Discrimination of Thai Melon Seeds Using Near-Infrared Spectroscopy and Adaptive Self-Organizing Maps

Sureerat Makmuang,[1] Tirayut Vilaivan,[2] Simon Maher,[3] Sanong Ekgasit,[1] and Kanet Wongravee[1] *

[1] Sensor Research Unit (SRU), Department of Chemistry, Faculty of Science, Chulalongkorn University, Bangkok, Thailand, 10330

[2] Organic Synthesis Research Unit, Department of Chemistry, Faculty of Science, Chulalongkorn University, Bangkok 10330, Thailand

[3] Department of Electrical Engineering & Electronics, University of Liverpool, Brownlow Hill, Liverpool L69 3GJ, UK

*Corresponding author : kanet.w@chula.ac.th, kanet.wongravee@gmail.com

Abstract

Melon (*Cucumis melo* L.) is a popular fruit consumed around the world. It has significant economic value as a crop, export product, and source of essential nutrients. Thus, using high-quality, authentic seed varieties is the first step toward achieving impactful agricultural production. Unfortunately, distinguishing between seed varieties using only human perception can be difficult because of their similar traits. Thus, dishonest distributors may trade low-quality seeds for high-quality seeds. In this study, seeds from five Thai melon varieties, Singapore Thai melon (ST), Nan Thai melon (NT), Round Thai melon (RT), Striped Singapore Thai melon (SST), and Golden and Long Thai melon (GLT), were classified using a distinctive discrimination method that combines modified self-organizing maps (SOMs) with near-infrared (NIR) spectroscopy. The physical characteristics, morphology, and thermal behavior of the seeds were also examined through optical microscopy, scanning electron microscopy, and thermogravimetric analysis, respectively. Attenuated total reflection–Fourier transform infrared, and NIR spectroscopy revealed that different varieties of melon seeds possess significant variations in lignin content and carbohydrate composition. Seed samples from the five Thai melon varieties were further classified using a modified SOM map created with optimized scaling value, map size, and a number of iteration parameters. Binary classification with the One vs Rest strategy and multiclass classification was performed to verify the constructed classifier model. The supervised SOMs developed herein can achieve the multiclassification of seed types effectively and efficiently, with a high accuracy of 95.52% for the training set and 91.59% for the test set, which were significantly superior to those of well-established discrimination models.

Key words: Near-infrared spectroscopy, Self-organizing maps, Chemometrics, Machine Learning, Multiclassification

1. Introduction

Muskmelon (*Cucumis melo* L.), or simply “melon,” is one of the world’s most important commercial fruit crops, with 1.3 million hectares of harvest area and 31 million tons in annual demand worldwide [1]. There is a large variety of melons, including netted varieties such as cantaloupes (*C. melo* Reticulatus Group) and smooth-skinned varieties such as honeydew melons (*C. melo* Inodorus Group). In addition to the melon’s richness in minerals and their health-promoting components [2], sweetness, flavor/aroma, texture, and phytonutrient contents, including potassium, vitamin C, and provitamin A (beta-carotene), have a significant impact on consumer purchasing decisions [3]. Thus, using high-quality melon seeds is among several essential factors for the production of high-quality crops with desirable product quality characteristics [4]. In addition, the issues of seed quality are important from other perspectives, such as agricultural output, quarantine processes, and local and worldwide seed mobility for economic and commercial considerations.

Seeds are obtained from certified agencies to ensure high quality. However, countries with underdeveloped agroeconomics may lack the necessary infrastructure, technology, and institutions to support agricultural development. Thus, farmers may retain historic cultivars using seeds from family, neighbors, or the local market. These informal seed supply systems have been referred to as “seed exchange networks,” “farmer seed systems,” “traditional seed systems,” and “informal seed systems” [5].

In Thailand, melon is one of the most costly fruits because of the difficulty of its cultivation [6]. Thai melon varieties vary greatly in flavor and price ranges, and the seeds of popular varieties that are in high demand are likely to be more expensive. According to the information obtained from seed exchange networks, which comprise the majority of Thailand’s agricultural communities, high commercial value seeds are often adulterated with low-quality and cheaper seeds. Because of the similar physical characteristics of seeds from different varieties, differentiating them through visual observation alone is virtually impossible. As it takes 3–4 months before the melon plants are fully grown and start to produce fruits, growing the wrong seeds means wasting time and resources. Thus, appropriate management methods are required to maintain and regulate seed quality to prevent the detrimental impacts of seed adulteration.

In recent years, various strategies have been employed to protect the interests of importing nations and consumers through explicit cultivar discrimination, exact adulterant measurement, and the identification of geographic cultivation areas [7, 8].

For instance, seed morphology analysis relies on physical methods for seed inspection to examine the macroscopic and microscopic characteristics of seeds and other seed features, such as solubility, bulk density, and texture [9]. However, despite its simple measurement and operation, this approach has substantial limitations, such as its subjectivity and phenological variance, which require expert interpretation. Meanwhile, more accurate and sensitive biotechnological methods, including polymerase chain reaction, probe hybridization, and sequencing, are also widely used [10]. However, these methods also suffer from limitations due to instruments and reagents costs and technical skill requirements. In comparison, chemical methods based on chromatographic techniques, such as gas chromatography or high-performance liquid chromatography, offer high performance in detecting seed adulteration with good reliability. However, they also have limitations in terms of their relatively high cost, complexity, and time consumption [7]. Table S1 summarizes the different methods available for agricultural product evaluation. Rapid, preferably real-time, effective, and affordable detection approaches are thus highly desirable for quality control and rapid adulteration detection in processed agricultural goods.

Among various existing techniques, the near-infrared (NIR) method provides several advantages, including its nondestructive nature, allowing the reuse of samples for further investigation. It also requires minimal sample preparation, offers rapid detection and applies to various sample types. Thus, it has been widely utilized as a noninvasive analytical method for various purposes, including controlling processes, undertaking qualitative and quantitative examinations, and detecting food product adulteration [11, 12]. The use of NIR spectroscopy in seed quality evaluation [13], seed adulteration [14], and seed purity analysis [15] is well known. However, because of the contribution of several factors, such as the physical state of the sample and testing environment, which can affect the quality of the spectra, discerning “relevant” information regarding the properties of target analytes from raw spectral data is incredibly challenging [16]. To solve this problem, mathematical and statistical techniques are required to extract relevant information (i.e., spectrum features related to the analyte's properties) from other irrelevant data (i.e., interfering parameters) [17].

Chemometrics is a well-known chemical discipline that uses mathematics, statistics and formal logic for extracting meaningful and important qualitative or quantitative information from large and complex data sets [16]. NIR spectroscopy, along with the polar qualifying system (PQS), can discriminate between melon genotypes and hybrids [18]. PQS, using automatic wavelength range optimization, has successfully differentiated various horticultural plant seeds, including melon (*C. melo*), watermelon (*Citrullus lanatus*), pepper (*Capsicum annuum*), and *Mathiola incana* varieties, as well as a watermelon hybrid with its parent lines. From measurements of a single seed to large volumes of samples, the NIR method with a hyperspectral approach has been proven valuable for differentiating and identifying different agricultural plants [19]. For instance, NIR hyperspectral imaging, using statistical models like partial least square discriminant analysis (PLS-DA) and least square support vector machines, distinguished between virus-infected and healthy watermelon seeds with 83.3% accuracy [20]. A discriminant PLS-DA model was also used to distinguish between viable and nonviable triploid watermelon seeds based on Fourier transform NIR spectroscopy (FT-NIR) data with a high level of classification accuracy for both viable (87.7%) and nonviable (82%) seeds [21]. More information on the NIR approach integrated with well-established chemometrics in seed quality assessment is shown in Table S2.

Although several studies focused on the categorization of agricultural seeds using the combination of NIR and chemometrics as mentioned above, the use of the NIR approach in conjunction with self-organizing maps (SOMs) to distinguish various classes of seeds has not been reported. As the first step toward ensuring high-quality seed production, this work reports the first successful discrimination of melon seeds from five different varieties grown in Thailand using NIR in conjunction with modified SOMs. The melon varieties in this study included the Singapore Thai melon (ST), Nan Thai melon (NT), Round Thai melon (RT), Stiped Singapore Thai melon (SST), and Golden and Long Thai melon (GLT). The surface topography and other physical/physicochemical properties of the seeds were investigated using optical microscopy, scanning electron microscopy (SEM), thermogravimetric analysis (TGA), and attenuated total reflection–Fourier transform infrared (ATR-FTIR) spectroscopy. These supplementary and validated techniques were employed as there has been no previous research on the application of NIR for the classification of Thai melon seeds. Next, a supervised self-organizing map (SOM) classifier was developed and optimized to accurately classify various kinds of Thai melon seeds based on the data collected from the NIR spectra of the seeds according to the conceptual framework proposed in Fig. 1. The adaptive SOMs could be used for both binary and multiclass classification and enabled the detection and comprehension of nonlinear data relationships which exceeded the capabilities of conventional linear-based chemometric techniques. The great performance and nondestructive nature of this technique, its ability to perform multiclassification, which overcomes the limitations of the current dichotomous system, and its potential economic scale-up should make it easily accessible to agro-dealers and farmers in various disciplines.

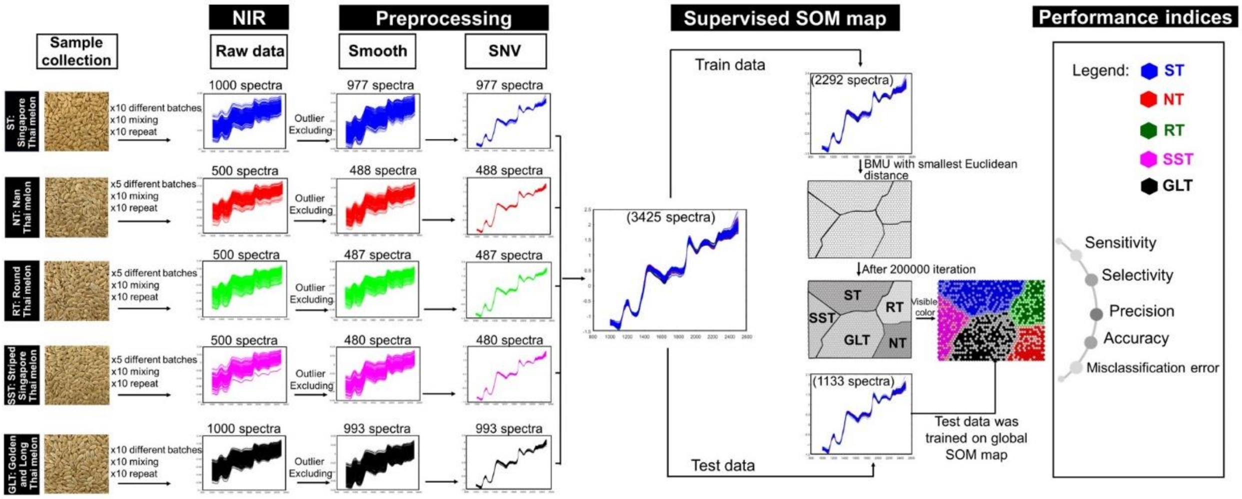


Fig. 1 The proposed multiclassification approach based on supervised self-organizing maps (SOMs) to distinguish five Thai melon seeds directly.

2. Materials and methods

2.1. Sample collection and preparation

The seeds of five distinct Thai melon cultivars were collected from various trusted sources in Thailand. ST seeds were collected from honest and credible local vendors (Phatum-thani and Phitsanulok

Provinces, Thailand). NT seeds were collected from reliable suppliers in Phatum-thani Province. The other Thai melon seeds (RT, SST, and GLT) were collected from trusted vendors (Bangkok and Nonthaburi provinces, Thailand). All seeds were collected and tested between January and June of 2022. During the trial, these seeds were roughly 6 and 10 months of age. The information on all samples (common name, source, harvest date, and production date) is summarized in Table 1.

Table 1 Information on the collected Thai melon seeds from various local markets and distributors in Thailand

Variety of Thai melon	Abbreviation	Source	Harvest date	Collection date	NIR acquisition	Number of data points (spectra)
Singapore Thai melon	ST	Phatum-thani and Phitsanulok	Feb 10, 2022	Mar 15, 2022	May 29, 2022	1,000
Nan Thai melon	NT	Phatum-thani	Jan 5, 2022	Jan 15, 2022	May 14, 2022	500
Round Thai melon	RT	Bangkok	Jan 1, 2022	Mar 1, 2022	May 17, 2022	500
Striped Singapore Thai melon	SST	Nontaburi	Aug 1, 2021	Jan 1, 2022	May 23, 2022	500
Golden and Long Thai melon	GLT	Bangkok	Jan 1, 2022	Feb 2, 2022	May 25, 2022	1,000

2.2 NIR Spectral acquisition

The NIR spectra of the seed samples were collected on a Thermo Scientific™ Nicolet™ iS5N FT-NIR spectrometer with an extended range indium gallium arsenide detector, high-intensity halogen light source, and temperature-stabilized solid-state NIR diode laser. Each type of melon seed sample was randomly dispersed into the quartz cup holder to ensure that all variances in the obtained spectra were collected. Fig. S1 displays the details of the data collection process. The samples were placed at identical distances from the probe, and their surfaces were flattened before the measurement to eliminate undesirable interference from scattering effects. During the spectrum sample collection, the sample holder was covered by a black box to eliminate interferences from external light. The NIR spectra of the samples were acquired over the range of 1,000–2,500 nm in the reflection mode, and the average data obtained from 32 scans were recorded. Throughout the experiment, the temperature was maintained between 27°C and 29°C.

2.3. Data analysis

2.3.1 Pre-processing algorithm of NIR spectra

In the initial data pre-processing stage, the interquartile range (IQR), which demonstrates the difference between the 75th and the 25th percentiles [22], was used to identify data points that deviated significantly from the norm or outliers. The average NIR spectrum of each sample class was calculated as a centroid of the data class. The Euclidean distance of the NIR spectra of samples within the same class was subsequently determined. Outliers were defined as samples with a Euclidean distance greater than 1.5IQR from the mean in-class NIR spectra and were thus removed, representing approximately 2% of the total data in this case. Then, the spectra were processed using Savitsky–Golay smoothing filter followed by an additional mathematical pre-processing algorithm based on standard normal variate (SNV) to compensate for the surface scattering of light, uneven sample particle size, and optical path fluctuation on the NIR spectra [23].

2.3.2 Adaptive SOMs for the discrimination approach

SOMs are unsupervised learning models whose architecture consists of a two-dimensional grid of neurons with interconnected multidimensional functions. The two fundamental steps in constructing SOMs and their algorithm are the learning of multidimensional space projection onto a two-dimensional map and the subsequent selection of the best matching unit (BMU) [24].

Step 1: An initial SOM map is generated with $M \times N = K$ units whereby each unit contains a weight vector $\boldsymbol{v}_k$ randomly generated from a uniform distribution between the maximum and minimum intensities in the dataset [23]. In this study, the size of the SOM map was carefully considered to cover most of the samples to be matched.

Step 2: In a supervised SOM model, this can be expanded for supervised learning by adding an extra set of variables denoting class labels to the input variables before the training process. For each random selection, a vector is generated; for instance, if a sample belongs to the third class out of five, the extra variables are $\boldsymbol{w}_k = [0,0,\omega,0,0]$, where ω is the scaling factor. The value of ω is used to determine if the sample belongs to the given class; a value of 0 demonstrates that the sample does not belong to the class. Here, the vector was randomly generated and added to the vector $\boldsymbol{v}_k$ in step 1 for each unit, given as $\boldsymbol{v}_u$ = [$\boldsymbol{v}_k$ $\boldsymbol{w}_k$]. In the study, the scaling value was carefully optimized because it determines the degree to which class membership affects the map; if the value is too large, the map may overfit the data; if the value is too small, the map could transition into an unsupervised state. This means that classes may not always be fully separated, which may contribute to inaccurate statistical analysis [24].

Step 3: The sample vector $\boldsymbol{x}_s$ with the supervised vector $\boldsymbol{w}_k$ resulting in $\boldsymbol{x}_{sk}$ = [$\boldsymbol{x}_s$ $\boldsymbol{w}_k$] in the dataset is then compared with the weight vector of each unit ($\boldsymbol{v}_u$) on the initial SOM map from step 2. The Euclidian distance between $\boldsymbol{x}_{sk}$ and $\boldsymbol{v}_u$ of each map unit k is calculated as follows (Eq. (1)):

$$d_{sk,u} = \sqrt{(\boldsymbol{x}_{sk} - \boldsymbol{v}_u)(\boldsymbol{x}_{sk} - \boldsymbol{v}_u)^{\mathrm{T}}} \tag{1}$$

This process is repeated until the distance of $\boldsymbol{x}_{sk}$ and K units on the map is calculated.

Step 4 : The map unit with the shortest Euclidean distance is announced as the BMU of the chosen sample weight vector $\boldsymbol{x}_{sk}$: BMU = $\overset{\min}{k}\{d_{sk,u}\}$ [25].

Step 5 : The training process is started for the BMU and the neighboring map units (N_{u}) within the length from the BMU. They are updated to become more similar to the sample weight vector $\boldsymbol{x}_{sk}$. The learning rate in each iteration is calculated:

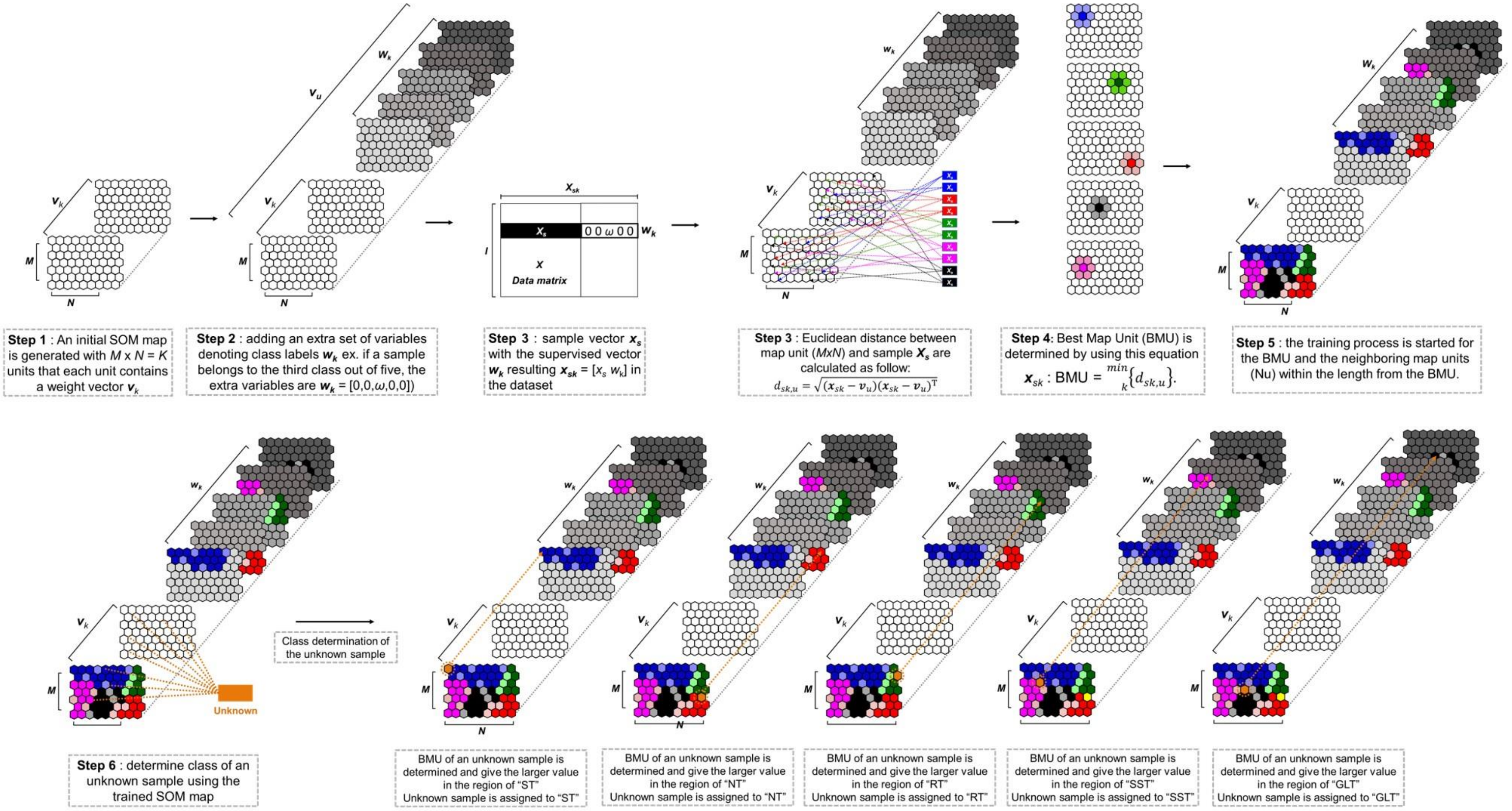
$$\boldsymbol{v}_u = \begin{cases} \boldsymbol{v}_u + \gamma\alpha(\boldsymbol{x}_{sk} - \boldsymbol{v}_u) \;\; k \in N_u \\ \boldsymbol{v}_u \;\; k \notin N_u \end{cases} \tag{2}$$

where α indicates the learning rate and γ is the neighborhood learning weight. Note that the amount of learning decreases with each iteration of the training process, as does the neighborhood learning rate with the distance from the BMU [26]. The number of iterations utilized in the SOM training process should exceed the number of map units (K) to guarantee that the map provides an adequate unit to learn from each sample. The clusters of samples are shown graphically using color map shading.

Step 6: After the reference samples, referred to as the training set, have been trained, the SOM map can then be obtained in step 5. For applying the trained SOM map to identify the class of an unknown sample, the BMU of the unknown sample is searched and allocated to the SOM unit with the shortest

Euclidean distance. The class of the unknown sample is allocated to the class with the greatest value in the part of the class weight vector (w_k); for instance, if the class vector of the BMU is [2 2.5 2.7 2.3 1.9], the class of the unknown is ascribed to the third class (with the highest value of 2.7) [24].

Other adaptable parameters, including the map size and the number of iterations, should be optimized to push the original SOM algorithm to deal with the specific applications at hand (i.e., Thai melon seeds in this case). In this work, we developed our software for the supervised SOMs in MATLAB (early findings have been described elsewhere [23]), enabling the creation of innovative approaches combined with hyperspectral imaging methodology for the multiclassification of Thai melon seeds. Fig. 2 depicts the adaptive supervised SOM conceptual model along with the details of the supervised SOM algorithm.



Step 1 : An initial SOM map is generated with M x N = K units that each unit contains a weight vector $\boldsymbol{v}_k$

Step 2 : adding an extra set of variables denoting class labels $\boldsymbol{w}_k$ ex. if a sample belongs to the third class out of five, the extra variables are $\boldsymbol{w}_k$ = [0,0,ω,0,0])

Step 3 : sample vector $\boldsymbol{x}_s$ with the supervised vector $\boldsymbol{w}_k$ resulting $\boldsymbol{x}_{sk}$ = [x_s w_k] in the dataset

Step 3 : Euclidean distance between map unit (*MxN*) and sample $\boldsymbol{X}_s$ are calculated as follow:

$$d_{sk,u} = \sqrt{(\boldsymbol{x}_{sk} - \boldsymbol{v}_u)(\boldsymbol{x}_{sk} - \boldsymbol{v}_u)^{\mathrm{T}}}$$

Step 4: Best Map Unit (BMU) is determined by using this equation

$$\boldsymbol{x}_{sk} : \text{BMU} = \min_{k}\{d_{sk,u}\}.$$

Step 5 : the training process is started for the BMU and the neighboring map units (Nu) within the length from the BMU.

Step 6 : determine class of an unknown sample using the trained SOM map

Class determination of the unknown sample

BMU of an unknown sample is determined and give the larger value in the region of "ST"
Unknown sample is assigned to "ST"

BMU of an unknown sample is determined and give the larger value in the region of "NT"
Unknown sample is assigned to "NT"

BMU of an unknown sample is determined and give the larger value in the region of "RT"
Unknown sample is assigned to "RT"

BMU of an unknown sample is determined and give the larger value in the region of "SST"
Unknown sample is assigned to "SST"

BMU of an unknown sample is determined and give the larger value in the region of "GLT"
Unknown sample is assigned to "GLT"

Fig. 2 Conceptual diagram for the multiclassification of Thai melon seeds using adaptive supervised self-organizing maps (SOMs) for K classes with a two-dimensional SOM map in the M × N dimension. The adaptive supervised SOMs can be implemented in two scenarios: training operation of a supervised SOM map to be used as a reference map for multiclass classification (Steps 1–5) and unknown class identification by mapping the unknown to the reference SOM map (Step 6).

2.3.3 Model validation

The discrimination performance was validated by dividing the whole dataset into training and test sets. Two-thirds of the samples in each class were assigned as the training set for developing the classifier model, whereas the remaining one-third were used as the test set for model validation. To ensure the model's robustness, the procedure was repeated 10 times. The performance of the classifier model was evaluated using the percentage of correctly classified (%CC) seeds (Eq. (3)), where the class of samples predicted by the generated model exactly matched the actual class.

$$\%\text{CC} = \frac{N_\text{p}}{N_\text{t}} \times 100 \tag{3}$$

where N_p and N_t are the numbers of correctly classified samples and the total number of samples, respectively. %CC was mainly used to evaluate the multiclass classification.

For evaluating the binary classification approach, the One vs Rest strategy was used by assigning the "in-class" as a positive identification and the remaining mixed class of melon seeds as the "out-class," which denoted a negative identification. For example, Case I has in-class members that are ST melon seeds and out-class members comprising the rest (mixed seeds). Case II holds in-class members of NT seeds against the rest of the seeds and vice versa, resulting in a total of five cases (Cases I–V). The five indicators involving sensitivity, specificity, precision, accuracy, and misclassification (ME) were used to assess the model performance [23].

$$\text{Sensitivity} = TP \,/\, (TP + FN) \tag{4}$$

$$\text{Precision} = TP \,/\, (TP + FP) \tag{5}$$

$$\text{Specificity} = TN \,/\, (TN + FP) \tag{6}$$

$$\text{Accuracy} = (TP + TN) \,/\, (TP + FP + TN + FN) \tag{7}$$

$$\text{ME} = (FP + FN) \,/\, (TP + FP + TN + FN) \tag{8}$$

Where TP is "true positive", indicating the number of correctly classified positive case; FP is "false positives", denoting the number of negative cases that were classified as positive; TN is "true negatives", representing the number of correctly classified negative cases; and FN is "false negative", representing the number of positive cases classified as negative. From these assigned indices, the classification performances, including sensitivity, specificity, precision, accuracy, and misclassification error (ME). Generally, a good classifier model is expected to exhibit high sensitivity and accuracy. All discrimination approaches are developed based on hard modelling as all seed samples will be categorized into one of the Thai melon varieties, without any seed samples remaining unclassified or defined as outliers [26-30].

2.4 ATR-FTIR

Attenuated Total Reflectance-Fourier Transform Infrared (ATR-FTIR) spectroscopy was employed to determine the IR spectral characteristics of the melon seed samples. The IR spectra in the functional group region (500–4,000 cm^{-1}) were recorded using a Nicolet™ iS™ 5 FTIR spectrometer (Thermo Fisher Scientific, USA) with a Diamond ATR at a resolution of 0.4 cm^{-1}.

2.5 Thermogravimetric analysis (TGA)

Thermogravimetric analysis (TGA) is a thermal analysis technique that measures changes in sample weight as a function of temperature. In the present study, it was employed to examine thermal stability and decomposition of chemical substitutes of Thai melon seeds. The TGA curves obtained provide valuable information which facilitates the evaluation of seed quality, determination of shelf life, and identification of potential contaminants or adulterants. Thermogravimetric experiments were conducted to illustrate the thermophysical properties of the samples using a Perkin Elmer Pyris1 TGA system. The system was operated under inert conditions with a steady nitrogen flow of 20 mL min^{-1}. Each type of melon sample was crushed into small pieces, and around 3–15 mg was pyrolyzed. The samples were first isothermally heated at 35°C for 1 min to keep the initial environment identical for all samples to remove the adsorbed water and moisture on the sample. Next, the samples were continuously heated from 50°C to 800°C at a heating rate of 20°C min^{-1}.

2.6 Scanning electron microscopy (SEM)

The morphology of the Thai melon seed samples was examined using SEM technique" change to "Scanning electron microscopy (SEM) is a highly effective tool for investigating the microstructure and surface morphology of materials. The present study utilized SEM to examine the surface characteristics of Thai melon seeds, thereby providing significant insights into the seed composition and structure, which enabled in the differentiation of distinct variety of Thai melon seeds. The samples were fixed on carbon tape and attached to an aluminum stub. The samples for SEM were vacuum-dried for 1 h before imaging. The SEM micrographs of the samples were acquired using a scanning electron microscope (JEOL JSM-6510) operated at 2–15 kV under a high vacuum mode of 6.7 × 10^{-2} Pa.

3. Results and discussion

3.1. Physical characteristics of Thai melon seeds

Photographs of the Thai melon seeds were taken using a digital camera to observe their morphologies, as presented in Fig. 3A1–3A5. All varieties of melon seeds showed similar morphologies in terms of their shape and color. Therefore, differentiating the seeds through merely visual inspection is difficult. Thus, the seeds were further examined using an optical microscope (AxioVision Viewer 4.8) with a high magnification optical microscope image of 100×, as demonstrated in Fig. 3B1–3B5. Again, even at such a microscopic level, no noticeable differences were observed regarding the physical features on the seed surfaces. After vacuum drying for 1 h, the surface topographical characteristics of the seeds were examined using SEM, as shown in Fig. 3C1–3C5. The SEM images revealed spherical particles of hemicellulose and lignin buried in the cellulose matrix, which was the main component of the melon seed cell wall [31]. As evident in Fig. 3C1–3C5, the outer surface of seed husks from different melon varieties exhibited remarkably distinctive patterns. ST and SST shared a similar endocarp pattern consisting of fiber lines; however, that of ST was more uniform and ordered. Meanwhile, NT and RT showed the same systematic square-shaped contours. GLT exhibited a combination of a linear pattern and a square contour on the surface endocarp. Minor differences in the surface morphology of the seeds might be associated with the varieties as well as other reasons, such as environmental circumstances (e.g., climate, temperature, light, soil kinds, and qualities) [6]. Although significant differences existed at such an extreme magnification image (100×), it was concluded at this point that the visual observation of seed morphologies could not provide sufficient input data for multiclassification purposes. The thermal degradation behaviors of the biomass from Thai melon seeds were assessed through TGA and derivative thermogravimetry (DTG) curves. Additionally, the chemical structure and functional properties were investigated using Attenuated Total Reflectance-Fourier Transform Infrared (ATR-FTIR) characterization, as illustrated in Fig. S2.

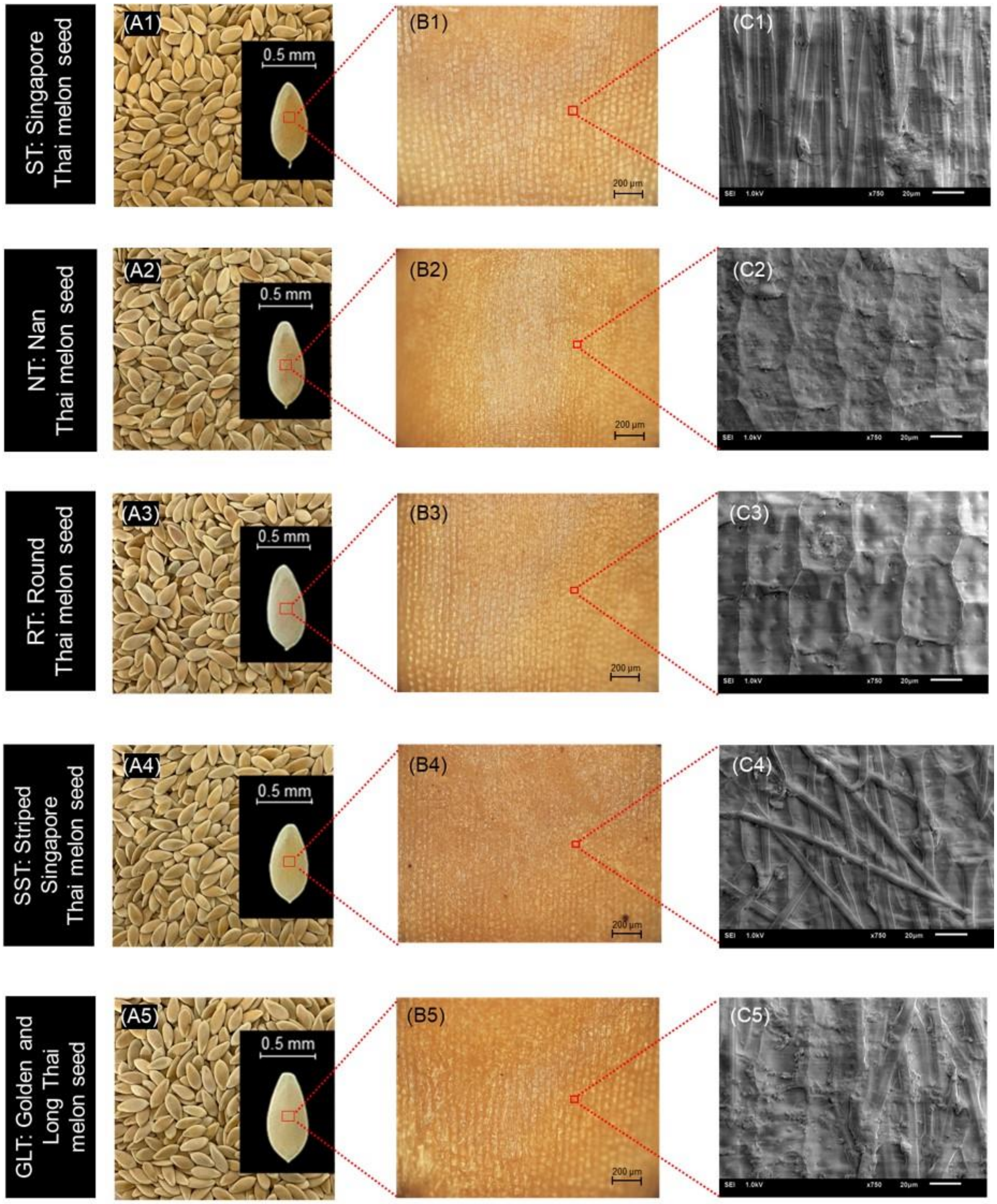
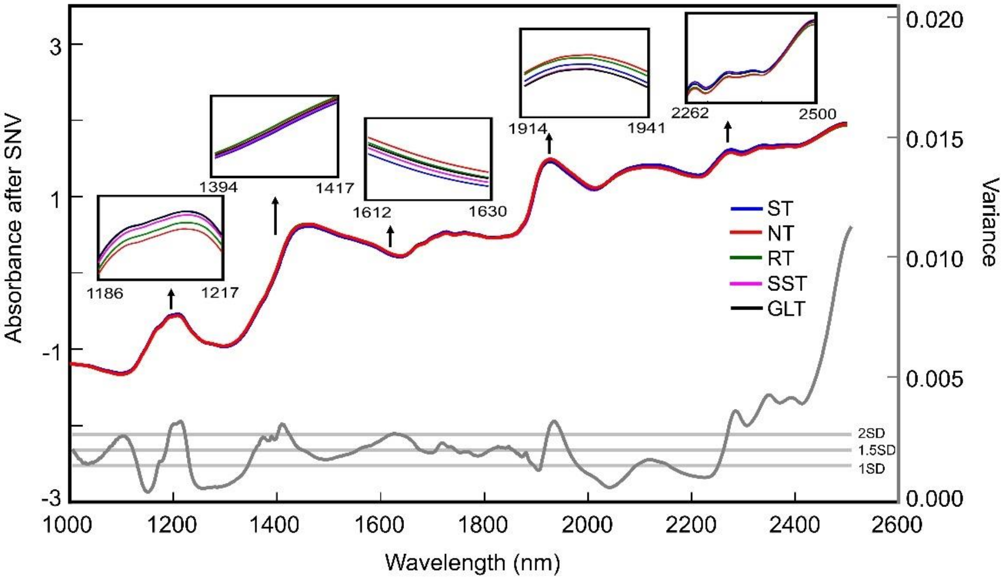


Fig. 3 Morphological features of Thai melon seeds. Digital images of the Thai melon samples ST, NT, RT, SST, and GLT are presented in the acquisition stage (A1)–(A5), respectively. Optical microscopy images (100×) and scanning electron microscopy (SEM) images (750×) are shown in (B1)–(B5) and (C1)–(C5). ST: Singapore Thai melon; NT: Nan Thai melon; RT: Round Thai melon; SST: Striped Singapore Thai melon; GLT: Golden and Long Thai melon.

3.2 NIR spectra of Thai melon seed

Assigning NIR bands is challenging because of the broad and overlapping bands. The visual examination of the NIR spectra within the 1,000–2,500 nm wavelength region revealed no obvious difference among different seed varieties. Yet, major spectral areas could still be identified using the variance value, a statistical measurement calculated by taking the average of squared deviations from the average spectra. Fig. 4 displays the average NIR spectra of the five varieties of Thai melon seeds after pre-processing using Savitsky–Golay smoothing filter to minimize noise and SNV to attenuate the unwanted fluctuations in the NIR dataset [32]. It also depicts the computed and displayed variance of the NIR spectra (bottom line). Any overtone areas with a variation larger than a twofold standard deviation (2SD) may serve as possible markers for Thai melon seed variants. These distinctive reflection bands are comparable with those of melon seeds reported by other studies [21, 33, 34]. Five key areas in the spectra, comprising carbohydrate, starch, moisture, and protein contributions, are summarized in the inset table of Fig. 4. The 1,200 nm band was assigned to the second overtone of C–H in carbohydrates, whereas the 1,450 nm band was attributed to the combination of the first overtones of the C–H bond in protein and O–H bond in moisture [33]. The absorption band between 1,612 and 1,630 nm corresponded to the first overtone of the C–H stretching vibration of the methyl and methylene groups [21]. The spectral region between 2,262 and 2,500 nm was related to the C–H stretch and CH_2 deformation of starch [34]. Evidently, the five kinds of Thai melon seeds had distinct NIR reflectance intensities at wavelengths between 1,000 and 2,500 nm, indicating that they contained varying amounts of lignocellulosic biomass components.



Wavelength (nm)	Band assignment	Structure
1186−1217	2nd overtone of the C−H bond	Carbohydrates or starch
1396−1417	Combination of the first overtones of the N−H bond and O−H in moisture	Protein or amino acids and O−H in moisture
1612−1630	First overtone of C−H stretching vibration	Methyl and methylene group
1914−1941	O−H stretch and H−OH deformation Combination	Starch, cellulose, and H_2O
2260−2500	C−H stretch and CH_2 deformation	Starch

Fig. 4 Mean absorbance near-infrared (NIR) spectra of Thai melon seeds, including ST (blue), NT (red), RT (green), SST (magenta), and GLT (black), after performing standard normal variate (SNV) with the variance plot on the bottom. The inset table demonstrates the band assignment of significant NIR regions for Thai melon discrimination chosen from the NIR region with high variance. ST: Singapore Thai melon; NT: Nan Thai melon; RT: Round Thai melon; SST: Striped Singapore Thai melon; GLT: Golden and Long Thai melon [33] [21] [34].

The current study involves a substantial number of samples with the objective of classifying five distinct types of Thai melon seeds. The use of a large sample size for seed classification offers various advantages, including improved accuracy, robustness, representation of seed variability, statistical significance, and the ability to identify subtle differences, as compared to single seed detection approaches. Particularly in practical scenarios, such as industrial contexts, it is common to employ a vast number of samples. However, single seed-by-seed classification has its own merits, providing a more detailed and focused analysis of each seed, which can be valuable when dealing with heterogeneous seed populations or when precise discrimination is required. We have previously report using single-seed classification approach based on our adaptive SOMs [35]. This methodology allowed for the prediction of individual seed features using data from the entire seed without the need to manually identify specific regions of interest (ROIs). It is crucial to be noted that this single-seed approach was based on the utilization of hyperspectral NIR imaging. Therefore, our developed method can serve for both single-seed and seed batch sample discrimination, depending on the user's specific purposes.

3.3 Multiclassification of Thai melon seeds

Modified SOMs for multiclassification were performed on the collected NIR spectra of the Thai melon seeds from five varieties: ST, NT, RT, SST, and GLT. Herein, the modified SOMs developed using an in-house coding algorithm were utilized to illustrate the underlying link and categorize the five different seed samples. The supervised SOM model typically functioned in two modes: (i) model creation and (ii) classified mapping. In this phase, the map was trained to utilize the training set input samples. A group of test set samples was categorized automatically using the created map. Five types of Thai melon seeds were distinguished using the modified SOM network via 2D mapping visualization. Although SOMs contain several configurable parameters, an optimization procedure is always required to reach the optimal network [36]. In this paper, the scaling value, the size of the map, and the number of iterations were considered as they have a significant impact on the prediction accuracy.

In examining the classification performance, each classification step was performed 10 times. In each replicate, samples were randomly split into the training set (two-thirds of all samples) and the test set (one-third of all samples). Therefore, the number of training samples for each class was proportional to the number of test samples for the class. The evaluation of classification performance was based on %CC (Percentage correctly classified). For predictive modeling, the overall %CC was simply the sum of correctly classified samples divided by the total number of samples [37, 38]. From a statistical aspect, a model with a high %CC is a good classifier, whereas a model with a low %CC is likely to be poor. A more in-depth explanation of the metrics can be found elsewhere [23].

First, the scaling value (ω) for the supervised SOMs was optimized. If ω was too small, it produced a nearly unsupervised map, whereas a high value might result in data overfitting [39]. Fig. 5A shows the overall %CC of the training and test sets when the supervised SOM model was created using various scaling parameters. Initially, when the ω was raised, %CC increased until the classification model gave a steady prediction. When the rate %CC either straightened out or stabilized, the best scaling value for each case was instantaneously determined. This resulted in the ideal scaling value of 2, which yielded the maximum %CC of 70 and 65.80 for the training and test sets, respectively. The corresponding SOM map using different scaling values is shown on the right-hand side in Fig. 6A. In addition to the scaling factor, the map size (number of units) is a critical parameter for classification effectiveness. A smaller map generates more comprehensive patterns, but may not sufficiently describe some substantial changes. Meanwhile, larger map sizes produce more sophisticated patterns but may cause model overtraining [40]. Consequently, determining the appropriate map size is crucial [41]. As illustrated in Fig. 6B, a larger map size resulted in a marginally more precise classification. From the five different map sizes used in this study (12 × 12, 20 × 20, 30 × 35, 35 × 40, and 65 × 70), the 35 × 40 (1400 unit cell) provided the greatest %CC. Additional information on generating supervised SOM maps of various sizes is shown on the right-hand side in Fig. 5B. Consequently, the chosen map size (35 × 40) together with the optimal scaling value (ω = 2) was further used to construct the SOM map to determine the ideal number of iterations. Next, the appropriate number of iterations corresponding to the number of samples must be indicated. The number of iterations was designed to be higher than the number of map units to ensure that the map has sufficient opportunities to be trained from the samples, resulting in sufficient accuracy [42]. However, the larger number of iterations resulted in higher computing demands of SOMs [41]. In other words, while %CC increased as the number of iterations increased, the training procedure time was substantially longer and the cost/benefit might not justify the efforts. Herein, the value of 200,000 (~142 times higher than the number of map units) was determined as the ideal number of iterations for creating a global SOM map as shown in Fig. S3. Fig. 5C illustrates the discrimination performance of the SOM map constructed using these optimized parameters in classifying five Thai melon seeds. Compared with the discrimination results of models constructed using nonoptimized parameters, the model with well-optimized parameters demonstrated significantly improved discrimination performance with a high percentage of correct classifications. The relevant SOM map with various number of iterations is depicted on the right-hand side in Fig. 5C.

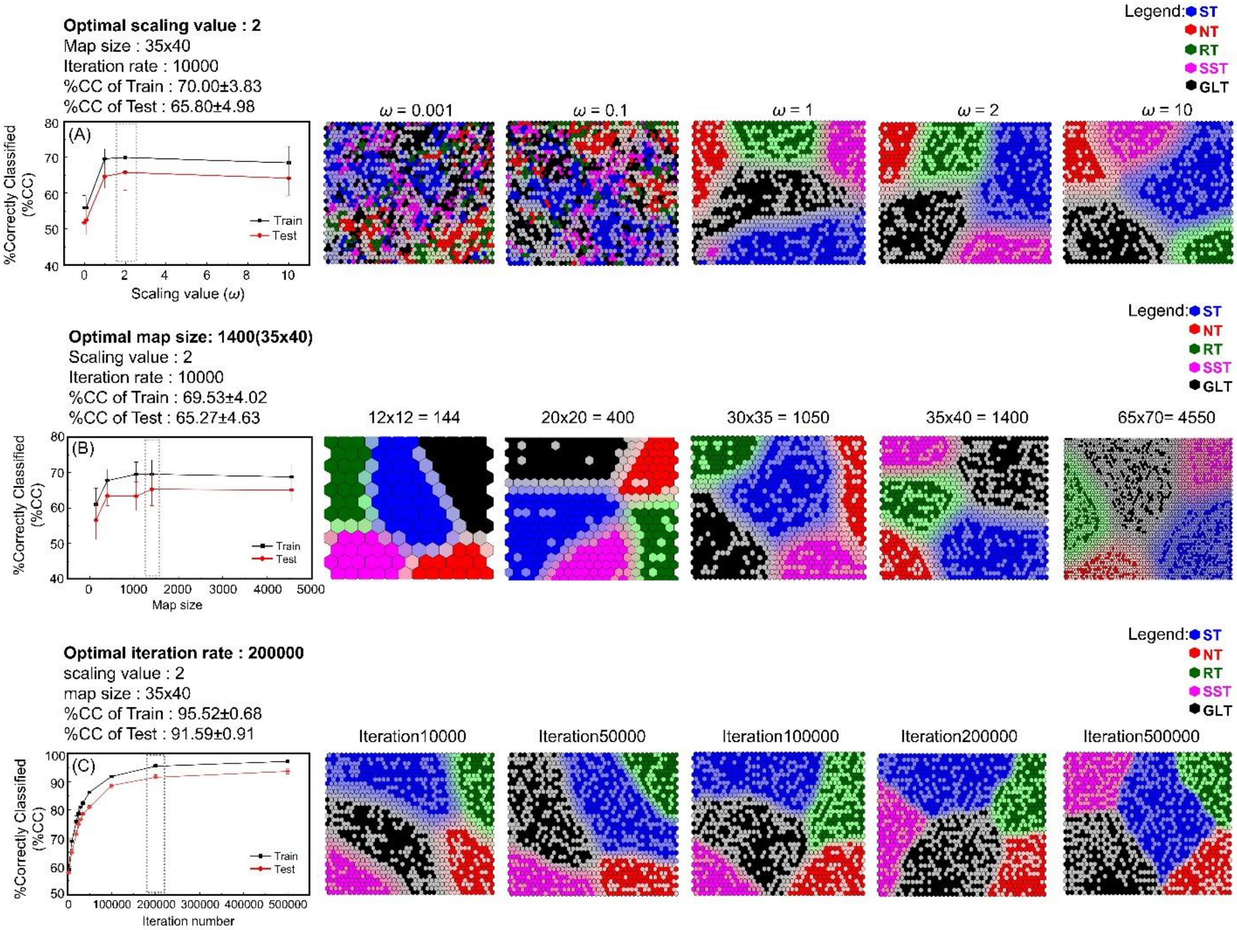


Fig. 5 Percentage of correct classifications (%CC) of the training and test sets (average from 10 iterations) with the optimization of different parameters used to create the supervised self-organizing map (SOM) model for the multiclassification of five classes of Thai melon seeds: (A) scaling value, (B) map size, and (C) number of iterations.

After the SOM map was constructed from the optimized parameters, including the scaling value (*ω* = 2), map size (35 × 40 units), and number of iterations (200,000), it was then used to classify the five varieties of Thai melon seeds. The binary classification using the One vs Rest strategy was first performed. The interaction of two classes (One vs Rest) was established beneath a contingency table to gain insight into the potential of the developed supervised SOMs for discriminating different Thai melon seed varieties. Thus, the discrimination efficacies based on different chemometric approaches including Euclidean distance (EDC), linear discrimination analysis (LDA), quadratic discrimination analysis (QDA), and our adaptive SOMs, were compared. The model performance was validated by five key indicators: sensitivity, specificity, precision, accuracy, and ME [23]. The leave-one-out cross-validation approach was used to validate the classifier models for case I−V. The optimized number of principal components (PCs) was carefully considered for LDA, QDA, and PLS-DA calculation, as shown in Fig. S4.

Fig 6 compares the efficacies and validities of different chemometric approaches for Thai melon seed discrimination. Generally, a good classifier model should give high values of sensitivity, specificity, precision, and accuracy and a low ME value. In all cases, the modified SOMs exhibited the best performance across all indices. Regarding the performance of the SOM discrimination, the sensitivity, specificity, precision, and accuracy were outstanding (>0.9), with a remarkably small ME (<0.01). Only the SOM discrimination model gave a balanced value of sensitivity and specificity, indicating an unbiased discrimination even though the number of samples in each class was extremely unequal. There were many possible reasons for this excellent performance. For example, SOMs can recognize and capture nonlinear relationships in data, whereas conventional chemometric techniques rely on linear assumptions. Furthermore, SOMs rely on a self-organizing process that can adapt to the structure of the data [43]. The results imply that the model was less affected by the unbalanced sample size dataset. On the basis of the performance indices from cases I to V, the developed classifier using the supervised SOMs can be used to classify and distinguish target Thai melon seeds with high precision and accuracy. The sample cluster of cases I−V when the supervised SOMs with optimal parameters (scaling value, map size, and number of iterations) were applied is illustrated in Fig. S5.

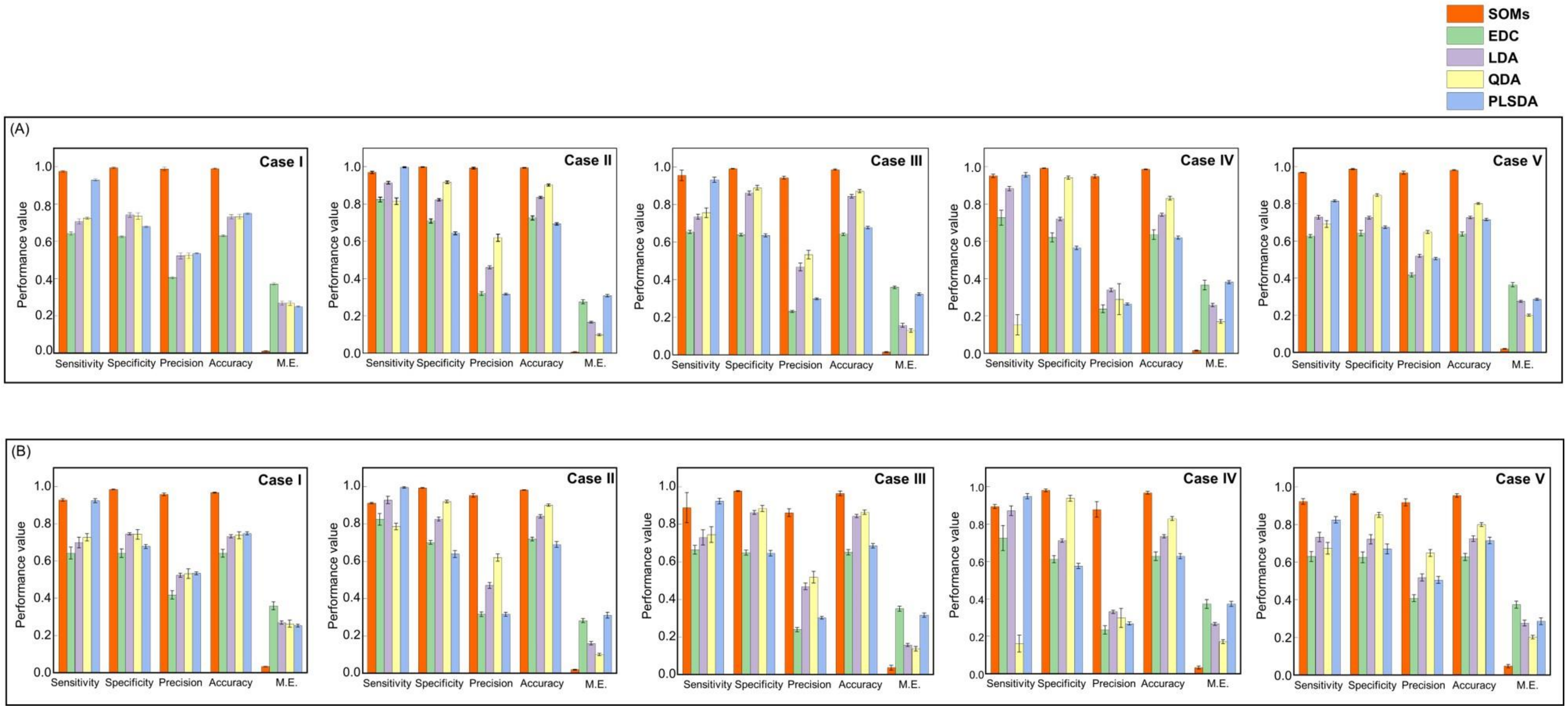


Fig. 6 Performance of the developed and modified self-organizing maps (SOMs) to classify one vs all classes of Thai melon seeds compared with different chemometric techniques, including Euclidean distance to centroids (EDC), Linear discriminant analysis (LDA), Quadratic discriminant analysis (QDA), and Partial least-squares discriminant analysis (PLS-DA) for the (A) training dataset and (B) test dataset. Cases I–V were generated to evaluate the binary classification using the One vs Rest strategy. Case I: Singapore Thai melon (ST) vs Rest; Case II: Nan Thai melon (NT) vs Rest; Case III: Round Thai melon (RT) vs Rest; Case IV: Striped Singapore Thai melon (SST) vs Rest; and Case V: Golden and Long Thai melon (GLT) vs Rest.

Fig. 7 compares the %CC values of our adaptive supervised SOMs and other chemometric approaches, including EDC, LDA, and QDA. The discrimination experiments were performed with 10 iterations on training and test sets to show the stability of the estimated %CC. This could provide the mean and standard deviation of the discrimination performance. The graphs in the diagonal axis exhibit the correct classification, whereas the off-diagonal graphs demonstrate the incorrect classifications (where the predicted class does not match the actual sample class). The %CC of the training set indicates how well the classifier model was optimized, whereas the %CC of the test set shows how well the model could predict the sample class. The %CC results indicate that the modified SOMs provide superior discrimination efficiency (high %CC) with great consistency compared with other approaches. A good balance between the prediction of all classes suggests that the classifier model based on the modified SOMs was not biased toward either group and that the SOMs parameters (scaling value, map size, and iterations) were well optimized.

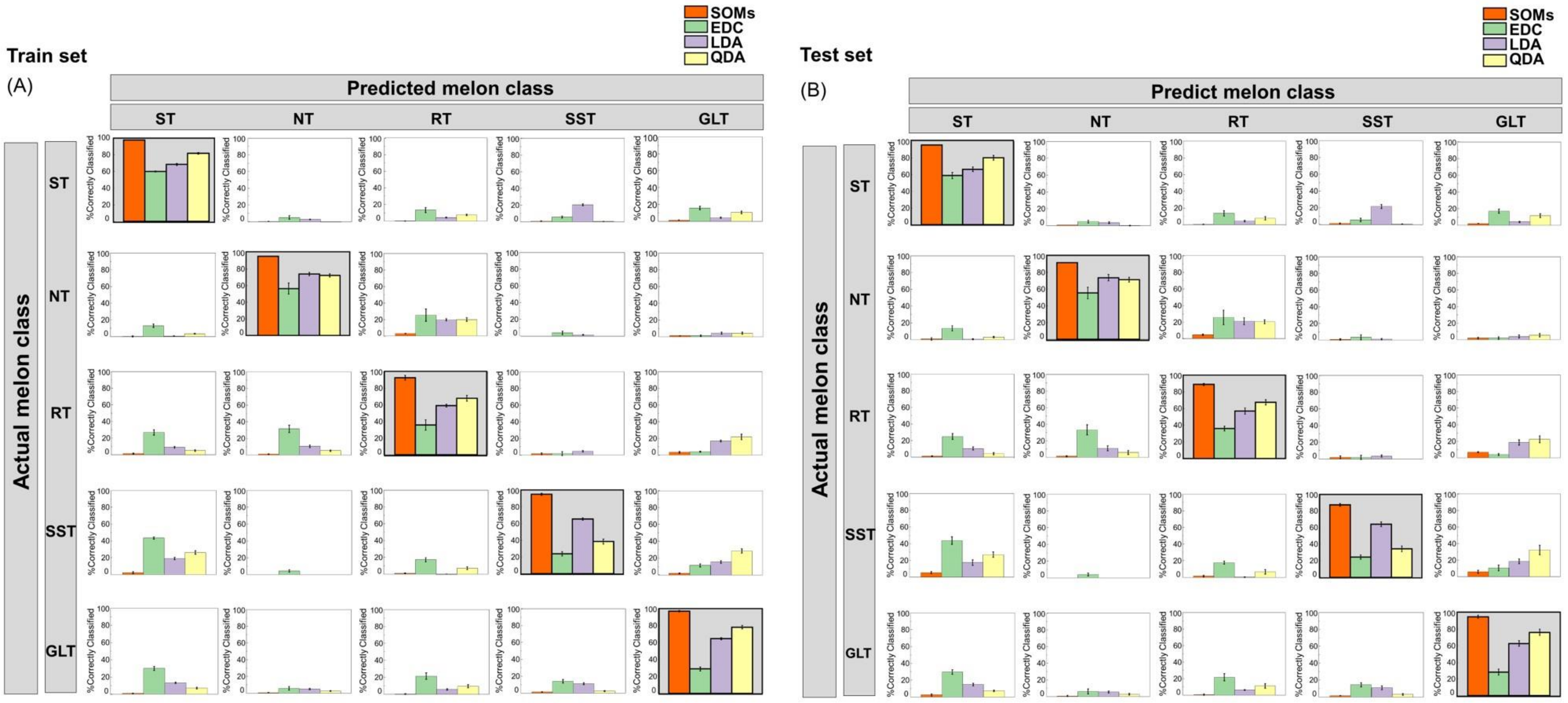


Fig. 7 Percentage of correct classifications (%CC) of five classes of Thai melon seeds using different chemometric models: Euclidean distance to centroids (EDC), Linear discriminant analysis (LDA), Quadratic discriminant analysis (QDA), and our adaptive supervised self-organizing maps (SOMs).

According to the above approach, SOMs are suitable for displaying data with numerous variables. Fig. 8A shows the score plots of the top three largest principal components (PC1–PC3) to provide empirical evidence for the discrimination. Unsupervised and supervised SOMs were used to compare sample group discrimination from the score plots, as demonstrated in Fig. 8B and 8C, respectively.

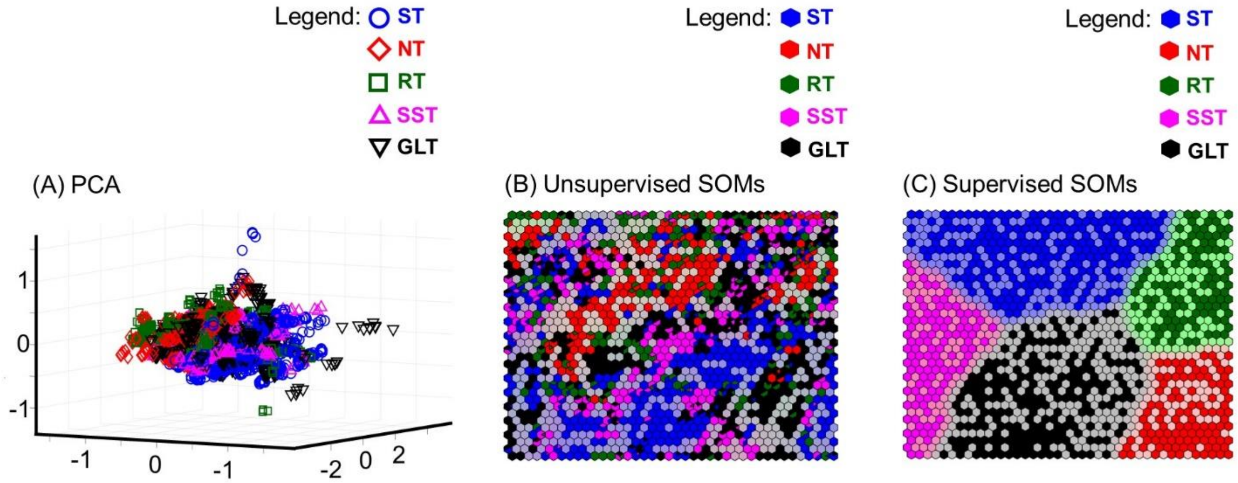


Fig. 8 (A) Principal component analysis (PCA) score plots (PC1–PC3), (B) unsupervised self-organizing maps (SOMs), and (C) supervised SOMs of the discrimination of five classes of Thai melon seeds using the optimal parameters (scaling value, map size, and number of iterations).

From a data visualization perspective, if samples fall into groups or classes, they can be used to shade the background on the SOM. The map unit is shaded in the color of its closest BMU. If more than one BMU is equidistant from the unit, it is shaded in a combination of colors, according to how many BMUs from each group it is closest to two [41]. In other words, any sample belonging to a similar class is projected in the same BMU, resulting in the same shade in color for that class. Meanwhile, if the samples have slightly different properties, they are projected in the combination of many BMUs, resulting in the combination of color shades for the samples (or light shades in case). In Fig. 8, the results indicate that the principal component analysis (PCA) score plots for the sample groups are heavily overlapping, resulting in a barrier that makes it difficult to distinguish between various groupings. A possible explanation for this is that many of the data points in our input data had similar chemical properties and thus overlapped excessively on the score plot. Besides that, SOM made use of the entire available space on the map, whereas PCA utilized just a fraction of it. The unsupervised SOMs reveal that the sample groups were not uniformly distributed, whereas our modified supervised SOMs significantly improved the separation of sample clusters. The possible reason behind this achievement was that our adaptive supervised SOM possessed the optimal scaling values, enabling it to proficiently group the five Thai melon seed samples into predetermined clusters on the map [39, 41]. Consequently, our modified SOMs, in combination with NIR, were highly effective in differentiating types of Thai melon seeds.

4. Conclusion

In this study, a novel multiclassification strategy for five Thai melon seeds based on NIR spectroscopy and adaptive supervised SOMs was presented. The morphological traits or visual appearance of the seeds showed no noticeable difference among different varieties. Thermal degradation profiles revealed the unique amounts of lignin content and carbohydrate content of the seeds with different varieties. The primary bands of the key biomass components, including hemicellulose, cellulose, and lignin, were detected using FTIR. An intense FTIR band was observed in the 3,000 cm^{-1} region, which was proportional to the number of intermolecular linked –OH groups in lignin and carbohydrates. IR and TGA data corroborated the hypothesis that the Thai melon seeds from the five varieties possessed different chemical characteristics. In the multiclassification process from the NIR spectra, the supervised SOMs' parameters, including the optimal scaling value (ω), map size, and number of iterations, were optimized to produce a global SOM map. By using the optimum parameters, exceptional classification results were achieved with an overall %CC of 95.52 ± 0.68% for the training set and 91.59 ± 0.91% for the test set, respectively. Our modified SOMs clearly outperformed other approaches in differentiating between the five classes of Thai melon seeds. Accordingly, the developed SOMs provide excellent multiclassification results and can be used as a nondestructive technique for discrimination Thai melon seeds.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research project is supported by the Second Century Fund (C2F), Chulalongkorn University, National Research Council of Thailand (NRCT) Grant/Contract Number N72A660881 and the National Science, Research and Innovation Fund (NSRF) via the Program Management Unit for Human Resources & Institutional Development, Research and Innovation (B16F640101).

References

[1] M. Martuscelli, C. Di Mattia, F. Stagnari, S. Speca, M. Pisante, D. Mastrocola, Influence of phosphorus management on melon (Cucumis melo L.) fruit quality, Journal of the Science of Food and Agriculture 96 (8) (2016) 2715–2722, https://doi.org/10.1002/jsfa.7390.

[2] G.E. Lester, Antioxidant, sugar, mineral, and phytonutrient concentrations across edible fruit tissues of orange-fleshed honeydew melon (Cucumis melo L.), Journal of Agricultural and Food Chemistry 56(10) (2008) 3694–3698, https://doi.org/10.1021/jf8001735.

[3] G.E. Lester, J.L. Jifon, D.J. Makus, Impact of potassium nutrition on postharvest fruit quality: Melon (Cucumis melo L) case study, Plant and soil 335(1) (2010) 117–131, https://doi.org/10.1007/s11104-009-0227-3.

[4] Z. Bishaw, A.A. Niane, Y. Gan, Quality seed production, In: Yadav, S.S., McNeil, D.L., Stevenson, P.C. (eds) Lentil, Springer, Dordrecht (2007) 349–383, https://doi.org/10.1007/978-1-4020-6313-8_21.

[5] H.R. Jensen, L. Belqadi, P. De Santis, M. Sadiki, D.I. Jarvis, D.J. Schoen, A case study of seed exchange networks and gene flow for barley (Hordeum vulgare subsp. vulgare) in Morocco. Genetic resources and crop evolution 60(3) (2013) 1119–1138, https://doi.org/10.1007/s10722-012-9909-4.

[6] Khomphet, T., et al., *Genetic Variability, Correlation, and Path Analysis of Thai Commercial Melon Varieties.* International Journal of Agronomy, 2022. **2022**.

[7] T. Khomphet, W. Intana, A. Promwee, S.S. Islam, Genetic Variability, Correlation, and Path Analysis of Thai Commercial Melon Varieties, International Journal of Agronomy 2022 (2022), https://doi.org/10.1155/2022/7877239.

[8] U. Kiran, S. Khan, K.J. Mirza, M. Ram, M.Z. Abdin, SCAR markers: a potential tool for authentication of herbal drugs, Fitoterapia, 81(8) (2010) 969–976, https://doi.org/10.1016/j.fitote.2010.08.002.

[9] N. Choudhary, B.S. Sekhon, An overview of advances in the standardization of herbal drugs, Journal of Pharmaceutical Education and Research, 2(2) (2011) 55.

[10] P.Y. Yip, C.F. Chau, C.Y. Mak, H.S. Kwan, DNA methods for identification of Chinese medicinal materials, Chinese Medicine 2 (2007) 1–19, https://doi.org/10.1186/1749-8546-2-9.

[11] M.M. Oliveira, J.P. Cruz-Tirado, J.V. Roque, R.F. Teófilo, D.F. Barbin, Portable near-infrared spectroscopy for rapid authentication of adulterated paprika powder. Journal of Food Composition and Analysis, 2020. 87: p. 103403.

[12] J.U. Porep, D.R. Kammerer, R. Carle, On-line application of near infrared (NIR) spectroscopy in food production, Trends in Food Science & Technology, 46(2, Part A) (2015) 211–230, https://doi.org/10.1016/j.tifs.2015.10.002.

[13] E. Teye, C.L. Amuah, T. McGrath, C. Elliott, Innovative and rapid analysis for rice authenticity using hand-held NIR spectrometry and chemometrics, Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, (217) (2019) 147–154, https://doi.org/10.1016/j.saa.2019.03.085.

[14] H.S. Park, K.C. Choi, J.H. Kim, M.J. So, S.H. Lee, K.W. Lee, Discrimination and quantification between annual ryegrass and perennial ryegrass seeds by near-infrared spectroscopy, JAPS: Journal of Animal & Plant Sciences 26(5) (2016).

[15] J. Zhang, M. Li, T. Pan, L. Yao, J. Chen, Purity analysis of multi-grain rice seeds with non-destructive visible and near-infrared spectroscopy 164 (2019) 104882, https://doi.org/10.1016/j.compag.2019.104882.

[16] P. Wang, Yu, Z. Species authentication and geographical origin discrimination of herbal medicines by near infrared spectroscopy: A review, Journal of pharmaceutical analysis 5(5) (2015) 277–284, https://doi.org/10.1016/j.jpha.2015.04.001.

[17] G. Reich, Near-infrared spectroscopy and imaging: basic principles and pharmaceutical applications, Advanced drug delivery reviews 57(8) (2005) 1109–1143, https://doi.org/10.1016/j.addr.2005.01.020.

[18] Z. Seregely, T. Deak, G.D. Bisztray, Distinguishing melon genotypes using NIR spectroscopy, Chemometrics and Intelligent Laboratory Systems 72(2) (2004) 195–203, https://doi.org/10.1016/j.chemolab.2004.01.013.

[19] G.D. Bisztray, T. Deak, Z.S. Seregély, K. Kaffka, NIR spectroscopy for distinction of horticultural plant seeds, in V International Symposium on In Vitro Culture and Horticultural Breeding 725 (2004), https://doi.org/10.17660/ActaHortic.2006.725.99.

[20] H. Lee, M.S. Kim, H.S. Lim, E. Park, W.H. Lee, B.K. Cho, Detection of cucumber green mottle mosaic virus-infected watermelon seeds using a near-infrared (NIR) hyperspectral imaging system: Application to seeds of the "Sambok Honey" cultivar, Biosystems Engineering 148 (2016) 138–147, https://doi.org/10.1016/j.biosystemseng.2016.05.014.

[21] J. Yasmin, M. Raju Ahmed, S. Lohumi, C. Wakholi, M.S. Kim, B.K. Cho, Classification method for viability screening of naturally aged watermelon seeds using FT-NIR spectroscopy, Sensors 19(5) (2019) 1190, https://doi.org/10.3390/s19051190.

[22] H.P. Vinutha, B. Poornima, B.M. Sagar, Detection of outliers using interquartile range technique from intrusion dataset, In: Satapathy, S., Tavares, J., Bhateja, V., Mohanty, J. (eds) Information and Decision Sciences. Advances in Intelligent Systems and Computing, Springer, Singapore 701 (2018) 511–518, https://doi.org/10.1007/978-981-10-7563-6_53.

[23] S. Makmuang, S. Nootchanat, S. Ekgasit, K. Wongravee, Non-destructive method for discrimination of weedy rice using near infrared spectroscopy and modified self-organizing maps (SOMs), Computers and Electronics in Agriculture 191 (2021) 106522, https://doi.org/10.1016/j.compag.2021.106522.

[24] S.F. Sim, V. Sági-Kiss, Multiple Self Organising Maps (mSOMs) for simultaneous classification and prediction: Illustrated by spoilage in apples using volatile organic profiles, Chemometrics Intelligent Laboratory Systems 109(1) (2011) 57–64, https://doi.org/10.1016/j.chemolab.2011.08.001.

[25] Y. Liu, R.H. Weisberg, C.N. Mooers, Performance evaluation of the self-organizing map for feature extraction, Journal of Geophysical Research: Oceans 111(C5) (2006), https://doi.org/10.1029/2005JC003117.

[26] K. Wongravee, M. Ishigaki, Y. Ozaki, Chemometrics as a Green Analytical Tool, in Challenges in Green Analytical Chemistry, S. Garrigues and M. de la Guardia, Editors, The Royal Society of Chemistry (2020).

[27] M. Cocchi, A. Biancolillo, F. Marini, Chapter Ten - Chemometric Methods for Classification and Feature Selection, in Comprehensive Analytical Chemistry, J. Jaumot, C. Bedia, and R. Tauler, Editors, Elsevier (2018) 265–299, https://doi.org/10.1016/bs.coac.2018.08.006.

[28] A.L. Pomerantsev, O.Y. Rodionova, Multiclass partial least squares discriminant analysis: Taking the right way—A critical tutorial, Journal of Chemometrics 32(8) (2018) e3030, https://doi.org/10.1002/cem.3030.

[29] A.L. Pomerantsev, O.Y. Rodionova, New trends in qualitative analysis: Performance, optimization, and validation of multi-class and soft models, TrAC Trends in Analytical Chemistry 143 (2021) 116372, https://doi.org/10.1016/j.trac.2021.116372.

[30] R.G. Brereton, Chemometrics for pattern recognition, John Wiley & Sons (2009).

[31] J.A. da Cunha, P.M. Rolim, K.S.F.D.S.C. Damasceno, F.C. de Sousa Júnior, R.C. Nabas, L.M.A.J. Seabra, From seed to flour: sowing sustainability in the use of cantaloupe melon residue (Cucumis melo L. var. reticulatus). PloS one 15(1) (2020) e0219229, https://doi.org/10.1371/journal.pone.0219229.

[32] B.T. Borille, M.C.A. Marcelo, R.S. Ortiz, K. de Cássia Mariotti, M.F. Ferrão, R.P. Limberger, Near infrared spectroscopy combined with chemometrics for growth stage classification of cannabis cultivated in a greenhouse from seized seeds, Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 173 (2017) 318–323, https://doi.org/10.1016/j.saa.2016.09.040.

[33] S. Yang, Q.B. Zhu, M. Huang, J.W. Qin, Hyperspectral image-based variety discrimination of maize seeds by using a multi-model strategy coupled with unsupervised joint skewness-based

wavelength selection algorithm, Food Analytical Methods 10(2) (2017) 424–433, https://doi.org/10.1007/s12161-016-0597-0.

[34] M.M. da Mata, P.D. Rocha, I.K.T. de Farias, J.L.B. da Silva, E.P. Medeiros, C.S. Silva, S. da Silva Simões, Distinguishing cotton seed genotypes by means of vibrational spectroscopic methods (NIR and Raman) and chemometrics, Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 266 (2022) 120399, https://doi.org/10.1016/j.saa.2021.120399.

[35] S. Makmuang, A. Terdwongworakul, T. Vilaivan, S. Maher, S. Ekgasit, K. Wongravee, Mapping hyperspectral NIR images using supervised self-organizing maps: Discrimination of weedy rice seeds, Microchemical Journal 190 (2023) 108599, https://doi.org/10.1016/j.microc.2023.108599.

[36] D. Ballabio, M. Vasighi, P. Filzmoser, Effects of supervised Self Organising Maps parameters on classification performance, Analytica Chimica Acta 765 (2013) 45–53, https://doi.org/10.1016/j.aca.2012.12.027.

[37] C.D. Brown, H.T. Davis, Receiver operating characteristics curves and related decision measures: A tutorial. Chemometrics and Intelligent Laboratory Systems 80(1) (2006) 24–38, https://doi.org/10.1016/j.chemolab.2005.05.004.

[38] G.R. Lloyd, S. Ahmad, M. Wasim, R.G. Brereton, Pattern recognition of inductively coupled plasma atomic emission spectroscopy of human scalp hair for discriminating between healthy and hepatitis C patients, Analytica chimica acta 649(1) (2009) 33–42, https://doi.org/10.1016/j.aca.2009.07.005.

[39] K. Wongravee, G.R. Lloyd, C.J. Silwood, M. Grootveld, R.G, Brereton, Supervised self organizing maps for classification and determination of potentially discriminatory variables: illustrated by application to nuclear magnetic resonance metabolomic profiling, Analytical Chemistry 82(2) (2010) 628–638, https://doi.org/10.1021/ac9020566.

[40] B.-H. Lee, M. Scholz, A comparative study: Prediction of constructed treatment wetland performance with k-nearest neighbors and neural networks, Water, Air, and Soil Pollution 174(1) (2006) 279-301, https://doi.org/10.1007/s11270-006-9113-2.

[41] R.G. Brereton, Self organising maps for visualising and modelling, Chemistry Central Journal 6(2) (2012) 1–15.

[42] Lek, S. and Y.S. Park, Artificial Neural Networks, in Encyclopedia of Ecology, S.E. Jørgensen and B.D. Fath, Editors, Academic Press: Oxford (2008) 237–245, https://doi.org/10.1016/B978-008045405-4.00173-7.

[43] T. Kohonen, The self-organizing map, Proceedings of the IEEE 78(9) (1990) 1464–1480.