


Leaf-based species classification of hybrid cherry tomato plants by using hyperspectral imaging

Journal of Near Infrared Spectroscopy
2023, Vol. 31(1) 41–51
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/09670335221148593
journals.sagepub.com/home/jns


Songhao Li¹ , Huilin Wu², Jing Zhao^{1,3}, Yu Liu¹, Yunpeng Li¹, Houcheng Liu⁴, Yiting Zhang⁴, Yubin Lan^{1,3,5}, Xinglong Zhang², Yutao Liu² and Yongbing Long^{1,3}

Abstract

Approaches based on near infrared hyperspectral imaging (NIR-HSI) technology combined with machine learning have been developed to classify the leaves of hybrid cherry tomatoes and then identify the species of hybrid cherry tomato plants. The near infrared (NIR) hyperspectral images of 400 cherry tomato leaves (100 per species) were collected in the wavelength range of 900–1700 nm. Machine learning algorithms such as linear discriminant analysis (LDA), random forest (RF), and support vector machine (SVM) were employed to construct leaf classification models with the hyperspectral data pre-processed by Savitzky-Golay (SG) smoothing filter, first derivative (first Der) and standard normal variate (SNV). Principle of Component Analysis (PCA) was also used to reduce the data dimension and extract spectral features. It is revealed that the LDA model reaches the highest classification accuracy among the three machine learning algorithms and SNV can lead to higher improvement in model accuracy than other preprocessing methods of SG smoothing and first Der. Analysis based on PCA spectral feature extraction demonstrates that differences occur in internal material content in the leaves of cherry tomato plants with different species, which renders the models being able to distinguish between the species. Another important work was performed to reveal the different effects of the mesophyll and vein regions (VR) on the accuracy of the leaf classification model. It is demonstrated that the classification accuracy is improved by a value of 0.033 or 0.042 when mesophyll substitutes vein or whole leaf as regions of interest (ROI) to extract reflectance spectra for modeling. As a result, the accuracy of the training and test set respectively reached a high value of 0.998 and 0.973 for the LDA classification model combined with the SNV preprocessing method. The results propose that the use of mesophyll region (MR) as ROI can improve the performance of the leaf classification model, which provides a new strategy for efficient and non-destructive classification of different hybrid cherry tomato plants.

Keywords

Cherry tomato plants, hyperspectral imaging, mesophyll, near infrared spectral analysis, classification model

Received 8 June 2023; accepted 8 December 2023

Introduction

Cherry tomato, *Solanum Lycopersicum* var. *Cerasiforme*, is a variety of tomato genus cultivated subspecies with both high economic value and high edible value. Cherry tomato is one of the most widely consumed vegetables in the world due to its unique flavor and richness of vitamins and antioxidants such as glutathione, lycopene, and beta-carotene polyphenols. It has become one of the “Four Major Fruits and Vegetables” promoted by the Food and Agriculture Organization.¹ Cherry tomato has hundreds of different cultivars due to the high genetic variation² and large difference occurs in flavor, moisture, and nutrient content for cherry tomatoes with different species. For example, Kavitha et al. has demonstrated that the content of lycopene varies significantly between the cherry tomato varieties.³ In addition, different water and nutrient supply conditions are required in the cultivation process of different cherry tomato varieties, which can lead to significant differences in

¹College of Electronic Engineering/College of Artificial Intelligence, South China Agricultural University, Guangzhou, China

²Guangzhou National Modern Agricultural Science and Technology Innovation Center, Guangzhou, China

³South China Intelligent Agriculture Public Research and Development Platform, Ministry of Agriculture and Rural Affairs, Guangzhou, China

⁴College of Horticulture, South China Agricultural University, Guangzhou, China

⁵Lingnan Modern Agricultural Science and Technology Guangdong Lab, Guangzhou, China

Corresponding authors:

Jing Zhao, College of Electronic Engineering/College of Artificial Intelligence, Natl Ctr Int Collaborat Res Precis Agr Aviat Pest, South China Agricultural University, Guangzhou 510642, China.
Email: edithzhao@scau.edu.cn

Yongbing Long, College of Electronic Engineering/College of Artificial Intelligence, South China Agricultural University, Guangzhou 510642, China.
Email: yongbinglong@126.com

antioxidant components for tomato fruits.⁴ Therefore, it is very important to classify the species of cherry tomato plants so that the cultivation patterns and irrigation strategies could be precisely controlled according to the varieties.

Precise cultivation in the seedling stage, flowering stage and fruit development stage of cherry tomato plants plays important but different roles in the determination of the yield and quality of tomato fruit. The research of Chen et al. showed that water deficit exerted a more obvious influence on the yield of tomato fruit during the flowering and fruit development stages than during the seedling stage.⁵ In addition, cherry tomato plants with different species have different sensitivity to the growing environment, especially during the flowering stage.⁶ Therefore, rapid and efficient classification of different species of tomato plants during the flowering stage helps change irrigation strategies to improve fruit yields. In order to classify different species of cherry tomato plants efficiently and accurately, this paper proposes a near infrared (NIR) hyperspectral imaging (HSI) technology to obtain the spectral information of cherry tomato leaves and establish an effective classification model based on machine learning.

HSI is an emerging platform technology that integrates traditional spectral and imaging technology to obtain both spectral and spatial information from samples.⁷ Due to the advantage of simultaneously obtaining spectral and spatial information, its application in the agricultural field has attracted more and more attention.⁸ The hyperspectral data of leaves are as regarded as human fingerprints, which can be used to reveal the nutritional status of the plant and the difference between the species of plants.^{9,10} Yang et al. utilized hyperspectral imaging technology combined with a recognition model based on particle swarm optimization-extreme learning machine to identify eight tree species at the leaf level.¹¹ Hideaki et al. developed an approach based on near infrared hyperspectral imaging (NIR-HSI) technology and deep convolutional neural network to identify 38 hardwood species with an accuracy of 90.5%.¹² Hyperspectral reflectance imaging techniques were applied to discriminate different lettuce varieties based on linear discriminant analysis (LDA) and principal component analysis.¹³

Previous studies showed that leaf-based HSI technology exhibited excellent performance in the classification of plant species.^{14,15} But most of these works used the whole leaf as the region of interest (ROI) to extract reflectance spectra for the classification models. The different effects of the mesophyll region (MR) and vein region (VR) of leaves are not distinguished in modeling although differences occur between the optical behaviors of the MR and VR. The primary goal of the paper is to propose a novel ROI selection strategy to classify the species of cherry tomato plants by HSI techniques, further to improve the classification accuracy. Besides, a classification model based on HSI technology and machine learning is established to classify four hybrid cherry tomato plants. The specific objectives are to (1) compare and analyze the optical behaviors of different ROIs of leaves: the whole leaf region (WLR), VR and MR; (2) compare and analyze the performance of classification models constructed based on

hyperspectral reflectance data extracted from different ROIs; (3) investigate the accuracy of the models based on different spectral preprocessing methods and classifiers.

Materials and methods

Sample preparation

Hybrid cherry tomato plants used for the classification experiment were grown under greenhouse conditions at South China Agriculture University, Guangzhou, China. Four hybrid species of Cherry tomato plants, named Baiyuxiangfei (Bai), Miying (Mi), Moka (Mo) and Huo-linglong (Huo) were cultivated in the greenhouse at 25°C/15°C (day/night temperature) on November 25. The seeds of these four hybrid cherry tomato plants were provided by the Chenhong Seed Company. All the cherry tomato plants were cultivated in pots using the Japanese Yamasaki nutrient solution.¹⁶ They were watered every 3 hours on sunny days and every 6 hours on cloudy days. To ensure that any observed difference exclusively originated from varieties, all the cherry tomato species were grown in the same environment with the same nutrient and water supply during planting. At the flowering stage (8 weeks after sowing), one hundred leaves were picked from the cherry tomato plants for each specie, and four hundred leaves were then obtained for the hyperspectral image collection experiment. Only the leaves with small curvature were selected to reduce the effects of curvature on hyperspectral imaging. The leaves were immediately packaged in plastic bags after being picked and transported to the laboratory for spectral imaging within 10 minutes.

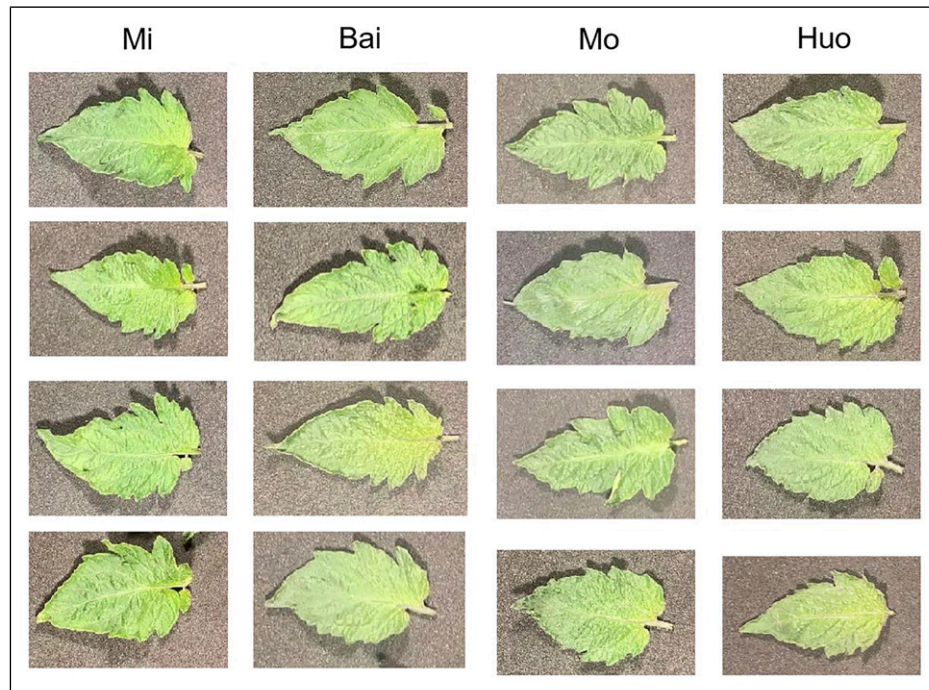
Table 1 shows the average plant height, stem thickness, and leaf area of the four cherry tomato species used in the experiment. Figure 1 shows RGB images of leaves for the four cherry tomato species, which were captured by RGB camera. The figure demonstrates that the shape, size, and colors of the four species were highly similar.

Hyperspectral imaging system and image acquisition

The hyperspectral image data of cherry tomato leaves were acquired with a NIR hyperspectral camera (GaiaField Pro-N17E, Sichuan Shuanglihepu Technology Co. Ltd, China). The NIR hyperspectral camera consists of an imaging spectrograph (ImSpector N17 E, Specim, Finland) with a wavelength range of 900–1700 nm and a spectral resolution of 5 nm, and an Indium Gallium Arsenide (InGaAs) charge-coupled device (CCD) with a resolution of 320 × 400 pixels. The hyperspectral camera has a built-in translation push scanning device without an additional scanning platform. The moving speed of the push-broom motor was set as 0.45 mm/s. The field of view for the hyperspectral camera was 17.6°. The height between the hyperspectral camera lens and samples was ca. 40 cm, and each pixel of CCD corresponded to about 0.0185 cm for the sample. Four tungsten halogen lamps (HSIA-LS-T-H-200W, Sichuan Shuanglihepu Technology Co. Ltd) with a spectral range of 350–2500 nm were adopted as light sources. To provide

Table 1. Details of the four cherry tomato varieties at sampling (8 weeks).

Species	Average plant height (cm)	Average stem thickness (mm)	Average single leaf area (cm ²)
Mi	148.76	16.24	20.36
Huo	150.88	16.53	19.92
Bai	149.05	16.17	18.73
Mo	152.22	16.67	20.22

**Figure 1.** RGB images of leaves for four cherry tomato species.

uniform lighting in the field of view, the angle between lamps and the horizontal plane was set to 45° and the center of halogen lamps was kept about 30 cm from the loading platform. To reduce the influence of external light, the hyperspectral camera, halogen lamps and samples were located in a dark box, as shown in Figure 2. The control software SpecVIEW (Version 2.9, Sichuan Shuanglihepu Technology Co. Ltd) was used to collect hyperspectral data and set acquisition parameters. Finally, the spectral data of each sample as collected by the hyperspectral camera with an exposure time of 30 ms. During hyperspectral image acquisition, the leaves were placed flat on a black platform with the adaxial side facing the camera.^{17,18} The shooting time of one hyperspectral image was about 20 s per sample. The effects of heat on leaf samples caused by the lights could be ignored in such a short time and no obvious changes were observed for leaves after hyperspectral image acquisition. All the reflectance spectral data was recorded as a three-dimensional cube with 256 bands.

Correction of hyperspectral image

To reduce the effects of dark current and illumination discrepancies,¹⁹ all raw hyperspectral images were corrected by black and white references using the following equation²⁰

$$I_c = \frac{I_{raw} - I_{black}}{I_{white} - I_{black}} \quad (1)$$

where I_c is the reflectance of images after correction; I_{raw} is the original intensity of the raw image; I_{black} is the dark reference image acquired from the standard blackboard and I_{white} is the white reference image obtained from the white Teflon board. The white Teflon board (HSIA-CT-150 × 150) with a reflectance of 99% was provided by Sichuan Shuanglihepu technology company.

Region of interest

Selection of the region of interest (ROI) always plays an important role in classification tasks based on hyperspectral image data since the optical behavior of different regions of the sample is always different. Thus, WLR, VR and MR of each hyperspectral image of leaves were respectively selected as ROIs for reflectance extraction and model construction. These three ROIs were obtained by threshold morphological operation and the process is shown in Figure 3.²¹ At first, the raw hyperspectral image at 1375 nm (Figure 3(a)) was converted into a grayscale image (Figure 3(b)). The ROI of the WLR as then created by applying the thresholding algorithm with a thresh-binary value of 550 to the grayscale image (Figure 3(c)). Thresh-to-zero operation with a value of 550 and inverse thresh-to-

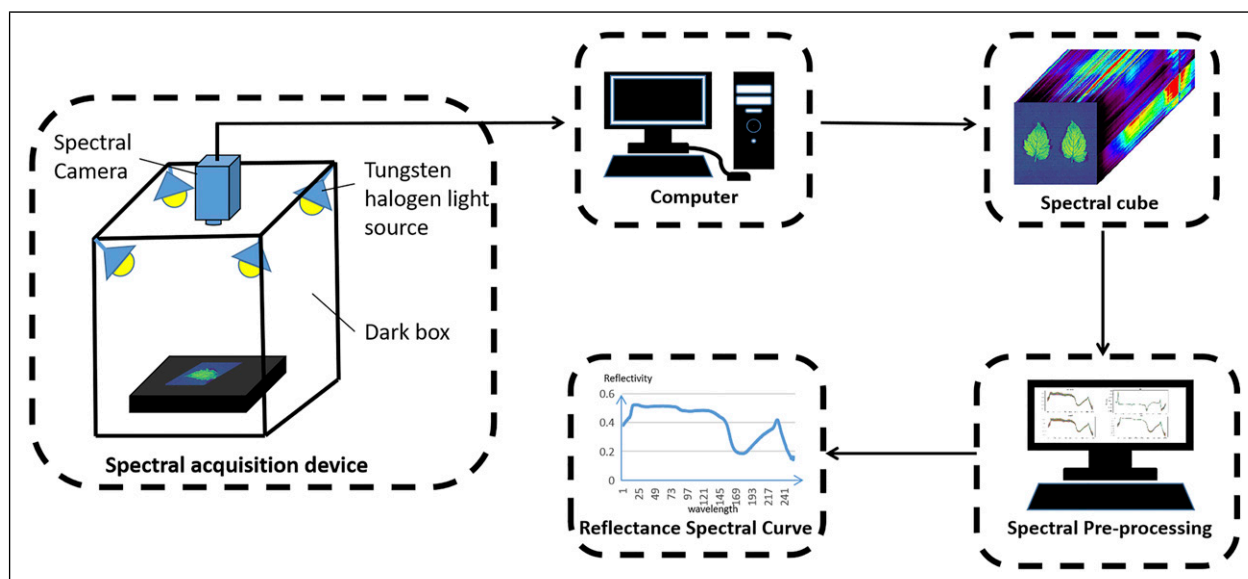


Figure 2. Schematics of the hyperspectral imaging system.

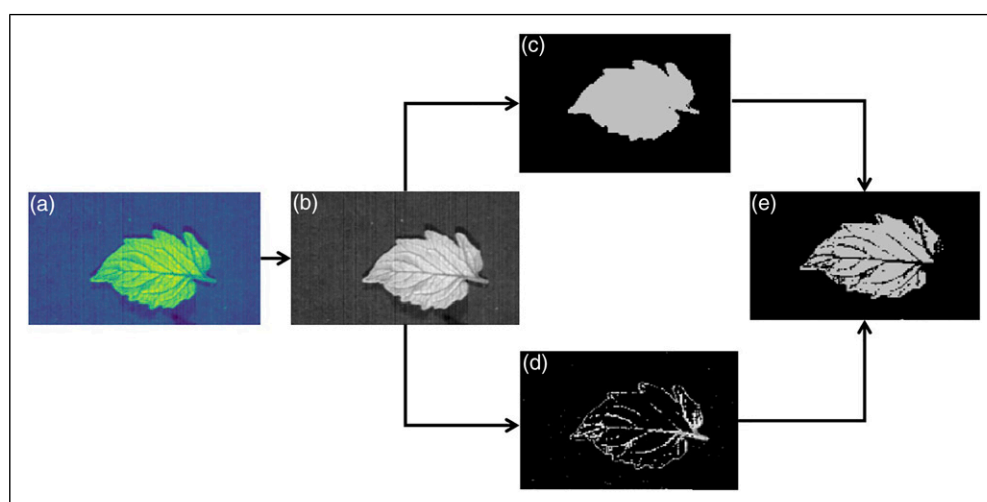


Figure 3. Process of ROI selection. (a) Raw hyperspectral image at 1375 nm; (b) grayscale image; (c) ROI-WLR; (d) ROI-VR; (e) ROI-MR.

zero operation with a value of 660 was performed to obtain the ROI of VR (Figure 3(d)). Finally, the ROI of MR was created by subtracting the ROI of VR from WLR (Figure 3(e)).

Spectral data processing

Spectral pretreatment methods. In hyperspectral image acquisition, dark current, stray light and other factors inevitably produce noise or baseline offset. So, it is necessary to preprocess the spectra to reduce the noise and minimize the influence of these factors on the discriminant model.^{22,23} For this purpose, the Savitzky-Golay smoothing filter (SG smoothing), first derivative (first Der), and standard normalized variable (SNV) were used to preprocess the hyperspectral data to eliminate random noises.

The SG smoothing algorithm is a standard preprocessing method for spectral analysis and has been widely used in data smoothing.²⁴ The SG smoothing is a weighted average

algorithm of a moving window, whose weight is obtained by the least-squares fitting of a given high-order polynomial. In this experiment, the window size of SG smoothing is set as 17 and the original data were fitted through a cubic polynomial for reducing the random noise in spectral data. First Der is commonly used for baseline correction, background elimination, and spectral resolution preprocessing in spectral analysis.²⁵ It can eliminate the effects of baseline changes and background interference. SNV is mainly used to eliminate the influence of solid particles, surface scattering and optical path change on the NIR diffuse reflectance spectrum.²⁶ SNV pretreatment can reduce the influence of optical path change on spectral absorption caused by the uneven size of single leaf and uneven surface of accumulated leaves.

Spectral feature extraction. Principal component analysis (PCA) is widely used for dimension reduction and band selection of hyperspectral images.^{27–29} To eliminate the

redundant part of the data and improve the operational efficiency of the model, PCA was used to extract the important features of reflectance spectra of the leaves. It can decompose high dimensional data into new orthogonal axis, known as principal components (PCs). The combination of a few PCs can represent original spectral data with minimal loss of information, which can be used to reveal the different physiological characteristics between hybrid cherry tomato species. In addition, the peaks or valleys in the loadings plot of the PCs show the importance of corresponding wavelengths or bands in the spectral matrix.³⁰ After extracting the feature, classification models were developed by using machine learning algorithms.

Classification models

Support vector machine. SVM is a classical classification and regression algorithm with excellent performance in hyperspectral data classification.³¹ Extensive studies indicated that SVM shows good performance in the application of pattern recognition and classification based on HSI technology.^{32–34} For classification, SVM attempts to find the optimal hyperplane that can separate the closest samples of two classes by maximizing the margin. Different kernel functions such as linear kernel, Sigmoid kernel, and radial basis function kernel have been developed for SVM to deal with different classification tasks. In this work, the grid search approach was applied to select the optimal kernel for SVM model and optimize the hyperparameters such as soft-margin constant (C).³⁵

Random forest. Random forests (RF) can be applied to address the problems of high-dimensional data and high feature-to-instance ratio.³⁶ RF combines several individual decision trees at the training stage, in which randomly selected features are used to split a leaf on each tree. The RF classifier can be applied in the hyperspectral data for its insensitivity to high-dimensional features. Different hyperparameters such as the number of estimators, minimum samples split, and the maximum depth of the tree are considered and optimized by the grid search approach.

Linear discriminant analysis. LDA, also known as Fisher LDA, is a classic algorithm commonly used for classification.³⁷ Its core idea is to project high-dimensional samples into the best discriminant vector space. LDA calculates the between-class and within-class variance and maximizes the ratio of the between-class variance and the within-class variance in the new subspace, which make the model have the best separability in the space.³⁸ In this paper, the least squares solution is used as the solver for LDA and the shrinkage parameter of LDA is optimized by the grid search approach.

Separability and similarity analysis of spectra. To evaluate the separability and similarity of four cherry tomato species, the Sum of Intra-class Distance (SID), Average Distance Between classes (ADB), and the ratio of SID/ADB were calculated.³⁹ The SID is used to reflect the intra-class similarity and calculated as the following equation

$$SID = \sum_{i=1}^m D(C_i) \quad (2)$$

where $D(C_i) = 1/n_{ci} \sqrt{\sum_{j=1}^{n_{ci}} \sum_{k=1}^d (x_{jk} - c_{ik})^2}$ denotes the intra-class distance of classes C1, C2, C3, C4, which respectively represents the specie *Mi*, *Huo*, *Mo*, and *Bai*; m represents the number of classes which equals four in this work; n_{ci} indicates the number of samples of the i class; d represents the dimension of the data, x_{jk} indicates the value of the k dimension for the j sample, and c_{ik} represents the class center (average spectral curve) c_i on the k dimension of the class.

The ADB is used to reflect the interclass separation of species *Mi*, *Huo*, *Mo*, and *Bai* and is calculated as the following equations

$$ADB = \sum_{i=1}^{m-1} \sum_{j=i+1}^m D(C_i - C_j) / C_m^2 \quad (3)$$

where $D(C_i - C_j) = \sqrt{\sum_{k=1}^d (C_{ik} - C_{jk})^2}$ represents the inter-class distance between C_i and C_j of any two classes, and C_{ik} represents the class center of class C_i on the k dimension.

In addition, the ratio of SID/ADB is used to reflect the separation of species when different ROIs are used.³⁹ The smaller the SID/ADB, the greater separation between species.

Model evaluation index. Sensitivity, specificity, and precision are important indicators to evaluate the performance of the classification model. And they are calculated as the following equations⁴⁰

$$Sensitivity = \frac{TP}{TP + FN} \quad (4)$$

$$Specificity = \frac{TN}{TN + FN} \quad (5)$$

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

where TP, TN, FP, and FN stand for true positives, true negatives, false positives, and false negatives, respectively. For a specific class of specie *Mi*, TP, TN, FP, and FN can be explained as: TP is the number of samples correctly classified as specie *Mi*; TN is the number of samples belonging to the other species such as *Mo*, *Bai*, and *Huo*; FN denotes the number of samples belonging to specie *Mi* which are misclassified as other species; FP denotes the number of samples of species *Mo*, *Bai* and *Huo* that are misclassified as specie *Mi*.

Results and discussion

Spectral analysis of different regions of interests

Figure 4(a–c) shows the reflectance spectra for different ROIs (WLR, VR, MR) of all the leaves and Figure 4(d)

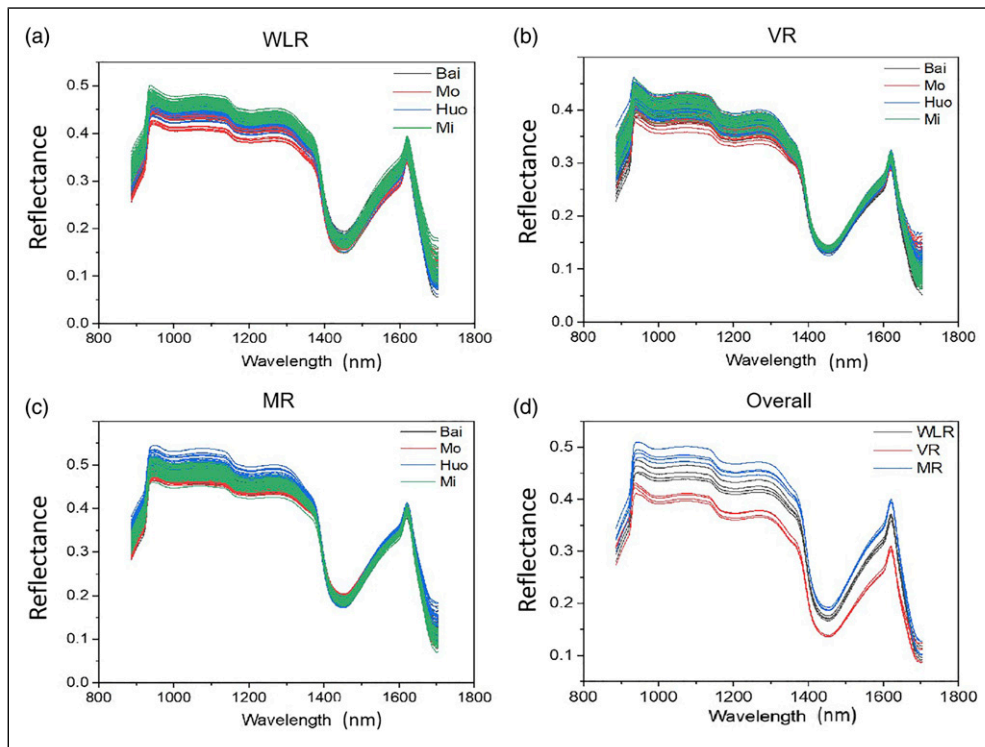


Figure 4. Reflectance spectra of different ROIs for all leaves (a): WLR-ROI; (b): VR-ROI; (c): MR-ROI; (d): Reflectance spectra for four leaves of different species.

shows the reflectance spectra of three ROIs of four leaves, each of which was randomly selected from one of the cherry tomato plant species. It was observed that the reflectance spectra show valleys around 980 nm, 1160 nm, and 1450 nm. The valley around 980 nm and 1160 nm mainly represents the absorption of symmetric and asymmetric stretching modes of the water molecules.⁴¹ And the absorption at 1450 nm may be related to the first harmonic O–H in water and the absorption of carbohydrates and protein.⁴²

By comparing the reflectance spectra extracted from different ROIs, it can be observed that the reflectance spectra of ROI-MR are higher than that of the other two ROIs. This may be attributed to the different structures between the MR and VR of the leaves.⁴³ The leaf veins are tubular structures composed of fibers, which increases the scattering effects and reduces the light reflected back to the hyperspectral camera. On the other hand, the mesophyll of the leaf is flatter and composed of many loosely arranged irregularly-shaped mesophyll cells⁴⁴ with large intercellular space, which renders higher reflectance.

To further evaluate the difference in optical behavior between the three ROIs for the leaves, the intra-class distance of four species (i.e. four classes), the inter-class distance between the four species, SID, ADB, and SID/ADB were calculated and the results are shown in Figure 5. It was observed that the inter-class distance for MR-ROI is much larger than that for the other two ROIs while the intra-class distance of MR-ROI is the smallest. The larger the inter-class distance, the greater the differences between the species. The smallest intra-class distance for the MR-ROI means that the reflectance spectra are the most concentrated. The ratio of SID/ADB can more intuitively reflect the

advantage of ROI-MR since the value of SID/ADB is much smaller than that of the other two ROIs. The smaller the ratio of SID/ADB, the larger the distance between classes, and so the better the separability of species.³⁹ These results demonstrate that the reflectance spectra from MR-ROI have higher intra-class similarity and larger inter-class segregation, which renders it more suitable for establishing classification models.

Modeling analysis

Machine learning algorithms such as SVM, RF, and LDA are employed to build the classification model for different hybrid species of cherry tomatoes. Prior to the model construction, the data set (400 samples) was divided into training and test sets with a ratio of 7:3 by using the train-test-split method in Scikit-learn. As a result, 280 samples were randomly selected as the training set and the remaining 120 samples were used as the test set. Then, K-fold cross-validation (CV) with $k = 5$ was implemented by a grid search approach to find the optimal hyperparameters for each classifier.⁴⁵ In this method, the training set was split into five groups of which each was used in turn to evaluate the model fitted on the other four groups of the training set. Thus, 1/5 of the samples in the training set were set as a validation set and the other 4/5 of the samples were set as a calibration set. Finally, the model with optimal hyperparameters was further evaluated on the test set (120 samples).⁴⁵

Table 2 shows the classification results for SVM, RF and LDA models constructed based on the full-wavelength reflectance spectra from different ROIs of the leaves. It can be found that the classification accuracy for all three

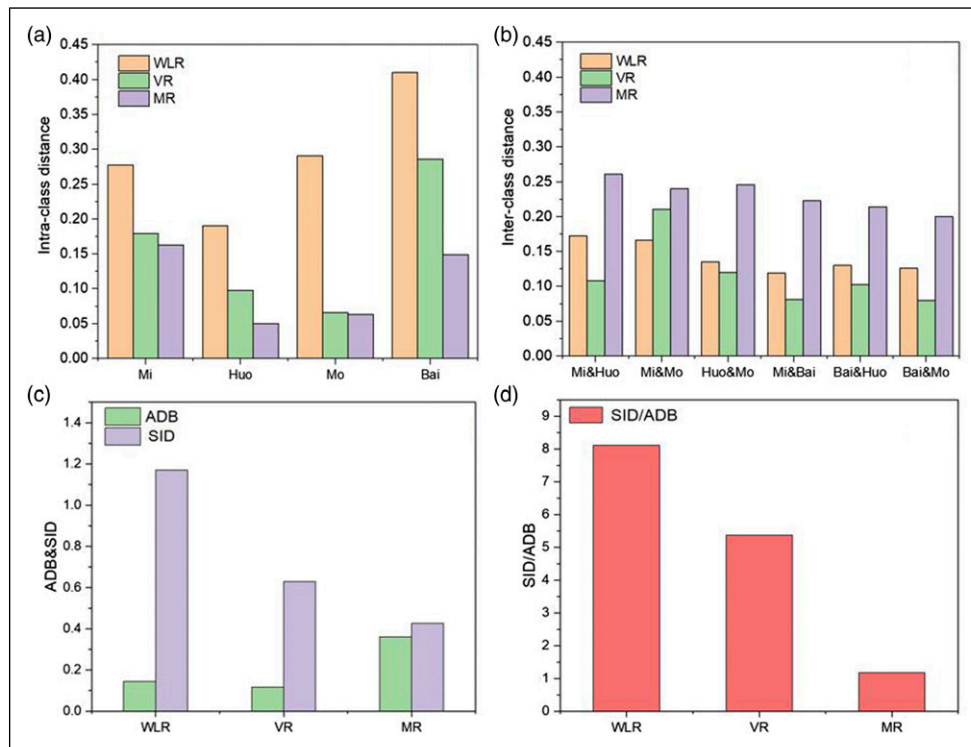


Figure 5. Separability and similarity analysis of ROI-WLR, ROI-VR, ROI-MR. (a) Intra-class distance within species; (b) inter-class distance between species; (c) ADB and SID of different ROIs; (d) SID/ADB of ROI-WLR, ROI-VR, and ROI-MR.

models always reaches the highest when MR was selected as ROI. For the LDA classifier with full wavelengths in modeling, the classification accuracy of the test set reaches a high value of 0.840 when MR is used as ROI. This value was much higher than that when ROI-WLR (0.760) and ROI-VR (0.807) were used. Similar results can be found when SVM and RF are selected as the classifier. The relatively higher accuracy for the ROI-MR occurs because the reflectance spectra for the ROI-MR have higher intra-class similarity and larger inter-class separability than other ROIs as discussed in the previous section (see Figure 5). This allows the species of cherry tomato plant to be easily classified and then improves the performance of the model.

PCA was used to improve the performance of the classification models by dimension reduction and spectral feature extraction. For the ROI-WLR, ROI-VR and ROI-MR, 20 PCs, 18 PCs and 15 PCs were respectively selected by

PCA to obtain a total contribution rate of 99.9% explained variance. Based on these extracted features, the classification models were reconstructed and the results are shown in Table 2. The accuracy of the SVM, RF and LDA models was significantly improved and the test accuracy of the LDA model reach the highest value of 0.94 when ROI-MR was used. The loadings for the first five PCs of PCA which explains the most variance in the spectra was calculated and the results are shown in Figure 6. From the loadings plot, 905 nm, 950 nm, 1060 nm, 1160 nm, 1400 nm, and 1450 nm were selected as the sensitive wavelengths since obvious peaks or valleys occur in the loadings plot. The peak at 905 nm, 1060 nm and the valley at 1160 nm represents the stretching and deformation of C-H.^{40,46} The wavelengths at 1400 nm also include a vibration of C-H caused by water, lipid, and protein, which can be explained by PC4 and PC5.⁴⁷ And the wavelength at 1450 nm is related to the first harmonic of O-H and the absorption of carbohydrates and protein, which can be explained by PC2, PC3, and PC4.^{48,49}

The results show that the different internal components such as protein and water are one of the reasons that the four varieties can be distinguished by hyperspectral technology combined with machine learning. In other words, differences occur in internal material composition for the leaves of four cherry tomato plants with different species. The differences result from the specie variation of the four tomato plants since all the plants were grown in the same environment with the same nutrient and water supply during planting. Previous studies also proposed that the metabolite composition such as water content is different between species under the same growing environment.^{50,51}

To further improve the accuracy of the model, spectral preprocessing methods such as SG smoothing, SNV, and

Table 2. Performance of classification models based on full wavelength spectra and PCA feature extraction.

ROI	Classifier	Full wavelength		PCA	
		Train	Test	Train	Test
WLR	SVM	0.71	0.64	0.86	0.84
	RF	0.73	0.65	0.79	0.73
	LDA	0.80	0.76	0.92	0.89
VR	SVM	0.80	0.71	0.80	0.79
	RF	0.71	0.70	0.83	0.77
MR	LDA	0.82	0.81	0.94	0.89
	SVM	0.81	0.80	0.86	0.84
	RF	0.80	0.75	0.90	0.82
	LDA	0.86	0.84	0.98	0.94

first Der were used to preprocess the raw reflectance spectra. After spectra preprocessing, PCA was used to extract spectral features and classification models were then re-established. The results of the models are shown in Table 3. It can be seen that SNV can lead to higher improvement in the model accuracy than other preprocessing methods of SG smoothing and first Der. When the raw spectral data of ROI-MR was used in modeling, the accuracy of the test set for the SVM, RF and LDA classifier model was about 0.84, 0.82, and 0.94, respectively. The values are rapidly increased up to 0.93, 0.91 and 0.97 when the spectral data is preprocessed by SNV, which are much higher than that when being preprocessed by SG smoothing and first Der (see Table 3). This happens because SNV can eliminate the influence of scattering of solid particles, surface scattering and optical path change on NIR spectra. In addition, SNV also reduces the effects of scattering and the difference in the global intensities of the signals.⁴⁰ The superiority of SNV is further confirmed by using a T-distributed stochastic neighbor embedding (T-SNE) map,⁵² which can visualize the features of the spectral datasets after different preprocessing methods. As shown in Figure 7, it can be clearly seen that the aggregation degree of samples with the SNV preprocessing method is higher than that with SG smoothing and first Der. The results also clearly show that using SNV

preprocessing can effectively improve the accuracy of the classification model.

The investigation above demonstrates that the LDA classifier reached the highest test accuracy of 0.973 when the reflectance spectra from MR-ROI are utilized, preprocessed by SNV and dimensionally reduced by PCA. The accuracy is much higher than that (0.93 and 0.91) for SVM ($C = 5$, kernel = linear) and RF (number of estimators = 30, maximum depth = 4, minimum samples split = 10) of which the hyperparameters are optimized by the grid search approach from Scikit-learn. Similar results were also observed when the spectra from ROI-WLR and ROI-VR are utilized in modeling and processed by other methods of SG smooth and first Der. These findings demonstrate that LDA is more suitable to classify the hybrid species of cherry tomato in this study.

The confusion matrix (aka error matrix),⁵³ is calculated to compare whether the classification result was the same as the actual specie of cherry tomato. Based on the confusion matrix, other evaluation criteria such as precision, sensitivity, and specificity were calculated and the results are shown in Table 4. The confusion matrix, precision, sensitivity, and specificity, all implied that LDA was a more accurate classifier than others. For the LDA and SVM model, the specie *Bai* as the most difficult to classify among the four species since the precision and specificity are relatively low. This happens because the inter-class distance between *Bai* and other species is relatively small due to the similar spectral features between *Bai* and other species (see Figure 5). Therefore, the species *Bai* is easier to mix with other varieties, which decreases the accuracy of the models.

Friedman test is applied to analyze whether there is a significant difference between the results of the three classification classifiers (SVM, RF, and LDA).⁵⁴ For this purpose, the train-test-split method with five different random seeds was used to divide the entire data into five pairs of the training set and test set with a ratio of 7:3. Based on each pair of the training set and test set, classification models based on SVM, RF and LDA were constructed and the results are used in Friedman test to evaluate the significant difference between the three classifiers. The average test accuracy for the classification models (referred to as 5-times-test) is shown in Table 4. In Friedman test, the null hypothesis is set as the three classification algorithm conclusions are not significantly different, and the significance level α of the Friedman test is set to 0.05. The result shows that p -value ($p = 0.032$)

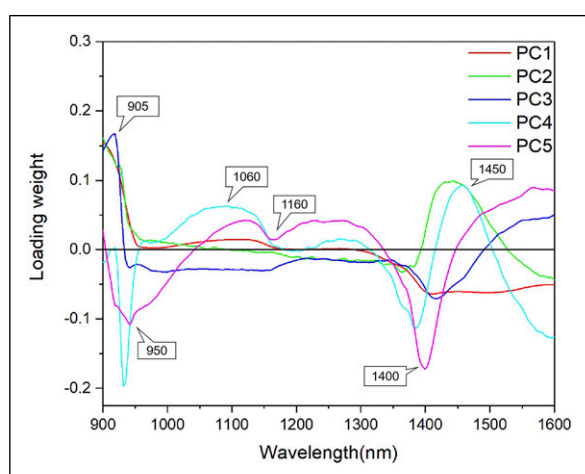


Figure 6. Loadings plot for the first five PCs of PCA.

Table 3. Performance of classification models with different spectral processing methods

Pretreatment		Raw spectral		SG smoothing		1st Der		SNV	
ROI	Classifier	Training	Test	Training	Test	Training	Test	Training	Test
WLR	SVM	0.86	0.84	0.88	0.84	0.78	0.72	0.9	0.89
	RF	0.79	0.73	0.86	0.82	0.91	0.84	0.87	0.79
	LDA	0.92	0.89	0.90	0.87	0.98	0.92	0.96	0.93
VR	SVM	0.80	0.79	0.83	0.76	0.78	0.69	0.86	0.77
	RF	0.83	0.77	0.84	0.76	0.85	0.90	0.82	0.77
	LDA	0.94	0.89	0.93	0.86	0.95	0.87	0.98	0.94
MR	SVM	0.86	0.843	0.87	0.81	0.67	0.62	0.95	0.93
	RF	0.89	0.824	0.91	0.86	0.89	0.87	0.996	0.91
	LDA	0.98	0.94	0.98	0.96	0.98	0.96	0.998	0.97

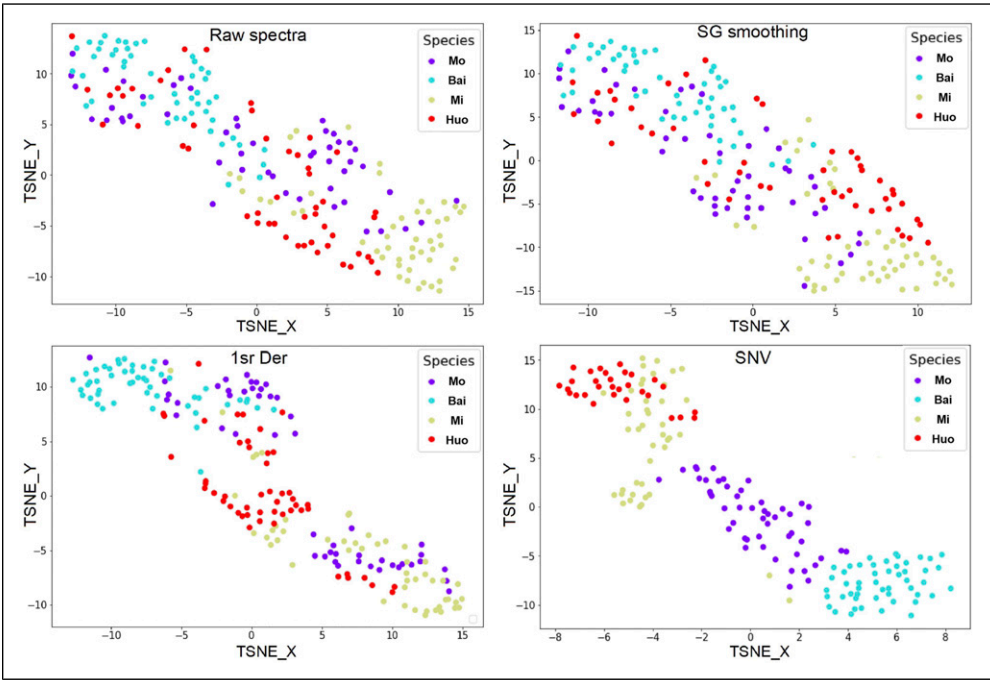


Figure 7. T-SNE map of samples with reflectance spectra preprocessed by different methods.

Table 4. Classification metrics for each classifier with SNV preprocessing

ROI	Classifier	Accuracy			Reference	Predicted				Precision	Sensitivity	Specificity
		Training	Test	5-times-test		Mi	Mo	Bai	Huo			
MR	SVM	0.95	0.93	0.93 ± 0.03	Mi	24	0	2	0	0.92	0.92	0.97
					Mo	0	21	1	1	0.91	0.91	0.97
					Bai	1	2	20	1	0.87	0.83	0.96
					Huo	1	0	0	26	0.93	0.96	0.98
	RF	0.996	0.91	0.90 ± 0.03	Mi	21	0	1	0	0.75	0.95	0.91
					Mo	2	19	0	0	0.96	0.90	0.98
					Bai	3	1	21	1	0.95	0.81	0.98
					Huo	2	0	0	29	0.97	0.94	0.98
	LDA	0.998	0.97	0.97 ± 0.02	Mi	23	0	0	0	1.0	1.0	1.0
					Mo	0	24	1	0	1.0	0.96	1.0
					Bai	0	0	23	0	0.96	1.0	0.98
					Huo	0	0	0	29	1.0	1.0	1.0

$< \alpha$, which means that the classification results of SVM, RF and LDA have significant differences.

Conclusions

This study proposed an effective approach based on NIR-HIS technology and machine learning to classify the species of cherry tomato plants. By analyzing the separability and similarity of reflectance spectra of different ROIs, it was found that MR-ROI was more suitable for establishing classification models than the WLR-ROI and VR-ROI. It also demonstrated that SNV pretreatment and PCA feature extraction method can significantly improve the accuracy of the classification models. Analysis based on PCA demonstrates that the difference occurs in internal material composition for the leaves of four cherry tomato plants with different species, enabling the models being able to distinguish the species. When the reflectance spectra from MR-ROI are utilized, preprocessed by SNV and dimensionally

reduced by PCA, the LDA classification model reaches the highest test accuracy of 0.97, which is much higher than that for the SVM and RF models. The high classification accuracy indicates that NIR HSI technology combined with machine learning is an efficient approach to classify hybrid cherry tomatoes rapidly and non-destructively. It will be significant in the future to employ the NIR technology to determine internal material content and classify seedlings for various species of cherry tomato plants.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by a grant from the Key-Area Research and

Development Program of Guangdong Province (2019B020219002, 2019B020214005). Leading talents of Guangdong province program (2016LJ06G689). Ministry of Education, China - 111 Project (D18019). Science and Technology Promoting Agriculture Project of Guangdong Provincial Department of Agriculture and Rural Affairs (2021KJ383).

ORCID iD

Songhao Li  <https://orcid.org/0000-0002-4763-3927>

References

1. Tieman D, Zhu G, Resende MFR, et al. Plant science a chemical genetic roadmap to improved tomato flavor. *Science* 2017; 355: 391–394.
2. Aguirre NC, López W, Orozco-Cárdenas M, et al. Use of microsatellites for evaluation of genetic diversity in cherry tomato. *Bragantia* 2017; 76: 220–228.
3. Kavitha P, Shivashankara KS, Rao VK, et al. Genotypic variability for antioxidant and quality parameters among tomato cultivars, hybrids, cherry tomatoes and wild species: variability for antioxidant quality parameters in tomato. *J Sci Food Agric* 2014; 94: 993–999.
4. Rapa M, Ciano S, Ruggieri R, et al. Bioactive compounds in cherry tomatoes (*Solanum lycopersicum* var. *cerasiforme*): cultivation techniques classification by multivariate analysis. *Food Chem* 2021; 355: 129630.
5. Chen J, Kang S, Du T, et al. Quantitative response of greenhouse tomato yield and quality to water deficit at different growth stages. *Agric Water Manag* 2013; 129: 152–162.
6. Ddamulira G, Idd R, Namazzi S, et al. Nitrogen and potassium fertilizers increase cherry tomato height and yield. *J Agric Sci* 2019; 11: 48.
7. Khulal U, Zhao J, Hu W, et al. Nondestructive quantifying total volatile basic nitrogen (TVB-N) content in chicken using hyperspectral imaging (HSI) technique combined with different data dimension reduction algorithms. *Food Chem* 2016; 197: 1191–1199.
8. Peng Y and Lu R. Analysis of spatially resolved hyperspectral scattering images for assessing apple fruit firmness and soluble solids content. *Postharvest Biol Technol* 2008; 48: 52–62.
9. Khan A, Vibhute AD, Mali S, et al. A systematic review on hyperspectral imaging technology with a machine and deep learning methodology for agricultural applications. *Ecol Inform* 2022; 69: 101678.
10. Singh L, Mutanga O, Mafongoya P, et al. Hyperspectral remote sensing for foliar nutrient detection in forestry: a near-infrared perspective. *Remote Sens Appl Soc Environ* 2022; 25: 100676.
11. Yang R and Kan J. Classification of tree species at the leaf level based on hyperspectral imaging technology. *J Appl Spectrosc* 2020; 87: 184–193.
12. Kanayama H, Ma T, Tsuchikawa S, et al. Cognitive spectroscopy for wood species identification: near infrared hyperspectral imaging combined with convolutional neural networks. *Analyst* 2019; 144: 6438–6446.
13. Furlanetto RH, Moriwaki T, Falcioni R, et al. Hyperspectral reflectance imaging to classify lettuce varieties by optimum selected wavelengths and linear discriminant analysis. *Remote Sens Appl Soc Environ* 2020; 20: 100400.
14. Liu KH, Yang MH, Huang ST, et al. Plant species classification based on hyperspectral imaging via a lightweight convolutional neural network model. *Front Plant Sci* 2022; 13: 855660.
15. Yuan S, Song G, Huang G, et al. Reshaping hyperspectral data into a two-dimensional image for a CNN model to classify plant species from reflectance. *Remote Sens* 2022; 14: 3972.
16. Li J, Zhou JM, Duan ZQ, et al. Effect of CO₂ Enrichment on the growth and nutrient uptake of tomato seedlings. *Pedosphere* 2007; 17: 343–351.
17. Dmitriev PA, Kozlovsky BL, Kupriushkin DP, et al. Identification of species of the genus *Acer* L. using vegetation indices calculated from the hyperspectral images of leaves. *Remote Sens Appl Soc Environ* 2022; 25: 100679.
18. Diago MP, Fernandes AM, Millan B, et al. Identification of grapevine varieties using leaf spectroscopy and partial least squares. *Comput Electron Agric* 2013; 99: 7–13.
19. Mishra P, Nordon A, Mohd Asaari MS, et al. Fusing spectral and textural information in near-infrared hyperspectral imaging to improve green tea classification modelling. *J Food Eng* 2019; 249: 40–47.
20. Gao D, Li M, Zhang J, et al. Improvement of chlorophyll content estimation on maize leaf by vein removal in hyperspectral image. *Comput Electron Agric* 2021; 184: 106077.
21. Ma Y, Huang M, Yang B, et al. Automatic threshold method and optimal wavelength selection for insect-damaged vegetable soybean detection using hyperspectral images. *Comput Electron Agric* 2014; 106: 102–110.
22. Pu YY and Sun DW. Vis–NIR hyperspectral imaging in visualizing moisture distribution of mango slices during microwave-vacuum drying. *Food Chem* 2015; 188: 271–278.
23. Su WH and Sun DW. Evaluation of spectral imaging for inspection of adulterants in terms of common wheat flour, cassava flour and corn flour in organic avatar wheat (*triticum* spp.) flour. *J Food Eng* 2017; 200: 59–69.
24. Savitzky A and Golay MJE. *Smoothing and differentiation of data by simplified least squares procedures*. 2002. ACS Publications.
25. Vidal M and Amigo JM. Pre-processing of hyperspectral images: essential steps before image analysis. *Chemom Intell Lab Syst* 2012; 117: 138–148.
26. Guo Y, Ni Y and Kokot S. Evaluation of chemical components and properties of the jujube fruit using near infrared spectroscopy and chemometrics. *Spectrochim Acta A Mol Biomol Spectrosc* 2016; 153: 79–86.
27. Amigo JM, Martí I and Gowen A. Chapter 9 - hyperspectral imaging and chemometrics: a perfect combination for the analysis of food structure, composition and quality. 2022. In: F Marini (ed). *Data handling in science and technology*. Elsevier, pp. 343–370.
28. Thien Pham Q and Liou NS. The development of on-line surface defect detection system for jujubes based on hyperspectral images. *Comput Electron Agric* 2022; 194: 106743.
29. Wold S, Esbensen K and Geladi P. Principal component analysis. *Chemom Intell Lab Syst* 1987; 2: 37–52.
30. Chu B, Yu K, Zhao Y, et al. Development of noninvasive classification methods for different roasting degrees of

- coffee beans using hyperspectral imaging. *Sensors* 2018; 18: 1259.
31. Bazi Y and Melgani F. Toward an optimal SVM classification system for hyperspectral remote sensing images. *IEEE Trans Geosci Remote Sens* 2006; 44: 3374–3385.
 32. Zhang X, Sun J, Li P, et al. Hyperspectral detection of salted sea cucumber adulteration using different spectral preprocessing techniques and SVM method. *LWT* 2021; 152: 112295.
 33. Xia J, Yang Y, Cao H, et al. Visible-near infrared spectrum-based classification of apple chilling injury on cloud computing platform. *Comput Electron Agric* 2018; 145: 27–34.
 34. Wu Y and Zhang X. Object-based tree species classification using airborne hyperspectral images and LiDAR data. *Forests* 2020; 11: 32.
 35. Zhang L, Sun J, Zhou X, et al. Classification detection of saccharin jujube based on hyperspectral imaging technology. *J Food Process Preserv* 2020; 44. DOI: [10.1111/jfpp.14591](https://doi.org/10.1111/jfpp.14591)
 36. Kuncheva LI, Rodriguez JJ, Plumpton CO, et al. Random subspace ensembles for fMRI classification. *IEEE Trans Med Imaging* 2010; 29: 531–542.
 37. Gao TF and Liu CL. High accuracy handwritten Chinese character recognition using LDA-based compound distances. *Pattern Recognit* 2008; 41: 3442–3451.
 38. Baek I, Kim MS, Cho BK, et al. Selection of optimal hyperspectral wavebands for detection of discolored, Diseased Rice Seeds. *Appl Sci* 2019; 9: 1027.
 39. Li X, Liang W, Zhang X, et al. A cluster validity evaluation method for dynamically determining the near-optimal number of clusters. *Soft Comput* 2020; 24: 9227–9241.
 40. Sharma S, Sumesh KC and Sirisomboon P. Rapid ripening stage classification and dry matter prediction of durian pulp using a pushbroom near infrared hyperspectral imaging system. *Measurement* 2022; 189: 110464.
 41. Salzer R. Practical guide to interpretive near-infrared spectroscopy: by Jerry Workman, Jr. and Lois Weyer. *Angew Chem Int Ed* 2008; 47: 4628–4629.
 42. Xiaobo Z, Jiewen Z, Povey MJW, et al. Variables selection methods in near-infrared spectroscopy. *Anal Chim Acta* 2010; 667: 14–32.
 43. Sims DA and Gamon JA. Relationships between leaf pigment content and spectral reflectance across a wide range of species, leaf structures and developmental stages. *Remote Sens Environ* 2002; 81: 337–354.
 44. Slaton MR, Hunt RE and Smith WK. Estimating near-infrared leaf reflectance from leaf structural characteristics. *Am J Bot* 2001; 88: 278–284.
 45. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *Mach Learn PYTHON* 2011; 6: 2825–2830.
 46. Chen S, Zhang F, Ning J, et al. Predicting the anthocyanin content of wine grapes by NIR hyperspectral imaging. *Food Chem* 2015; 172: 788–793.
 47. Prananto JA, Minasny B and Weaver T. Chapter one - near infrared (NIR) spectroscopy as a rapid and cost-effective method for nutrient analysis of plant leaf tissues. 2020. In: DL Sparks (ed). *Advances in agronomy*. Academic Press, pp. 1–49.
 48. Zhu S, Feng L, Zhang C, et al. Identifying freshness of spinach leaves stored at different temperatures using hyperspectral imaging. *Foods* 2019; 8: 356.
 49. Wang Y-J, Jin G, Li LQ, et al. NIR hyperspectral imaging coupled with chemometrics for nondestructive assessment of phosphorus and potassium contents in tea leaves. *Infrared Phys Technol* 2020; 108: 103365.
 50. Zamljen T, Jakopič J, Hudina M, et al. Influence of intra and inter species variation in chilies (capsicum spp.) on metabolite composition of three fruit segments. *Sci Rep* 2021; 11: 4932.
 51. Huang Z, Sanaeifar A, Tian Y, et al. Improved generalization of spectral models associated with Vis-NIR spectroscopy for determining the moisture content of different tea leaves. *J Food Eng* 2021; 293: 110374.
 52. Miao A, Zhuang J, Tang Y, et al. Hyperspectral image-based variety classification of waxy maize seeds by the t-SNE model and procrustes analysis. *Sensors* 2018; 18: 4391.
 53. Lv W and Wang X. Overview of hyperspectral image classification. *J Sens* 2020; 2020: 48172344.
 54. Demsar J. Statistical comparisons of classifiers over multiple data sets. *J Machin Learn Res* 2006; 7: 1–30.