applied sciences

MDPI

Article

Variety Identification of Chinese Walnuts Using Hyperspectral Imaging Combined with Chemometrics

Hongzhe Jiang [1,2,3,*], Liancheng Ye [1], Xingpeng Li [1] and Minghong Shi [1]

[1] College of Mechanical and Electronic Engineering, Nanjing Forestry University, Nanjing 210037, China; ylctx999@163.com (L.Y.); lxp@njfu.edu.cn (X.L.); mhshi@njfu.edu.cn (M.S.)
[2] Key Laboratory of on Site Processing Equipment for Agricultural Products, Ministry of Agriculture and Rural Affairs, Hangzhou 310058, China
[3] College of Biosystems Engineering and Food Science, Zhejiang University, 866 Yuhangtang Road, Hangzhou 310058, China
* Correspondence: jianghongzhe@njfu.edu.cn

Abstract: Chinese walnuts have extraordinary nutritional and organoleptic qualities, and counterfeit Chinese walnut products are pervasive in the market. The aim of this study was to investigate the feasibility of hyperspectral imaging (HSI) technique to accurately identify and visualize Chinese walnut varieties. Hyperspectral images of 400 Chinese walnuts including 200 samples of Ningguo variety and 200 samples of Lin'an variety were acquired in range of 400–1000 nm. Spectra were extracted from representative regions of interest (ROIs), and principal component analysis (PCA) of spectra showed that the characteristic second principal component (PC_2) was potentially effective in variety identification. The PC transformation was also conducted to hyperspectral images to make an exploratory visualization according to pixel-wise PC scores. Three different modeling methods including partial least squares-discriminant analysis (PLS-DA), *k*-nearest neighbor (KNN), and support vector machine (SVM) were individually employed to develop classification models. Results indicated that raw full spectra constructed PLS-DA model performed best with correct classification rates (CCRs) of 97.33%, 95.33%, and 92.00% in calibration, cross-validation, and prediction sets, respectively. Successful projects algorithm (SPA), competitive adaptive reweighted sampling (CARS), and PC loadings were individually used for effective wavelengths selection. Subsequently, simplified PLS-DA model based on wavelengths selected by CARS yielded the best 96.33%, 95.67% and 91.00% CCRs in the three sets. This optimal CARS-PLS-DA model acquired a sensitivity of 93.62%, a specificity of 88.68%, the area under the receiver operating characteristic curve (AUC) value of 0.91, and Kappa coefficient of 0.82 in prediction set. Classification maps were finally generated by classifying the varieties of each pixel in multispectral images at CARS-selected wavelengths, and the general variety was then readily discernible. These results demonstrated that features extracted from HSI had outstanding ability, and could be applied as a reliable tool for the further development of an on-line identification system for Chinese walnut variety.

Keywords: hyperspectral imaging; Chinese walnuts; variety classification; identification models; visualization

Citation: Jiang, H.; Ye, L.; Li, X.; Shi, M. Variety Identification of Chinese Walnuts Using Hyperspectral Imaging Combined with Chemometrics. *Appl. Sci.* **2021**, *11*, 9124. https://doi.org/10.3390/app11199124

Academic Editor: Sungho Kim

Received: 19 July 2021
Accepted: 26 September 2021
Published: 30 September 2021

1. Introduction

Walnuts are nutrient-dense foods, which are extremely rich in unsaturated fatty acids and phytochemicals like proteins and antioxidants [1]. Moreover, several studies have demonstrated that the consumption of walnuts is closely related to the reduction of metabolic syndrome and heart disease risk [2,3]. In the human diet, walnuts are widely consumed or used to produce liquor or oil. In addition, they are often used as food additives in a variety of foods such as baked items, ice cream, pastries, etc. [4]. The health benefits and extensive consumption of walnuts have led to the establishment of an important walnut market.

Chinese walnuts (*Carya cathayensis* Sarg.) are more and more popular among consumers due to their special organoleptic characteristics and benefits to human health. Quality characteristics are strongly influenced by the Chinese walnut varieties. Certain Chinese walnuts from small harvesting areas such as the Lin'an walnut (Lin'an city is known as the capital of Chinese walnut) have high nutrient values, delicious taste and commonly fetch high prices. The different prices based on different varieties provide a financial incentive for undeclared adulteration or variety substitution by unscrupulous businessmen [5]. Therefore, the identification of Chinese walnut varieties is becoming a very important task to provide consumers exact information about the Chinese walnut products they purchase.

Morphological detection based on appearance features is frequently used in identifying walnut varieties. However, this method is not only subjective but also time-consuming. Several other methods have been proved to be feasible in the identification, such as gas chromatography (GC) [6], high performance liquid chromatography (HPLC) [7], nuclear magnetic resonance (NMR) [8], etc. However, these above techniques often require long-time sample preparation, and are high-cost, laborious and difficult for performance interpretation.

In recent years, fast imaging and spectroscopic techniques have been applied as non-destructive alternatives to detect various quality and safety traits of agricultural products [9–11]. Among them, visible and near-infrared spectroscopy (NIRS) has been widely investigated as an effective tool in evaluating the quality of nuts [1,12]. Yi et al. [13] collected the NIR reflectance spectra of walnut kernel, and successfully predicted the contents of moisture, fat, and protein. Wang et al. [14] investigated the feasibility of NIR spectral data to detect the internal moisture of freshly harvested in-shell walnuts during drying, and results showed that NIRS could be used to sort the walnuts into different moisture degrees. However, one of the main shortcomings is that NIRS is specially limited to the single-point detection of pre-selected areas in homogeneous products, which can hardly represent the whole nut sample [15].

The hyperspectral imaging (HSI) technique simultaneously integrates conventional spectroscopy and imaging to provide both spectral and spatial information from one target [16]. The hyperspectral image consists of a large amount of information (a 3D hypercube) which can be used to characterize the food more reliably than single imaging or spectroscopy technique. This leads to its special ability for visualizing the category or chemical composition distribution. In the last decades, HSI has received ample attention in agricultural products including meats [17,18], cereals [19], fruits [20], vegetables [21], and nuts [22]. In the nut industry, HSI technique has been successfully applied to assess mildew damage [23], fungal contamination [24], moisture [25], and internal damage [26]. Recently, the quantitative or qualitative analysis of HSI mostly needs to be combined with chemometrics [27,28]. With the application of chemometrics such as modeling, spectral preprocessing, and image processing, the performance may be improved [29]. However, to best of our knowledge, little work has been executed on identifying or authenticating different Chinese walnut varieties using this rapid and non-destructive HSI tool combined with chemometrics.

Therefore, the specific objectives of this study are to develop and optimize methodology for identifying Ningguo and Lin'an Chinese walnut varieties by using a HSI system in the range of 400–1000 nm. Moreover, different chemometric methods including spectral pretreatments, wavelengths selection and modeling algorithms, will be tested to obtain a robust, reliable but simplified model with the optimal discriminatory ability. Consequently, the optimal simplified model will be attempted to apply back to visualize the classification maps of Chinese walnut varieties.

2. Materials and Methods

2.1. Chinese Walnut Samples

Samples of two typical Chinese walnut varieties, Lin'an and Ningguo, were prepared. The reason was that Ningguo walnuts are similar to Lin'an walnuts in both size and color, and hard to distinguish by the naked eye. The Lin'an walnuts were harvested from the Lin'an district, Hangzhou city in Zhejiang Province in China, and the Ningguo walnuts were harvested at Ningguo city in Anhui Province in China. In order to be sure about the authenticity of walnuts, Lin'an walnuts were directly purchased from farmers in Shichangcheng village (Qingliangfeng town, Lin'an district, Hangzhou city, 30.14° N, 118.95° E), and Ningguo walnuts were purchased directly from farmers in Jiaxiang Mountain (Ningguo city, 30.67° N, 119.14° E). Defective specimens, including insect-damaged samples, were picked out and eliminated in the study. A total of 400 samples including 200 Ningguo walnuts and 200 Lin'an walnuts harvested in September 2020 were randomly collected. Then, the randomly selected 300 samples including 150 samples for each variety (150 samples × 2 groups = 300 samples) were used to calibrate and cross-validate the classification models, and the residual 100 samples (50 samples × 2 groups = 100 samples) were used as a prediction set to verify the models. After the samples were transported to our laboratory, they were individually sealed in plastic sealing bags and labeled by corresponding number of 1 to 400 for convenient recording. All the samples were stored at 0–4 °C in the refrigerator until their hyperspectral images were captured.

2.2. Hyperspectral Images Collection

In this study, hyperspectral images were acquired using a Headwall Hyperspec MV.X instrument at NBL Imaging System Ltd. (Guangzhou, China). The push-broom line-scan hyperspectral system was composed of a 12-bit CMOS detector, a conveyor motivated by a motor with a speed controller, a uniform illumination unit, a computer, and an image collection software. The spectral sampling interval was 1.75 nm/pixel, and the number of spatial pixels was 1024. The protection level was IP67, and the onboard hardware included 8 GB RAM and 128 GB SSD. All the collected hyperspectral images contained 301 channels with a spectral resolution of 2 nm.

Prior to the images capture, walnut samples were gently rubbed with soft tissues to remove any residues of dust or soil possibly present. The walnut samples were then neatly placed on a dark non-reflective conveyor first. After that, hyperspectral images will be captured with the movement of a platform from left to right. In order to eliminate the noise generated by the system and the influence of surrounding environmental factors such as humidity, white and black reference images collection was conducted before the hyperspectral images were captured. A white reference image (W) was acquired using a white Teflon with the reflectivity close to 100%. Dark reference image (close to 0% reflectance, D) was obtained by completely closing camera lens with its own opaque cap. After raw hyperspectral images were collected, the calibrated hyperspectral image (C) was calculated using the following equation:

$$C = (R - D)/(W - D) \times 100\% \quad (1)$$

where C and R indicate the calibrated and raw hyperspectral image, respectively, D denotes the dark reference image, and W represents the white reference image. The subsequent analyses were conducted based on the calibrated hyperspectral image, that is, the above C.

2.3. Regions of Interest Identification

In order to extract pure spectral information of Chinese walnut samples, a region of interest (ROI) was constructed to accurately isolate the sample from background. First, channels with very high (950 nm) and very low (416 nm) reflectance intensity were selected. After that, a resulting image was formed by subtracting image at 416 nm (band 1) from image at 950 nm (band 2) using 'Band Math' function in ENVI (Vision 5.1, ITT Visual

Information Solutions, Boulder, CO, USA) software. In the resulting image, walnut sample with strong contrast in background was shown. Secondly, 'Build Mask' function was achieved in this image to build a binary mask based on a simple threshold segmentation at constant value of 0.08. The mask was subsequently applied to corresponding hyperspectral image to segment the whole sample without the background. Finally, the average spectrum of each sample was extracted by repeating the above steps. As a result, a total of 400 spectra were obtained to form a 400 × 301 spectral data cube. All the above processes were performed and analyzed in the ENVI environment.

2.4. Principal Component Analysis

Principal component analysis (PCA) is a powerful method in solving multi-collinearity or eliminating potential collinearity hidden in spectral data. In this step, PCA is applied to decompose raw spectra into a series of new variables named principal components (PCs) which are orthogonally projected. Then, the first few PCs with high interpretation representing most of the samples' information will be considered. Data variability will be observed to analyze the distribution of sample data. PC loading lines are also commonly used to investigate the relationship among the variables.

In current study, PCA was first employed to visualize the spectral data in different groups. Characteristic PC loading lines were drawn to see the useful wavelengths in identifying Chinese walnut varieties. In addition, PCA was also applied to the hyperspectral images to select the corresponding characteristic images by implementing PC transformation to the spectra of each pixel. After that, transformed PC score values of each pixel were obtained, and thus pixel-wise PC score images were formed. As a result, similarities, and differences among different samples were observed in effective PC score images. Spectra PCA was conducted in Matlab (Vision 2013b, The Mathworks Inc., Natick, MA, USA), and PC transformation was carried out in the ENVI software.

2.5. Spectral Preprocessing

The extracted raw full spectra contained external noises including scatter and transmitted light variations due to uneven particle size which were unrelated to the chemical compositions. Therefore, five appropriate different preprocessing methods were independently applied to the raw full spectra, namely standard normal variate (SNV), detrending, normalization, first-order derivative (1st derivative), and second-order derivative (2nd derivative). SNV was used to eliminate the solid particle size, surface scattering, and the change of optical path of diffuse reflection spectra [30]. Normalization was applied to eliminate multiplicative spectral effects by transforming the spectral vector into unit length [31]. Detrending was used to suppress the curvilinearity and baseline shifting following SNV application. Derivatives (1st and 2nd derivatives) were conducted to remove baseline shifts and improve resolution based on Savitzky-Golay smoothing algorithm with a gap of five points. The aim of spectral preprocessing herein was to remove phenomena in spectra, and improve the subsequent classification performance. All the preprocessings were conducted employing the Unscrambler X10.1 software (CAMO, Trondheim, Norway).

2.6. Feature Variables Screening

Two different effective methods below for feature variables screening were employed in our work. Successful projects algorithm (SPA) is a variable selection algorithm designed to eliminate the collinearity by selecting new variables with minimal redundancy [32]. Optimal wavelengths will be chosen based on the smallest root mean squared error (RMSE) of calibration set. In the SPA vector space, variables with the largest projection value on the orthogonal subspace will be retained.

Competitive adaptive reweighted sampling (CARS) selects effective wavelengths based on Monte-Carlo sampling and regression coefficients (RC) in partial least squares regression (PLSR) model [33]. In each iteration, CARS evaluates the significance of each wavelength based on absolute RC values in the PLSR model. The wavelengths with little

effect will be removed, and next iteration will start with the remaining variables until all the iterations are finished.

2.7. Modeling Methods

Partial least squares-discriminant analysis (PLS-DA) is a supervised learning feature extraction method frequently used in spectral analysis. PLS-DA predicts the class for each sample based on PLSR step [34]. The optimum number of latent variables (LVs) at minimum value of the corresponding prediction errors of sum of squares (PRESS) will be selected. In current study, to reduce the overfitting risk, calibration and prediction sets were randomly divided based on the ratio of 3:1 (300 samples vs. 100 samples). The optimized number of LVs in developed PLS-DA models were determined based on the value of RMSE under two-fold 'venetian blinds' cross-validation. The variety values herein were replaced using pseudo-variables of 1 and 2 (1 for Ningguo walnut, and 2 for Lin'an walnut).

The *k*-nearest neighbor (KNN) algorithm is a commonly used supervised pattern recognition method, which automatically describes nonlinear relationship according to the K values [35]. The general steps are (1) evaluate the distance between the sample to be classified and other samples in calibration set, (2) look for the K samples closest to the targeted sample, (3) see the classification performance of these K samples, and finally (4) take the group with the most occurrences as the category of the sample to be classified. In our study, Euclidean distance evaluation method was employed, and the range of K values was set from 1 to 10 with a cross-validation step of two based on 'venetian blinds' method.

Support vector machine (SVM) is a kernel-based nonlinear method which has been proven to exhibit good performance in spectral modeling. In SVM classification, a hyperplane is explored to segment nonlinear spectral data of prepared samples [36]. In this study, radial basis function (RBF) was used as the kernel function of SVM, which has strong ability in addressing nonlinear problems. The parameters of c (the penalty coefficient) and g (the radial width of the kernel function) for RBF-SVM were chosen in automatic optimization process. The value ranges were set to 2^{-8} to 2^{8} for both c and g searching. All the three modeling procedures were performed in Matlab software with PLS-DA, KNN and SVM toolboxes.

2.8. Model Performance Assessment

Performance of classification model was assessed using correct classification rate (CCR) calculated using the following equation:

$$\mathrm{CCR} = \frac{\mathrm{N_1}}{\mathrm{N_2}} \tag{2}$$

where CCR indicates correct classification rate, $\mathrm{N_1}$ is the number of precisely sorted samples in calibration, cross-validation, or prediction set, and $\mathrm{N_2}$ represents the corresponding total number of samples in calibration, cross-validation, or prediction set.

To further evaluate the results of selected models, other indicators including sensitivity, specificity, and Kappa coefficient were included in this study. The definitions of sensitivity and specificity are listed below:

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{3}$$

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}} \tag{4}$$

Kappa coefficient is used to indicate the differences between reference and predicted categories in the confusion matrix, which ranges from -1 to 1. The higher the Kappa coefficient is, the more reliable the classification model is. Furthermore, receiver operating characteristic (ROC) curve is usually constructed to assess and visualize the classification

performance [37]. The area under the ROC curve (AUC) is also employed as an index to estimate the robustness of models. Generally, a model with AUC values between 0.5 to 1 indicates that a good classification model is obtained. If the AUC is lower than 0.5, the classification involves a random guess. The above evaluative indicators of models were record and analyzed using the SPSS v21.0 software (Statistical Product and Service Solutions, IBM Corporation, Armonk, NY, USA).

3. Results and Discussion

3.1. Spectral Properties

The variation of the average reflectance spectra in the spectral range of 400–1000 nm with their individual standard deviation (SD) is illustrated in Figure 1. It could be noticed that there were similar profiles of spectral curves between the two Chinese walnut varieties, but spectral intensity differed especially in rang of 800–1000 nm. The general similar spectral curves should be derived from their similar tissue composition, structure, and color presentation in the sample surface. The intensity differences were mainly due to the small differences in contents of surface chemical compositions (such as chlorophyll content, water content, etc.). However, spectral reflectivity was overlapped between different groups, so further data-driven analysis should be conducted to make a clear classification.

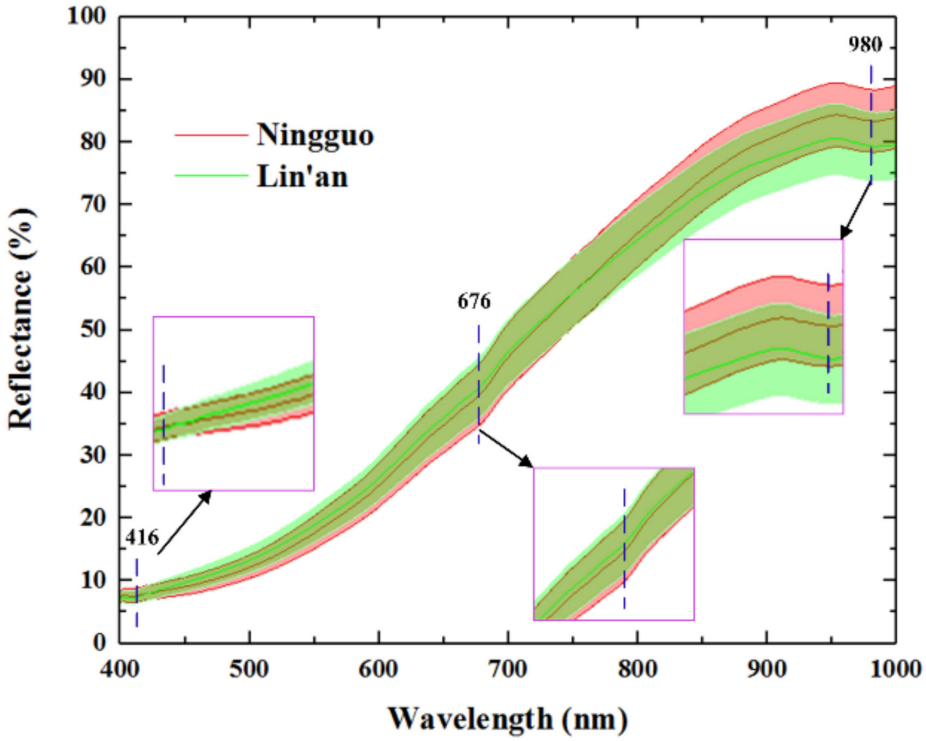

Figure 1. The average reflectance spectral curves of Chinese walnut samples of Ningguo and Lin'an varieties with standard deviation (SD).

In detail, the reflectance valley at 416 nm was associated with the Soret absorption band, which was closely related to the contents of porphyrins in chlorophyll [38]. The clear downwards wavelength at 980 nm corresponded to the 2nd O-H stretching overtone of water [18,39]. Another weak valley at 676 nm was responsible for chlorophyll-a [40]. As shown in sub-windows in Figure 1, differences at these three wavelengths could be further applied to be a basis for classifying samples of different varieties.

3.2. Exploratory Analysis

PCA was applied to visualize the spectral similarities and differences of samples of different varieties. The PCA results based on raw spectra showed that the first two PCs

accounted for a total of 98.19% of the cumulative spectral variances (86.16% for PC_1 and 12.03% for PC_2). These two PCs together showed great potential in differentiating the two Chinese walnut varieties, so only PC_1 and PC_2 were retained in this study. The total 400 samples are plotted in Figure 2 based on their first two PC scores. Figure 2a shows the score plot of PC_1 vs. PC_2 in the PC space, this preliminary evaluation indicated that there was a separable observation of samples of the two different varieties. Ellipses were drawn herein to denote their covered regions to intuitively see the distribution. Although there was little overlap, the samples in these two groups tended to gather, respectively. That is, Ningguo samples tended to have positive PC_2 values, while Lin'an samples had negative PC_2 values. PC_1 and PC_2 loading lines are further shown in Figure 2b. Peaks and valleys with high absolute coefficients were deemed to be effective. This comprehensive analysis showed that wavelengths centered at 676 nm, 760 nm, and 980 nm can be selected as effective wavelengths. The wavelengths of 676 nm and 980 nm were consistent with the ones selected in spectral analysis in Section 3.1. The band at 760 nm was related to the third stretching O-H overtone of water [41].

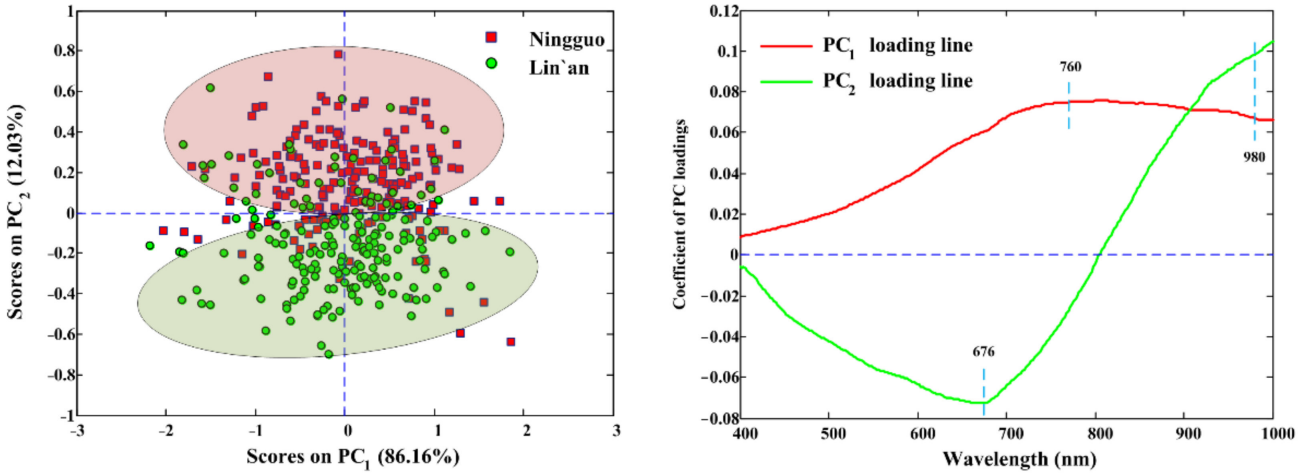

Figure 2. Spectra PCA for Chinese walnut samples of different varieties. (**a**) Score plot of PC_1 against PC_2, (**b**) PC_1 and PC_2 loading lines. PCA: principal component analysis; PC_1: first principal component; PC_2: second principal component.

As PC_1 and PC_2 were able to represent a total of 98.19% spectral variations, the corresponding score images were introduced to visualize and identify the inherent rules in spectra. Figure 3 shows the first two PC score images that transformed from calibrated hyperspectral images ranging from 400 to 1000 nm. Pixels were given different PC score values and displayed in different colors. It could be seen that there was no obvious difference between Ningguo and Lin'an walnuts in PC_1 score images. That is, the PC_1 score image can hardly be used for Chinese walnut varieties classification. This observation is consistent with the results presented in spectra PCA in Figure 2a. As is also can be observed that samples of these two varieties presented slightly different colors in PC_2 score image. The PCA and score images have demonstrated the potential separability between Ningguo and Lin'an Chinese walnuts. A high misclassification rate was also observed due to the similar spectral curves, and more accurate identification methods were still required.

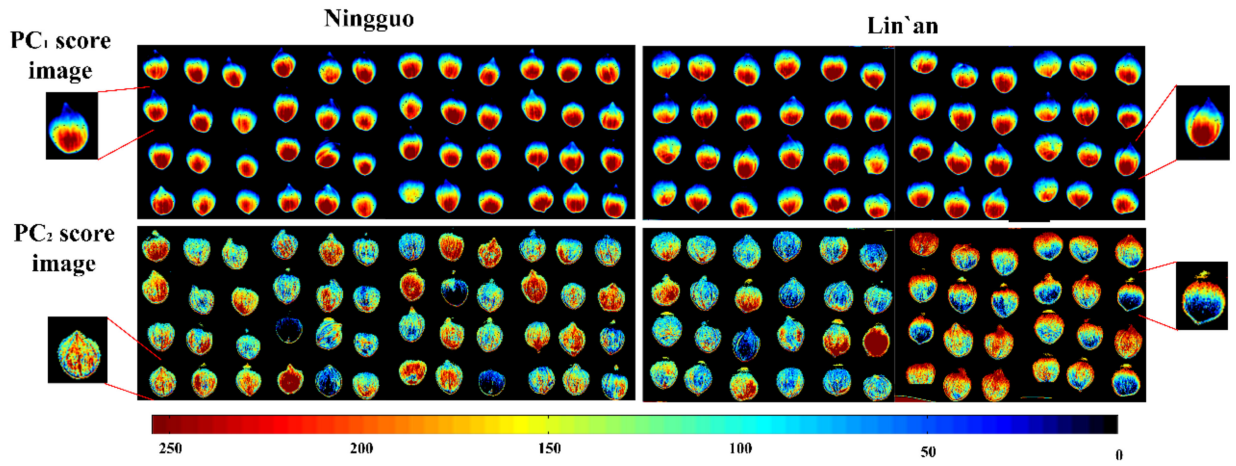

Figure 3. First two PC score images of Ningguo and Lin'an Chinese walnuts. PC: principal component; PC_1: first principal component; PC_2: second principal component.

3.3. Performance of Classification Models Based on Full Spectra

The performance of developed PLS-DA, KNN, and RBF-SVM models with their optimal parameters are shown in Table 1. These models were established based on raw or preprocessed spectra (400–1000 nm) including SNV, SNV+detrending, normalization, 1st, and 2nd derivatives, respectively. Results showed that raw full spectra performed not worse than other preprocessed spectra, and the 2nd derivative slightly decreased the performance. The developed optimal KNN model, which had the advantage of simple implementation, was capable in performing not bad identification with CCR of 79.00% in the prediction set. It could be also observed that PLS-DA and RBF-SVM models presented overall better results than KNN models with predicted CCRs above 89% in all sets. Among them, the PLS-DA model based on raw full spectra was found to perform best with the optimal LVs number of 13, calibration set CCR of 97.33%, cross-validation set CCR of 95.33%, and prediction set CCR of 92.00%.

As for different preprocessing methods, they were tentatively used for each model to reduce spectral noise and scatter effects, and the preprocessing method with the best results can be the most suitable one. The overall results implied that raw spectra were informative enough and any extra spectral preprocessing steps done to the raw spectral data was not necessary. Raw full spectra constructed PLS-DA and RBF-SVM models gave both the best results compared to other preprocessed full spectra. Furthermore, since the absolute difference among the three sets was small, PLS-DA model developed by raw full spectra was a more robust one than the RBF-SVM model. Therefore, the PLS-DA modeling method developed by raw spectra was chosen for further evaluation.

To further assess the capabilities of the optimal models, confusion matrices exhibiting the specific group characteristics and affinity are presented in Table 2. As for the selected PLS-DA model, Ningguo walnuts seemed to be more easily misclassified into Lin'an variety. There were 5, 8, and 6 Ningguo walnuts being misclassified as Lin'an in calibration, cross-validation, and prediction set, respectively, while the numbers that Lin'an misclassified as Ningguo were 3, 6, and 2. The RBF-SVM models also performed better for Lin'an variety identification than Ningguo variety in different sets. Especially in prediction set, RBF-SVM model gave the same classification results with the PLS-DA model.

Table 1. Performance of established classification models developed by full spectra with or without preprocessings.

Modeling Methods	Preprocessings	Corrected Classification Rate (%)			Parameters
		Calibration Set	Cross-Validation Set	Prediction Set	
PLS-DA	None	97.33	95.33	92.00	LVs = 13
	SNV	98.67	96.00	91.00	LVs = 12
	SNV + detrend	98.00	96.00	92.00	LVs = 11
	Normalization	96.33	93.00	90.00	LVs = 13
	1st derivative	97.67	96.00	92.00	LVs = 8
	2nd derivative	96.33	88.00	86.00	LVs = 5
KNN	None	78.67	77.00	72.00	K = 6
	SNV	81.67	76.67	70.00	K = 3
	SNV + detrend	90.00	86.67	79.00	K = 4
	Normalization	76.67	73.33	69.00	K = 7
	1st derivative	87.67	84.33	70.00	K = 5
	2nd derivative	72.67	69.67	62.00	K = 6
RBF-SVM	None	99.67	90.00	92.00	/
	SNV	99.33	89.67	92.00	/
	SNV + detrend	100.00	89.33	89.00	/
	Normalization	99.67	89.33	89.00	/
	1st derivative	99.33	95.00	91.00	/
	2nd derivative	99.00	89.67	89.00	/

PLS-DA: partial least squares-discriminant analysis; KNN: k-nearest neighbor; RBF-SVM: radial basis function-support vector machine; SNV: standard normal variate; LVs: latent variables.

Table 2. Confusion matrices of the optimal PLS-DA, SVM and KNN models based on full spectra.

Models	Group	Calibration Set			Cross-Validation Set			Prediction Set		
		Ningguo	Lin'an	Total	Ningguo	Lin'an	Total	Ningguo	Lin'an	Total
PLS-DA	Ningguo	145	5	97.33%	142	8	95.33%	44	6	92.00%
	Lin'an	3	147		6	144		2	48	
KNN	Ningguo	135	15	90.00%	129	21	86.67%	43	7	79.00%
	Lin'an	15	135		19	131		14	36	
RBF-SVM	Ningguo	150	0	99.33%	139	11	89.67%	44	6	92.00%
	Lin'an	2	148		20	130		2	48	

PLS-DA: partial least squares-discriminant analysis; KNN: *k*-nearest neighbor; RBF-SVM: radial basis function-support vector machine.

3.4. Effective Wavelengths Extraction

The usage of full spectra may bring the risk of overfitting, noise and nonlinearities that result in models with low accuracy [17]. Thus, effective wavelength selection from the full spectra (301 bands) was carried out to search for the wavelengths carrying feature information for the identification. In this study, SPA and CARS algorithms were individually used, and the results are shown in Figure 4. For SPA method, the minimum and the maximum numbers were individually set to 5 and 30 based on experiences [42]. The RMSE plot is presented in Figure 4a, and the RMSE reduced to 0.2948 when the number of variables reached 10. As a result, by comprehensively considering the number of variables and RMSE values, 10 wavelengths including 960, 930, 994, 472, 790, 906, 404, 402, 416, and 588 nm were retained in order of importance. In further steps, these 10 variables were applied as inputs to establish SPA-PLS-DA model. As indicated in Figure 4b, a total of 15 sensitive wavelengths selected by CARS had particular importance for the variety identification. Specifically, in the first subplot in Figure 4b, the number of sampled variables decreased as the sampling runs increased. It could be observed that the number began with the largest drop, and then tapered off gradually. Redundant variables in the full spectra were gradually eliminated in this step. In the second subplot, the RMSE value in cross-validation

(RMSECV) was used as an evaluation indicator for the number of sampling runs. The RMSECV value continued to decrease until the lowest value reached 296 and thereafter, the value increased. Consequently, the minimum RMSECV value at the 296th sampling run was used to denote a combination of effective variables. The third subplot shows the regression coefficients path of all the 301 variables at different sampling runs. As the RMSECV values changed, the coefficient of variables that dropped to 0 could be considered as a feature wavelength. Therefore, based on CARS, 15 (402, 456, 458, 472, 474, 650, 812, 842, 868, 932, 952, 956, 962, 968, and 996 nm) out of 301 variables were selected. Further classification of Chinese walnut varieties will be performed using PLS-DA based on SPA- and CARS-selected wavelengths.

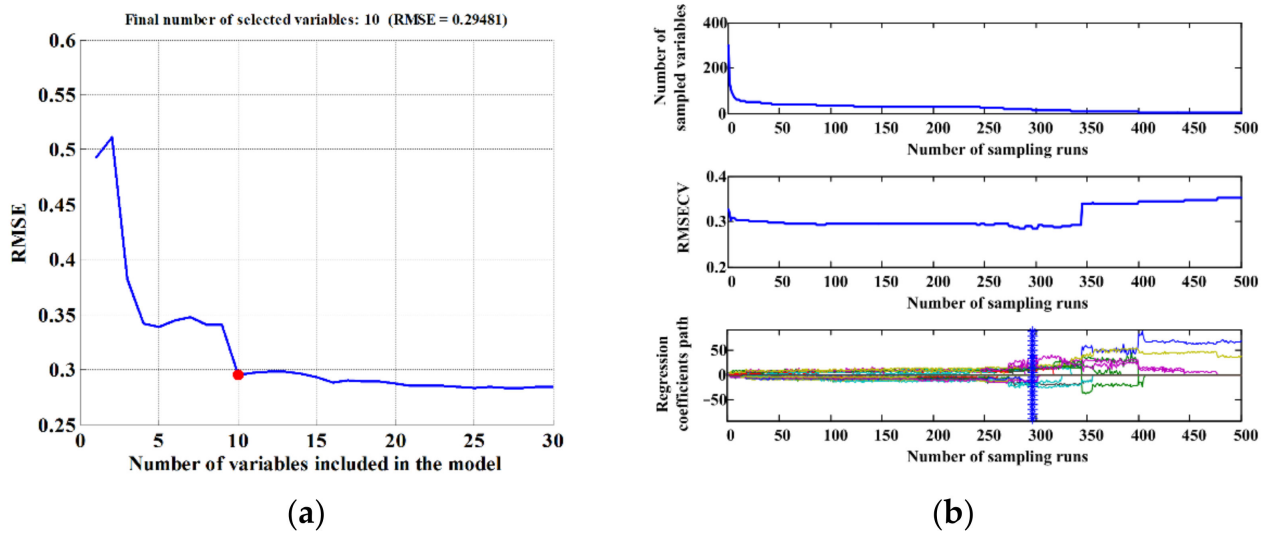

Figure 4. Wavelengths selection using SPA and CARS algorithms, (**a**) distribution maps of feature wavelengths selected by SPA; (**b**) parameter changes with sampling runs in wavelengths selection by CARS. RMSE: root mean squared error; RMSECV: root mean squared error in cross-validation; SPA: successful projects algorithm; CARS: competitive adaptive reweighted sampling.

3.5. Models Developed Using Feature Wavelengths

Table 3 displays the variety identification results of the simplified models based on selected wavelengths using three different methods. As shown, PC-PLS-DA model exhibited the worst performance, and the main reason was that a small number of only three variables could hardly contained all the valid information in the identification. When SPA-PLS-DA was considered, not bad classification results were obtained that CCRs equaled to 91.33%, 92.00%, and 89.00% in calibration, cross-validation, and prediction sets, respectively. Our study indicated that variable selection methods for Chinese walnut variety identification obtained different results. Similar results of using different variable selection methods for optimal wavelength selection could be found in the literature [43]. The optimal variable selection methods would be determined and used for identifying the variety. Moreover, previous study related to identify walnut varieties using NIRS showed that only a total classification accuracy of 77.00% ± 1.60% was achieved based on models developed by selected wavelengths [5]. This study showed good classification performances using both full spectra and selected wavelengths.

Table 3. The optimal PLS-DA modeling results developed by selected wavelengths based on different methods.

Models	Number	LVs	Corrected Classification Rate (%)		
			Calibration Set	Cross-Validation Set	Prediction Set
PC-PLS-DA	3	2	68.67	65.00	66.00
CARS-PLS-DA	15	11	96.33	95.67	91.00
SPA-PLS-DA	10	7	91.33	92.00	89.00

PC: principal component; PLS-DA: partial least squares-discriminant analysis; CARS: competitive adaptive reweighted sampling; SPA: successful projects algorithm; LVs: latent variables.

Above all, the PLS-DA model using CARS features achieved best performance with the CCRs of 96.33%, 95.67%, and 91.00% in calibration, cross-validation, and prediction sets. To further assess the performance, the confusion matrix is shown in Table 4. The CARS-PLS-DA model yielded a sensitivity of 93.62%, a specificity of 88.68%, AUC of 0.91, and Kappa coefficient of 0.82 in prediction set. These results suggested that CARS-PLS-DA model had great potential to identify Chinese walnut varieties without any chemical or physical information.

Table 4. Confusion matrix in the three sets predicted using the optimal simplified model.

	Calibration Set		Cross-Validation Set		Prediction Set	
	Ningguo	Lin'an	Ningguo	Lin'an	Ningguo	Lin'an
Ningguo	146	4	143	7	44	6
Lin'an	7	143	6	144	3	47
Sensitivity	97.28%	95.42%	95.36%	95.97%	88.68%	93.62%
Specificity	95.42%	97.28%	95.97%	95.36%	93.62%	88.68%

3.6. Classification Visualization of Chinese Walnut Varieties

The advantage of HSI to acquire both spatial and spectral information makes it possible to show the classification results of Chinese walnuts using intuitive classification maps. In our study, the simplified CARS-PLS-DA model was transferred to predict the varieties of each pixel. Compared to original hyperspectral images, the visual classification map could be eventually formed. The original false-color images of Ningguo and Lin'an walnuts are shown in Figure 5a,b. It could be seen that most of the walnuts differed little in texture or morphology characteristics. It was difficult to discriminate different varieties by naked eye. As shown in Figure 5c,d, the corresponding predicted varieties were marked using different colors (red for Ningguo walnuts, green for Lin'an walnuts, and black for background). The misclassified and well-classified pixels were clearly displayed in the classification maps. Most of the pixels within one sample were accurately classified, however, a distribution structured in band misclassification could be seen especially for Lin'an variety. The main reason was that samples shook slightly as the conveyor belt slipped, and noise was brought into the line-scanning images. Further application will focus on holding the walnuts with a belt with sockets in images collection procedure to obtain more accurate results. Anyway, the sample-level varieties were successfully identified with 100% accuracy, and more than 75% of the pixels were correctly identified. The visualization results indicated that HSI together with the optimal CARS-PLS-DA model also had great potential in visualizing the Chinese walnut varieties. This methodology was expected to be applied in modern nut industry as a powerful tool for large-scale qualitative detection of walnuts.

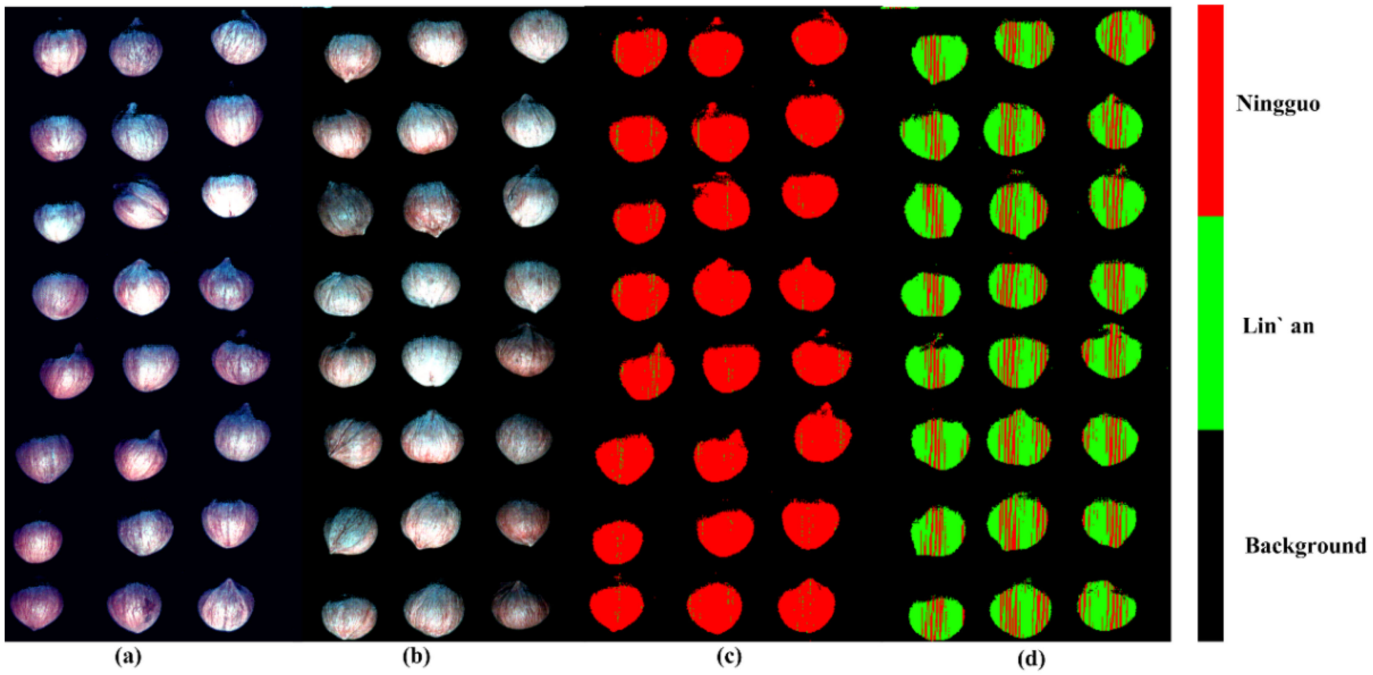

Figure 5. Classification visualization of Ningguo and Lin'an walnut varieties, (**a**) false-color images of Ningguo walnuts, (**b**) false-color images of Lin'an walnuts, (**c**) classification maps of Ningguo walnuts, and (**d**) classification maps of Lin'an walnuts.

The overall results indicated that HSI could become a good way for identifying Chinese walnut varieties. Coincidentally, a recent work [44] related to this subject also confirmed that walnut varieties can be evaluated by HSI technique. However, no attempts of HSI-based classification visualization have been conducted previously. Our study shows the potential of HSI for visualizing the variety categorization of Chinese walnuts. It provides a good example for widespread application of this technique on different kinds of nuts. The nut industry will benefit from this visual and rapid method to categorize nut according to their variety, and direct the nut of different varieties to appropriate end-uses.

4. Conclusions

Since the traditional methods for identifying the Chinese walnut varieties are highly time-consuming, destructive, or subjective, a method based on the hyperspectral imaging (HSI) technique coupled with chemometrics was successfully applied for classifying Ningguo and Lin'an Chinese walnuts. The first two principal component (PC) score plots and transformed PC score visualization images were drawn to clearly show the differences between Ningguo and Lin'an varieties. The partial least squares-discriminant analysis (PLS-DA) based models showed strong ability for varieties classification (accuracies of 97.33%, 95.33% and 92.00% in calibration set, cross-validation set, and prediction set). In order to reduce data dimensionality and further reduce the modeling time, the most effective 15 wavelengths selected by competitive adaptive reweighted sampling (CARS) were used to build a new PLS-DA model yielding 96.33%, 95.67%, and 91.00% accuracies in the three sets. These 15 wavelengths can be likely the most potential ones in further developing a multispectral instrument to discriminate Chinese walnut varieties rapidly and non-destructively. Final visualization map was successfully generated to observe the specific variety of each sample in an intuitive way. Visual differences between walnut samples of different varieties as well as different areas of the same sample were clearly displayed. In summary, all the results showed that HSI was feasible and beneficial in further building online large-scale Chinese walnut varieties identification system. More

Chinese walnuts originated from different planting areas will be further considered to improve the models' robustness.

Author Contributions: Conceptualization, H.J.; methodology, H.J.; software, H.J. and L.Y.; validation, H.J., L.Y. and X.L.; formal analysis, M.S.; investigation, M.S.; resources, H.J.; data curation, L.Y.; writing—original draft preparation, H.J.; writing—review and editing, M.S.; visualization, L.Y.; supervision, X.L.; project administration, H.J.; funding acquisition, H.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by Jiangsu Agriculture Science and Technology Innovation Fund, grant number CX(20)3040, and the Key Laboratory of on Site Processing Equipment for Agricultural Products, Ministry of Agriculture and Rural Affairs, China.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: The authors acknowledge the device support provided by NBL Imaging Systems Ltd.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Buthelezi, N.M.D.; Tesfay, S.Z.; Ncama, K.; Magwaza, L.S. Destructive and non-destructive techniques used for quality evaluation of nuts: A review. *Sci. Hortic.* **2019**, *247*, 138–146. [CrossRef]
2. Kris-Etherton, P.M.; Shaomei, Y.P.; Joan, S.; Ratcliffe, H.E.; Zhao, G.; Etherton, T.D. Nuts and their bioactive constituents: Effects on serum lipids and other factors that affect disease risk. *Am. J. Clin. Nutr.* **1999**, *70*, 504. [CrossRef] [PubMed]
3. Davis, L.; Stonehouse, W.; Loots, D.T.; Mukuddem-Petersen, J.; Westhuizen, F.H.V.D.; Hanekom, S.M.; Jerling, J.C. The effects of high walnut and cashew nut diets on the antioxidant status of subjects with metabolic syndrome. *Eur. J. Nutr.* **2007**, *46*, 155–164. [CrossRef] [PubMed]
4. Rong, D.; Xie, L.; Ying, Y. Computer vision detection of foreign objects in walnuts using deep learning. *Comput. Electron. Agric.* **2019**, *162*, 1001–1010. [CrossRef]
5. Arndt, M.; Drees, A.; Ahlers, C.; Fischer, M. Determination of the geographical origin of walnuts (*Juglans regia* L.) using near-infrared spectroscopy and chemometrics. *Foods* **2020**, *9*, 1860. [CrossRef]
6. Esteki, M.; Farajmand, B.; Amanifar, S.; Barkhordari, R.; Ahadiyan, Z.; Dashtaki, E.; Mohammadlou, M.; Heyden, Y.V. Classification and authentication of Iranian walnuts according to their geographical origin based on gas chromatographic fatty acid fingerprint analysis using pattern recognition methods. *Chemom. Intell. Lab. Syst.* **2017**, *171*, 251–258. [CrossRef]
7. Lavedrine, F.; Ravel, A.; Poupard, A.; Alary, J. Effect of geographic origin, variety and storage on tocopherol concentrations in walnuts by HPLC. *Food Chem.* **1997**, *58*, 135–140. [CrossRef]
8. Popescu, R.; Ionete, R.E.; Botoran, O.R.; Costinel, D.; Bucura, F.; Geana, E.I.; Alabedallat, Y.F.J.; Botu, M. 1H-nmr profiling and carbon isotope discrimination as tools for the comparative assessment of walnut (*Juglans regia* L.) cultivars with various geographical and genetic origins—A preliminary study. *Molecules* **2019**, *24*, 1378. [CrossRef]
9. Jensen, P.N.; Sørensen, G.; Engelsen, S.B.; Bertelsen, G. Evaluation of quality changes in walnut kernels (*Juglans regia* L.) by Vis/NIR spectroscopy. *J. Agric. Food Chem.* **2001**, *49*, 5790–5796. [CrossRef]
10. Jiang, X.S.; Zhao, T.X.; Liu, X.; Zhou, Y.C.; Shen, F.; Ju, X.R.; Liu, X.Q.; Zhou, H.P. Study on method for on-line identification of wheat mildew by array fiber spectrometer. *Spectrosc. Spectral. Anal.* **2018**, *38*, 3729–3735.
11. Ge, Y.; Atefi, A.; Zhang, H.; Miao, C.; Ramamurthy, R.K.; Sigmon, B.; Yang, J.; Schnable, J.C. High-throughput analysis of leaf physiological and chemical traits with VIS–NIR–SWIR spectroscopy: A case study with a maize diversity panel. *Plant Methods* **2019**, *15*, 66. [CrossRef] [PubMed]
12. Huang, Z.; Zhu, T.; Li, Z.; Ni, C. Non-destructive testing of moisture and nitrogen content in *Pinus Massoniana* seedling leaves with NIRS based on MS-SC-CNN. *Appl. Sci.* **2021**, *11*, 2754. [CrossRef]
13. Yi, J.; Sun, Y.; Zhu, Z.; Liu, N.; Lu, J. Near-infrared reflectance spectroscopy for the prediction of chemical composition in walnut kernel. *Int. J. Food Prop.* **2017**, *20*, 1633–1642. [CrossRef]
14. Wang, X.; Atungulu, G.G.; Khir, R.; Gao, Z.; Pan, Z.; Wilson, S.A.; Olatunde, D.; Slaughter, D. Sorting in-shell walnuts using near infrared spectroscopy for improved drying efficiency and product quality. *Int. Agric. Eng. J.* **2017**, *26*, 165–172.
15. ElMasry, G.; Kamruzzaman, M.; Sun, D.-W.; Allen, P. Principles and applications of hyperspectral imaging in quality evaluation of agro-food products: A review. *Crit. Rev. Food Sci. Nutr.* **2012**, *52*, 999–1023. [CrossRef]

16. Jiang, H.; Ru, Y.; Chen, Q.; Wang, J.; Xu, L. Near-infrared hyperspectral imaging for detection and visualization of offal adulteration in ground pork. *Spectrochim. Acta Part A* **2021**, *249*, 119307. [CrossRef] [PubMed]
17. Zhang, Y.; Jiang, H.; Wang, W. Feasibility of the detection of carrageenan adulteration in chicken meat using visible/near-infrared (Vis/NIR) hyperspectral imaging. *Appl. Sci.* **2019**, *9*, 3926. [CrossRef]
18. Jiang, H.; Cheng, F.; Shi, M. Rapid identification and visualization of jowl meat adulteration in pork using hyperspectral imaging. *Foods* **2020**, *9*, 154. [CrossRef]
19. Zhao, X.; Li, C.; Zhao, Z.; Wu, G.; Xia, L.; Jiang, H.; Wang, T.; Chu, X.; Liu, J. Generic models for rapid detection of vanillin and melamine adulterated in infant formulas from diverse brands based on near-infrared hyperspectral imaging. *Infrared Phys. Technol.* **2021**, *116*, 103745. [CrossRef]
20. Huang, Y.; Yang, Y.; Sun, Y.; Zhou, H.; Chen, K. Identification of apple varieties using a multichannel hyperspectral imaging system. *Sensors* **2020**, *20*, 5120. [CrossRef]
21. Huang, Y.; Si, W.; Chen, K.; Sun, Y. Assessment of tomato maturity in different layers by spatially resolved spectroscopy. *Sensors* **2020**, *20*, 7229. [CrossRef] [PubMed]
22. Moscetti, R.; Saeys, W.; Keresztes, J.C.; Goodarzi, M.; Cecchini, M.; Danilo, M.; Massantini, R. Hazelnut quality sorting using high dynamic range short-wave infrared hyperspectral imaging. *Food Bioprocess. Technol.* **2015**, *8*, 1593–1604. [CrossRef]
23. Feng, L.; Zhu, S.; Lin, F.; Su, Z.; Yuan, K.; Zhao, Y.; He, Y.; Zhang, C. Detection of oil chestnuts infected by blue mold using near-infrared hyperspectral imaging combined with artificial neural networks. *Sensors* **2018**, *18*, 1944. [CrossRef]
24. Kalkan, H.; Beriat, P.; Yardimci, Y.; Pearson, T.C. Detection of contaminated hazelnuts and ground red chili pepper flakes by multispectral imaging. *Comput. Electron. Agric.* **2011**, *77*, 28–34. [CrossRef]
25. Mohammadi-Moghaddam, T.; Razavi, S.M.A.; Taghizadeh, M.; Pradhan, B.; Sazgarnia, A.; Shaker-Ardekani, A. Hyperspectral imaging as an effective tool for prediction the moisture content and textural characteristics of roasted pistachio kernels. *J. Food Meas. Charact.* **2018**, *12*, 1493–1502. [CrossRef]
26. Nakariyakul, S.; Casasent, D.P. Classification of internally damaged almond nuts using hyperspectral imagery. *J. Food Eng.* **2011**, *103*, 62–67. [CrossRef]
27. Yang, Y.; Wang, W.; Zhuang, H.; Yoon, S.C.; Jiang, H. Prediction of quality traits and grades of intact chicken breast fillets by hyperspectral imaging. *Br. Poult. Sci.* **2021**, *62*, 46–52. [CrossRef]
28. Jiang, H.; Jiang, X.; Ru, Y.; Wang, J.; Xu, L.; Zhou, H. Application of hyperspectral imaging for detecting and visualizing leaf lard adulteration in minced pork. *Infrared Phys. Technol.* **2020**, *110*, 103467. [CrossRef]
29. Mishra, P.; Passos, D. A synergistic use of chemometrics and deep learning improved the predictive performance of near-infrared spectroscopy models for dry matter prediction in mango fruit. *Chemom. Intell. Lab. Syst.* **2021**, *212*, 104287. [CrossRef]
30. Dotto, A.C.; Dalmolin, R.S.D.; Caten, S.T.A.; Grunwald, S. A systematic study on the application of scatter-corrective and spectral-derivative preprocessing for multivariate prediction of soil organic carbon by Vis-NIR spectra. *Geoderma* **2018**, *314*, 262–274. [CrossRef]
31. Swierenga, H.; De Weijer, A.P.; Van Wijk, R.J.; Buydens, L.M.C. Strategy for constructing robust multivariate calibration models. *Chemom. Intell. Lab. Syst.* **1999**, *49*, 1–17. [CrossRef]
32. Araujo, M.C.U.; Saldanha, T.C.B.; Galvao, R.K.H.; Yoneyama, T.; Chame, H.C.; Visani, V. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. *Chemom. Intell. Lab. Syst.* **2001**, *57*, 65–73. [CrossRef]
33. Li, H.; Liang, Y.; Xu, Q.; Cao, D. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. *Anal. Chim. Acta* **2009**, *648*, 77–84. [CrossRef] [PubMed]
34. Barker, M.; Rayens, W. Partial least squares for discrimination. *J. Chemom.* **2003**, *17*, 166–173. [CrossRef]
35. Zhang, L.; Sun, H.; Rao, Z.; Ji, H. Hyperspectral imaging technology combined with deep forest model to identify frost-damaged rice seeds. *Spectrochim. Acta Part A* **2020**, *229*, 117973. [CrossRef] [PubMed]
36. Chen, Q.; Zhao, J.; Fang, C.H.; Wang, D. Feasibility study on identification of green, black and Oolong teas using near-infrared reflectance spectroscopy based on support vector machine (SVM). *Spectrochim. Acta Part A* **2007**, *66*, 568–574. [CrossRef]
37. Lorente, D.; Aleixos, N.; Gómez-Sanchis, J.; Cubero, S.; Blasco, J. Selection of optimal wavelength features for decay detection in citrus fruit using the ROC curve and neural networks. *Food Bioprocess. Technol.* **2013**, *6*, 530–541. [CrossRef]
38. Yu, K.; Fang, S.; Zhao, Y. Heavy metal Hg stress detection in tobacco plant using hyperspectral sensing and data-driven machine learning methods. *Spectrochim. Acta Part A* **2021**, *245*, 118917. [CrossRef]
39. Bowker, B.; Hawkins, S.; Zhuang, H. Measurement of water-holding capacity in raw and freeze-dried broiler breast meat with visible and near-infrared spectroscopy. *Poult. Sci.* **2014**, *93*, 1834–1841. [CrossRef]
40. Qin, J.; Lu, R. Measurement of the optical properties of fruits and vegetables using spatially resolved hyperspectral diffuse reflectance imaging technique. *Postharvest Biol. Technol.* **2008**, *49*, 355–365. [CrossRef]
41. Wu, D.; Sun, D.W. Application of visible and near infrared hyperspectral imaging for non-invasively measuring distribution of water-holding capacity in salmon flesh. *Talanta* **2013**, *116*, 266–276. [CrossRef] [PubMed]
42. Wu, X.; Song, X.; Qiu, Z.; He, Y. Mapping of TBARS distribution in frozen-thawed pork using NIR hyperspectral imaging. *Meat Sci.* **2016**, *113*, 92–96. [CrossRef] [PubMed]

43. Zhang, C.; Jiang, H.; Liu, F.; He, Y. Application of near-infrared hyperspectral imaging with variable selection methods to determine and visualize caffeine content of coffee beans. *Food Bioprocess. Technol.* **2017**, *10*, 213–221. [CrossRef]
44. Nogales-Bueno, J.; Feliz, L.; Baca-Bocanegra, B.; Hernández-Hierro, J.M.; Heredia, F.J.; Barroso, J.M.; Rato, A.E. Comparative study on the use of three different near infrared spectroscopy recording methodologies for varietal discrimination of walnuts. *Talanta* **2020**, *206*, 120189. [CrossRef]