** *93* (2), 161-171.
38. Su, W.-H., Advanced Machine Learning in Point Spectroscopy, RGB- and Hyperspectral-Imaging for Automatic Discriminations of Crops and Weeds: A Review. *Smart Cities* **2020,** *3* (3).
39. ElMasry, G. M.; Nakauchi, S., Image analysis operations applied to hyperspectral images for non-invasive sensing of food quality–a comprehensive review. *Biosystems engineering* **2016,** *142*, 53-82.
40. Caporaso, N.; Whitworth, M. B.; Fisk, I. D., Near-Infrared spectroscopy and hyperspectral imaging for non-destructive quality assessment of cereal grains. *Applied Spectroscopy Reviews* **2018,** *53* (8), 667-687.
41. Barbedo, J. G. A.; Guarienti, E. M.; Tibola, C. S., Detection of sprout damage in wheat kernels using NIR hyperspectral imaging. *Biosystems Engineering* **2018,** *175*, 124-132.
42. Guo, L. B.; Yu, Y. X.; Yu, H. Y.; Tang, Y.; Li, J.; Du, Y.; Chu, Y. W.; Ma,

S. X.; Ma, Y. Y.; Zeng, X. Y., Rapid quantitative analysis of adulterated rice with partial least squares regression using hyperspectral imaging system. *J. Sci. Food Agric.* **2019,** *99* (12), 5558-5564.

43. He, X. T.; Feng, X. P.; Sun, D. W.; Liu, F.; Bao, Y. D.; He, Y., Rapid and Nondestructive Measurement of Rice Seed Vitality of Different Years Using Near-Infrared Hyperspectral Imaging. *Molecules* **2019,** *24* (12), 14.
44. Siripatrawan, U.; Makino, Y., Monitoring fungal growth on brown rice grains using rapid and non-destructive hyperspectral imaging. *Int. J. Food Microbiol.* **2015,** *199*, 93-100.
45. Bauriegel, E.; Giebel, A.; Geyer, M.; Schmidt, U.; Herppich, W. B., Early detection of Fusarium infection in wheat using hyper-spectral imaging. *Computers and Electronics in Agriculture* **2011,** *75* (2), 304-312.
46. Badaró, A. T.; Garcia-Martin, J. F.; del Carmen Lopez-Barrera, M.; Barbin, D. F.; Alvarez-Mateos, P., Determination of pectin content in orange peels by near infrared hyperspectral imaging. *Food chemistry* **2020,** *323*, 126861.
47. Faqeerzada, M. A.; Perez, M.; Lohumi, S.; Lee, H. S.; Kim, G.; Wakholi, C.; Joshi, R.; Cho, B. K., Online Application of a Hyperspectral Imaging System for the Sorting of Adulterated Almonds. *Applied Sciences* **2020,** *10* (18), 6569.
48. Singh, T.; Garg, N. M.; Iyengar, S. R. S., Nondestructive identification of barley seeds variety using near-infrared hyperspectral imaging coupled with convolutional neural network. *Journal of Food Process Engineering* **2021**, e13821.
49. Wang, J.; Zhang, C.; Shi, Y.; Long, M. J.; Islam, F.; Yang, C.; Yang, S.; He, Y.; Zhou, W. J., Evaluation of quinclorac toxicity and alleviation by salicylic acid in rice seedlings using ground-based visible/near-infrared hyperspectral imaging. *Plant Methods* **2020,** *16* (1), 16.
50. Malegori, C.; Oliveri, P.; Mustorgi, E.; Boggiani, M. A.; Pastorini, G.; Casale, M., An in-depth study of cheese ripening by means of NIR hyperspectral imaging: Spatial mapping of dehydration, proteolysis and lipolysis. *Food Chemistry* **2021,** *343*, 128547.
51. Wu, D.; Chen, J.; Lu, B.; Xiong, L.; He, Y.; Zhang, Y., Application of near

infrared spectroscopy for the rapid determination of antioxidant activity of bamboo leaf extract. *Food Chemistry* **2012,** *135* (4), 2147-2156.

52. Martelo-Vidal, M. J.; Vázquez, M., Classification of red wines from controlled designation of origin by ultraviolet-visible and near-infrared spectral analysis. *Ciência e técnica vitivinícola* **2014,** *29* (1), 35-43.
53. Cortés, V.; Blasco, J.; Aleixos, N.; Cubero, S.; Talens, P., Visible and near-infrared diffuse reflectance spectroscopy for fast qualitative and quantitative assessment of nectarine quality. *Food and Bioprocess Technology* **2017,** *10* (10), 1755-1766.
54. Firmani, P.; Bucci, R.; Marini, F.; Biancolillo, A., Authentication of "Avola almonds" by near infrared (NIR) spectroscopy and chemometrics. *J. Food Compos. Anal.* **2019,** *82*, 5.
55. dos Santos Costa, D.; Oliveros Mesa, N. F.; Santos Freire, M.; Pereira Ramos, R.; Teruel Mederos, B. J., Development of predictive models for quality and maturation stage attributes of wine grapes using vis-nir reflectance spectroscopy. *Postharvest Biology and Technology* **2019,** *150*, 166-178.
56. Liu, P.; Wen, Y.; Huang, J.; Xiong, A.; Wen, J.; Li, H.; Huang, Y.; Zhu, X.; Ai, S.; Wu, R., A novel strategy of near-infrared spectroscopy dimensionality reduction for discrimination of grades, varieties and origins of green tea. *Vib. Spectrosc.* **2019,** *105*, 102984.
57. Li, B.; Martin, E.; Morris, J., Latent variable selection in partial least squares modelling. *IFAC Proceedings Volumes* **2001,** *34* (25), 463-468.
58. Chtioui, Y.; Panigrahi, S.; Backer, L. F., Self–organizing map combined with a fuzzy clustering for color image segmentation of edible beans. *Transactions of the ASAE* **2003,** *46* (3), 831.
59. Wongravee, K.; Lloyd, G. R.; Silwood, C. J.; Grootveld, M.; Brereton, R. G., Supervised Self Organizing Maps for Classification and Determination of Potentially Discriminatory Variables: Illustrated by Application to Nuclear Magnetic Resonance Metabolomic Profiling. *Analytical Chemistry* **2010,** *82* (2), 628-638.
60. Lloyd, G. R.; Wongravee, K.; Silwood, C. J.; Grootveld, M.; Brereton, R. G.,

Self Organising Maps for variable selection: Application to human saliva analysed by nuclear magnetic resonance spectroscopy to investigate the effect of an oral healthcare product. *Chemometrics and Intelligent Laboratory Systems* **2009,** *98* (2), 149-161.

61. Wongsaipun, S.; Krongchai, C.; Jakmunee, J.; Kittiwachana, S., Rice Grain Freshness Measurement Using Rapid Visco Analyzer and Chemometrics. *Food Analytical Methods* **2018,** *11* (2), 613-623.
62. Wongravee, K.; Ishigaki, M.; Ozaki, Y., Chemometrics as a Green Analytical Tool. In *Challenges in Green Analytical Chemistry*, **2020**, 277-336.
63. Ishigaki, M.; Maeda, Y.; Taketani, A.; Andriana, B. B.; Ishihara, R.; Wongravee, K.; Ozaki, Y.; Sato, H., Diagnosis of early-stage esophageal cancer by Raman spectroscopy and chemometric techniques. *Analyst* **2016,** *141* (3), 1027-1033.
64. Lin, G.-F.; Wang, C.-M., Performing cluster analysis and discrimination analysis of hydrological factors in one step. *Advances in water resources* **2006,** *29* (11), 1573-1585.
65. Siripatrawan, U., Self-organizing algorithm for classification of packaged fresh vegetable potentially contaminated with foodborne pathogens. *Sensors and Actuators B: Chemical* **2008,** *128* (2), 435-441.
66. Meunkaewjinda, A.; Kumsawat, P.; Attakitmongcol, K.; Srikaew, A. In *Grape leaf disease detection from color imagery using hybrid intelligent system*, 2008 5th international conference on electrical engineering/electronics, computer, telecommunications and information technology, *IEEE* **2008**, 513-516.
67. Luna, A. S.; da Silva, A. P.; Alves, E. A.; Rocha, R. B.; Lima, I. C. A.; de Gois, J. S., Evaluation of chemometric methodologies for the classification of Coffea canephora cultivars via FT-NIR spectroscopy and direct sample analysis. *Anal. Methods* **2017,** *9* (29), 4255-4260.
68. Theanjumpol, P.; Wongzeewasakun, K.; Muenmanee, N.; Wongsaipun, S.; Krongchai, C.; Changrue, V.; Boonyakiat, D.; Kittiwachana, S., Non-destructive identification and estimation of granulation in 'Sai Num Pung' tangerine fruit using near infrared spectroscopy and chemometrics. *Postharvest*

Biology and Technology **2019,** *153*, 13-20.

69. Su, W.-H.; He, H.-J.; Sun, D.-W.; nutrition, Non-destructive and rapid evaluation of staple foods quality by using spectroscopic techniques: a review. *Critical reviews in food science* **2017,** *57* (5), 1039-1051.
70. Batten, G. D., An appreciation of the contribution of NIR to agriculture. *Journal of Near Infrared Spectroscopy* **1998,** *6* (1), 105-114.
71. Pasquini, C., Near infrared spectroscopy: fundamentals, practical aspects and analytical applications. *Journal of the Brazilian chemical society* **2003,** *14* (2), 198-219.
72. Qu, J.-H.; Liu, D.; Cheng, J.-H.; Sun, D.-W.; Ma, J.; Pu, H.; Zeng, X.-A., Applications of near-infrared spectroscopy in food safety evaluation and control: A review of recent research advances. *Critical Reviews in Food Science and Nutrition* **2015,** *55* (13), 1939-1954.
73. Kays, S. E.; Barton, F. E.; Windham, W. R., Predicting protein content by near infrared reflectance spectroscopy in diverse cereal food products. *Journal of Near Infrared Spectroscopy* **2000,** *8* (1), 35-43.
74. Fu, X.; Ying, Y.; Lu, H.; Xu, H., Comparison of diffuse reflectance and transmission mode of visible-near infrared spectroscopy for detecting brown heart of pear. *Journal of Food Engineering* **2007,** *83* (3), 317-323.
75. Magwaza, L. S.; Opara, U. L.; Nieuwoudt, H.; Cronje, P. J.; Saeys, W.; Nicolaï, B. NIR spectroscopy applications for internal and external quality analysis of citrus fruit—a review. *Food and Bioprocess Technology* **2012,** *5* (2), 425-444.
76. Nicolaï, B. M.; Beullens, K.; Bobelyn, E.; Peirs, A.; Saeys, W.; Theron, K. I.; Lammertyn, J., Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review. *Postharvest Biology and Technology* **2007,** *46* (2), 99-118.
77. Mahesh, S.; Jayas, D. S.; Paliwal, J.; White, N. D. G., Hyperspectral imaging to classify and monitor quality of agricultural materials. *Journal of Stored Products Research* **2015,** *61*, 17-26.
78. M. ElMasry, G.; Nakauchi, S., Image analysis operations applied to

hyperspectral images for non-invasive sensing of food quality – A comprehensive review. *Biosystems Engineering* **2016,** *142*, 53-82.

79. Leiva-Valenzuela, G. A.; Lu, R.; Aguilera, J. M., Prediction of firmness and soluble solids content of blueberries using hyperspectral reflectance imaging. *Journal of Food Engineering* **2013,** *115* (1), 91-98.
80. Herrero-Langreo, A.; Scannell, A. G. M.; Gowen, A., Chapter 3.5 - Hyperspectral imaging for food-related microbiology applications. In *Data Handling in Science and Technology*, Amigo, J. M., Ed. *Elsevier*, **2020**, *32*, 493-522.
81. Kamruzzaman, M.; ElMasry, G.; Sun, D.-W.; Allen, P., Application of NIR hyperspectral imaging for discrimination of lamb muscles. *Journal of Food Engineering* **2011,** *104* (3), 332-340.
82. Manley, M., Near-infrared spectroscopy and hyperspectral imaging: non-destructive analysis of biological materials. *Chemical Society Reviews* **2014,** *43* (24), 8200-8214.
83. Siripatrawan, U.; Makino, Y., Monitoring fungal growth on brown rice grains using rapid and non-destructive hyperspectral imaging. *International Journal of Food Microbiology* **2015,** *199*, 93-100.
84. Jirakittiwut, N.; Munkongdee, T.; Wongravee, K.; Sripichai, O.; Fucharoen, S.; Praneenararat, T.; Vilaivan, T., Visual genotyping of thalassemia by using pyrrolidinyl peptide nucleic acid probes immobilized on carboxymethylcellulose-modified paper and enzyme-induced pigmentation. *Microchim. Acta* **2020,** *187* (4), 9.
85. Perez, H.; Tah, J. H., Improving the accuracy of convolutional neural networks by identifying and removing outlier images in datasets using t-SNE. *Mathematics* **2020,** *8* (5), 662.
86. Schafer, R. W., What is a Savitzky-Golay filter?[lecture notes]. *IEEE Signal processing magazine* **2011,** *28* (4), 111-117.
87. Zeaiter, M.; Rutledge, D., 3.04 - Preprocessing Methods. In *Comprehensive Chemometrics*, Brown, S. D.; Tauler, R.; Walczak, B., *Eds. Elsevier: Oxford*, **2009**, 121-231.

88. Guo, Q.; Wu, W.; Massart, D., The robust normal variate transform for pattern recognition with near-infrared data. *Analytica chimica acta* **1999,** *382* (1-2), 87-103.

89. Xu, X.; Chen, S.; Xu, Z.; Yu, Y.; Zhang, S.; Dai, R., Exploring Appropriate Preprocessing Techniques for Hyperspectral Soil Organic Matter Content Estimation in Black Soil Area. *Remote Sens.* **2020,** *12* (22), 3765.

90. Zhang, H.-Z.; Zeng, W.; Rutman, M.; Lee, T.-C., Simultaneous determination of moisture, protein and fat in fish meal using near-infrared spectroscopy. *Food Science Technology Research* **2000,** *6* (1), 19-23.

91. Jolliffe, I. T.; Cadima, J., Principal component analysis: a review and recent developments. *Philos Trans A Math Phys Eng Sci* **2016,** *374* (2065), 20150202-20150202.

92. Brereton, R. G., Chemometrics: data analysis for the laboratory and chemical plant. *John Wiley & Sons*, **2003**.

93. Dixon, S. J.; Heinrich, N.; Holmboe, M.; Schaefer, M. L.; Reed, R. R.; Trevejo, J.; Brereton, R. G., Use of cluster separation indices and the influence of outliers: application of two new separation indices, the modified silhouette index and the overlap coefficient to simulated data and mouse urine metabolomic profiles. *Journal of Chemometrics: A Journal of the Chemometrics Society* **2009,** *23* (1), 19-31.

94. Brereton, R. G.; Lloyd, G. R., Partial least squares discriminant analysis: taking the magic away. *Journal of Chemometrics* **2014,** *28* (4), 213-225.

95. Saccenti, E.; Timmerman, M. E., Approaches to Sample Size Determination for Multivariate Data: Applications to PCA and PLS-DA of Omics Data. *Journal of Proteome Research* **2016,** *15* (8), 2379-2393.

96. Morais, C. L.; Lima, K. M.; Singh, M.; Martin, F. L., Tutorial: multivariate classification for vibrational spectroscopy in biological samples. *Nature Protocols* **2020,** *15* (7), 2143-2162.

97. Kohonen, T., *Construction of similarity diagrams for phonemes by a self-organizing algorithm*. Teknillinen korkeakoulu: Helsinki University of Technology, Espoo, 1981.

98. Kohonen, T., Self-organized formation of topologically correct feature maps. *Biological Cybernetics* **1982,** *43* (1), 59-69.

99. Lloyd, G. R.; Brereton, R. G.; Duncan, J. C., Self Organising Maps for distinguishing polymer groups using thermal response curves obtained by dynamic mechanical analysis. *Analyst* **2008,** *133* (8), 1046-1059.

100. Saraswati, A.; Nguyen, V. T.; Hagenbuchner, M.; Tsoi, A. C., High-resolution Self-Organizing Maps for advanced visualization and dimension reduction. *Neural Networks* **2018,** *105*, 166-184.

101. Brereton, R. G., Self organising maps for visualising and modelling. *Chemistry Central Journal* **2012,** *6* (2), 1-15.

102. Baqueta, M. R.; Coqueiro, A.; Março, P. H.; Valderrama, P., Quality control parameters in the roasted coffee industry: a proposal by using microNIR spectroscopy and multivariate calibration. *Food Analytical Methods* **2020,** *13* (1), 50-60.

103. Sun, Y. M.; Wang, Y. J.; Huang, J.; Ren, G. X.; Ning, J. M.; Deng, W. W.; Li, L. Q.; Zhang, Z. Z., Quality assessment of instant green tea using portable NIR spectrometer. *Spectroc. Acta Pt. A-Molec. Biomolec. Spectr.* **2020,** *240*, 118576.

104. Kohonen, T., The self-organizing map. *Proceedings of the IEEE* **1990,** *78* (9), 1464-1480.

105. Wongravee, K.; Heinrich, N.; Holmboe, M.; Schaefer, M. L.; Reed, R. R.; Trevejo, J.; Brereton, R. G., Variable Selection Using Iterative Reformulation of Training Set Models for Discrimination of Samples: Application to Gas Chromatography/Mass Spectrometry of Mouse Urinary Metabolites. *Analytical Chemistry* **2009,** *81* (13), 5204-5217.

106. Alias, N.; Ibrahim, N.; Hamid, A.; Hasbullah, H.; Ali, R. R.; Sadikin, N. A.; Asli, A. U., Thermogravimetric analysis of rice husk and coconut pulp for potential biofuel production by flash pyrolysis. *The Malaysian Journal of Analytical Sciences* **2014,** *18* (3), 705-710.

107. Liu, X.; Jiang, J.; Zhang, H.; Li, M.; Wu, Y.; Guo, L.; Wang, W.; Duan, P.; Zhang, W.; Zhang, Z., Thermal stability and microstructure of metakaolin-based

geopolymer blended with rice husk ash. *Applied Clay Science* **2020,** *196*, 105769

108. Mohamad Ibrahim, M. N.; Ahmed-Haras, M. R.; Sipaut, C. S.; Aboul-Enein, H. Y.; Mohamed, A. A., Preparation and characterization of a newly water soluble lignin graft copolymer from oil palm lignocellulosic waste. *Carbohydrate Polymers* **2010,** *80* (4), 1102-1110.
109. Singh, S. K.; Dhepe, P. L., Isolation of lignin by organosolv process from different varieties of rice husk: Understanding their physical and chemical properties. *Bioresource Technology* **2016,** *221*, 310-317.
110. Ummah, H.; Suriamihardja, D. A.; Selintung, M.; Wahab, A. W., Analysis of chemical composition of rice husk used as absorber plates sea water into clean water. *ARPN Journal of Engineering and Applied Sciences* **2015,** *10* (14), 6046-6050.
111. Yeng, L.; Wahit, M. U.; Othman, N., Thermal and flexural properties of regenerated cellulose(RC)/poly(3- hydroxybutyrate)(PHB)biocomposites. *Jurnal Teknologi* **2015,** *75* (11), 107-112.
112. Lin, L.; He, Y.; Xiao, Z.; Zhao, K.; Dong, T.; Nie, P., Rapid-detection sensor for rice grain moisture based on NIR spectroscopy. *Applied Sciences* **2019,** *9* (8), 1654.
113. López, M. G.; García-González, A. S.; Franco-Robles, E., Carbohydrate Analysis by NIRS-Chemometrics. In *Developments in Near-Infrared Spectroscopy* **2017**, *10*, 67208.
114. Naik, S.; Goud, V. V.; Rout, P. K.; Jacobson, K.; Dalai, A. K., Characterization of Canadian biomass for alternative renewable biofuel. *Renew. Energy* **2010,** *35* (8), 1624-1631.
115. Rambo, M. K. D.; Schmidt, F. L.; Ferreira, M. M. C., Analysis of the lignocellulosic components of biomass residues for biorefinery opportunities. *Talanta* **2015,** *144*, 696-703.
116. Krongtaew, C.; Messner, K.; Ters, T.; Fackler, K., Characterization of key parameters for biotechnological lignocellulose conversion assessed by FT-NIR spectroscopy. Part I: Qualitative analysis of pretreated straw. *BioResources* **2010,** *5* (4), 2063-2080.

117. Zhang, Z.; Yin, X.; Ma, C., Development of simplified models for the nondestructive testing of rice with husk starch content using hyperspectral imaging technology. *Anal. Methods* **2019,** *11* (46), 5910-5918.

118. Abe, H.; Murata, Y.; Kubo, S.; Watanabe, K.; Tanaka, R.; Sulaiman, O.; Hashim, R.; Ramle, S. F. M.; Zhang, C.; Noshiro, S.; Yutaka, M., Estimation of the ratio of vascular bundles to parenchyma tissue in oil palm trunks using NIR spectroscopy. *BioResources* **2013,** *8* (2), 1573-1581.

119. Nakawajana, N.; Posom, J.; Paeoui, J., The prediction of higher heating value, lower heating value and ash content of rice husk using FT-NIR spectroscopy. *Engineering Journal* **2018,** *22* (5), 45-56.

120. Azevedo, C. R.; Boos, C. F.; de Azevedo, F. M. In *Classification of epileptiform events in EEG signals using neural classifier based on SOM*, 2015 International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), *IEEE* **2015,** 1-5.

121. Cuadros-Rodríguez, L.; Valverde-Som, L.; Jiménez-Carvelo, A. M.; Delgado-Aguilar, M., Validation requirements of screening analytical methods based on scenario-specified applicability indicators. *TrAC Trends in Analytical Chemistry* **2020,** *122*, 115705

122. Li, X.; He, Y.; Wu, C., Non-destructive discrimination of paddy seeds of different storage age based on Vis/NIR spectroscopy. *Journal of Stored Products Research* **2008,** *44* (3), 264-268.

123. Aznan, A.; Rukunudin, I.; Shakaff, A.; Ruslan, R.; Zakaria, A.; Saad, F., The use of machine vision technique to classify cultivated rice seed variety and weedy rice seed variants for the seed industry. *International Food Research Journal* **2016,** *23*.

124. Feng, X.; Peng, C.; Chen, Y.; Liu, X.; Feng, X.; He, Y., Discrimination of CRISPR/Cas9-induced mutants of rice seeds using near-infrared hyperspectral imaging. *Scientific reports* **2017,** *7* (1), 1-10.

125. Cui, Y.; Xu, L.; An, D.; Liu, Z.; Gu, J.; Li, S.; Zhang, X.; Zhu, D., Identification of maize seed varieties based on near infrared reflectance spectroscopy and chemometrics. *International Journal of Agricultural and*

Biological Engineering **2018,** *11* (2), 177-183.

126. Ruslan, R.; Bejo, S.; Rukunuddin, I.; Aznan, A. In *Selection of Morphological Features in Classifying Weedy Rice and Rice Seed Varieties using Discriminant Function Analysis*, IOP Conference Series: Materials Science and Engineering, IOP Publishing: 2019; p 012014.
127. Xiao, R.; Liu, L.; Zhang, D. J.; Ma, Y.; Ngadi, M. O., Discrimination of organic and conventional rice by chemometric analysis of NIR spectra: a pilot study. *Journal of Food Measurement and Characterization* **2019,** *13* (1), 238-249.
128. Feng, L.; Zhu, S.; Liu, F.; He, Y.; Bao, Y.; Zhang, C., Hyperspectral imaging for seed quality and safety inspection: A review. *Plant Methods* **2019,** *15* (1), 1-25.
129. Liu, Z.-y.; Cheng, F.; Ying, Y.-b.; Rao, X.-q., Identification of rice seed varieties using neural network. *Journal of Zhejiang University. Science. B* **2005,** *6* (11), 1095.
130. Singh, K. R.; Chaudhury, S., Efficient technique for rice grain classification using back-propagation neural network and wavelet decomposition. *IET Computer Vision* **2016,** *10* (8), 780-787.
131. Huang, K.-Y.; Chien, M.-C., A novel method of identifying paddy seed varieties. *Sensors* **2017,** *17* (4), 809.
132. Sun, J.; Lu, X.; Mao, H.; Jin, X.; Wu, X., A method for rapid identification of rice origin by hyperspectral imaging technology. *Journal of Food Process Engineering* **2017,** *40* (1), e12297.
133. Vu, H.; Duong, V. N.; Nguyen, T. T. In *Inspecting rice seed species purity on a large dataset using geometrical and morphological features*, Proceedings of the Ninth International Symposium on Information and Communication Technology, **2018**, 321-328.
134. Wu, N.; Jiang, H.; Bao, Y.; Zhang, C.; Zhang, J.; Song, W.; Zhao, Y.; Mi, C.; He, Y.; Liu, F., Practicability investigation of using near-infrared hyperspectral imaging to detect rice kernels infected with rice false smut in different conditions. *Sensors and Actuators B: Chemical* **2020,** *308*, 127696.

135. Amanah, H. Z.; Wakholi, C.; Perez, M.; Faqeerzada, M. A.; Tunny, S. S.; Masithoh, R. E.; Choung, M.-G.; Kim, K.-H.; Lee, W.-H.; Cho, B.-K., Near-Infrared Hyperspectral Imaging (NIR-HSI) for Nondestructive Prediction of Anthocyanins Content in Black Rice Seeds. *Applied Sciences* **2021,** *11* (11).

136. Kim, M. S.; Chen, Y.; Mehl, P.M., Hyperspectral reflectance and fluorescence imaging system for food quality and safety. *Transactions of the ASAE* **2001,** *44* (3), 721.

137. Bouveresse, E.; Massart, D. L., Improvement of the piecewise direct standardisation procedure for the transfer of NIR spectra for multivariate calibration. *Chemometrics and Intelligent Laboratory Systems* **1996,** *32* (2), 201-213.

138. Zhu, J.; Agyekum, A. A.; Kutsanedzie, F. Y. H.; Li, H.; Chen, Q.; Ouyang, Q.; Jiang, H., Qualitative and quantitative analysis of chlorpyrifos residues in tea by surface-enhanced Raman spectroscopy (SERS) combined with chemometric models. *LWT* **2018,** *97*, 760-769.

139. Zhang, L.; Rao, Z. H.; Ji, H. Y., Hyperspectral imaging technology combined with multivariate data analysis to identify heat-damaged rice seeds. *Spectr. Lett.* **2020,** *53* (3), 207-221.

140. Doshi, R. A.; King, R. L.; Lawrence, G. W., Classification of Rotylenchulus reniformis Numbers in Cotton Using Remotely Sensed Hyperspectral Data on Self-Organizing Maps. *J. Nematol.* **2010,** *42* (3), 179-193.

141. Kohonen, T., Analysis of a simple self-organizing process. *Biological cybernetics* **1982,** *44* (2), 135-140.

142. Kittiwachana, S.; Wangkarn, S.; Grudpan, K.; Brereton, R. G., Prediction of liquid chromatographic retention behavior based on quantum chemical parameters using supervised self organizing maps. *Talanta* **2013,** *106*, 229-236.

143. Sim, S. F.; Sági-Kiss, V., Multiple Self Organising Maps (mSOMs) for simultaneous classification and prediction: Illustrated by spoilage in apples using volatile organic profiles. *Chemometrics and Intelligent Laboratory Systems* **2011,** *109* (1), 57-64.

144. Céréghino, R.; Park, Y. S., Review of the Self-Organizing Map (SOM) approach

in water resources: Commentary. *Environmental Modelling & Software* **2009,** *24* (8), 945-947.

145. Wankhede, S. B., Analytical study of neural network techniques: SOM, MLP and classifier-a survey. *IOSR Journal of Computer Engineering* **2014,** *16* (3), 86-92.

146. Weng, S.; Tang, P.; Yuan, H.; Guo, B.; Yu, S.; Huang, L.; Xu, C., Hyperspectral imaging for accurate determination of rice variety using a deep learning network with multi-feature fusion. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **2020,** *234*, 118237.

147. Olawoyin, R.; Nieto, A.; Grayson, R. L.; Hardisty, F.; Oyewole, S., Application of artificial neural network (ANN)–self-organizing map (SOM) for the categorization of water, soil and sediment quality in petrochemical regions. *Expert Systems with Applications* **2013,** *40* (9), 3634-3648.

148. Lee, B.-H.; Scholz, M. J. W., Air,; Pollution, S., A comparative study: Prediction of constructed treatment wetland performance with k-nearest neighbors and neural networks. **2006,** *174* (1), 279-301.

149. Metz, C. E. In *Basic principles of ROC analysis*, Seminars in nuclear medicine, *Elsevier*, **1978**, 283-298.

150. Zhang, X. R.; Pan, Z. B.; Hu, B. L.; Zheng, X.; Liu, W. H., Target detection of hyperspectral image based on spectral saliency. *IET Image Processing* **2019,** *13* (2), 316-322.

151. Barnabei, L.; Marazia, S.; De Caterina, R., Receiver operating characteristic (ROC) curves and the definition of threshold levels to diagnose coronary artery disease on electrocardiographic stress testing. Part I: The use of ROC curves in diagnostic medicine and electrocardiographic markers of ischaemia. *Journal of Cardiovascular Medicine* **2007,** *8* (11), 873-881.

152. Sidike, P.; Asari, V. K.; Alam, M. S., Multiclass Object Detection With Single Query in Hyperspectral Imagery Using Class-Associative Spectral Fringe-Adjusted Joint Transform Correlation. *IEEE Transactions on Geoscience and Remote Sensing* **2016,** *54* (2), 1196-1208.

153. Shen, S.; Zhang, H.; Huang, K.; Chen, H.; Shen, W.; Fang, X., Differentiation

of cultivation areas and crop years of milled rice using single grain mass spectrometry. *New Journal of Chemistry* **2019,** *43* (5), 2118-2125.

VITA

NAME	Sureerat Makmuang
DATE OF BIRTH	16 September 1993
PLACE OF BIRTH	Phitsanulok
INSTITUTIONS ATTENDED	Chulalongkorn University
HOME ADDRESS	223/2 Moo 10 Phrompiram, Phrompiram district, Phitsanulok 65150
PUBLICATION	Makmuang, S., Nootchanat, S., Ekgasit, S., & Wongravee, K., 2021. Non-destructive method for discrimination of weedy rice using near infrared spectroscopy and modified self organizing maps (SOMs). Computers and Electronics in Agriculture, 191, pp.106522
	Makmuang, S., Ekgasit, S., & Wongravee, K., 2020. Non-invasive discrimination of milled rice grains by using near infrared spectroscopy combined with chemometrics. Proceeding of pure and Applied Chemistry International Conference 2020 (PACCON2020). “Chemistry for Catalyzing Sustainability and Prosperity” Impact Forum, Impact Muang Thong Thani, Nonthaburi, Thailand.
	Kikuchi, M., Makmuang, S., Izawa, S., Wongravee, K. and Hiramoto, M., 2019. Doped organic single-crystal photovoltaic cells. Organic Electronics, 64, pp.92-96.
	Makmuang, S., Sricharoen, N., Pienpinijtham, P., Ekgasit, S., & Wongravee, K., 2018. Global calibration model for determination of glucose in non-alcoholic beverages using near infrared spectroscopy combined with chemometrics. Proceeding of Pure and Applied Chemistry International Conference 2018 (PACCON 2018). The 60th Anniversary of His Majesty the King’s Accession to the Throne International Convention Center, Hat Yai, Songkhla, Thailand, pp.AN22-AN27.
	Chainok, K., Makmuang, S. and Kielar, F., 2016. Crystal structures of (E)-N′-(2-hydroxy-5-methylbenzylidene) isonicotinohydrazide and (E)-N′-(5-fluoro-2-hydroxybenzylidene) isonicotinohydrazide. Acta Crystallographica Section E: Crystallographic Communications, 72(7), pp.980-983.