Hindawi
Journal of Food Quality
Volume 2023, Article ID 6104038, 11 pages
https://doi.org/10.1155/2023/6104038

Research Article

Rapid and Nondestructive Identification of Origin and Index Component Contents of Tiegun Yam Based on Hyperspectral Imaging and Chemometric Method

Yue Zhang,[1,2] **Yuan Li**,[1] **Cong Zhou**,[1] **Junhui Zhou**,[1,3] **Tiegui Nan**,[1] **Jian Yang**,[1,3] **and Luqi Huang**[1]

[1]*State Key Laboratory Breeding Base of Dao-di Herbs, National Resource Center for Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing 100700, China*
[2]*School of Traditional Chinese Medicine, Yunnan University of Chinese Medicine, Kunming 650500, China*
[3]*Dexing Research and Training Center of Chinese Medical Sciences, Dexing 334220, China*

Correspondence should be addressed to Jian Yang; yangchem2012@163.com and Luqi Huang; huangluqi01@126.com

Received 5 July 2022; Revised 7 September 2022; Accepted 24 November 2022; Published 22 February 2023

Academic Editor: Daniel Cozzolino

Tiegun yam is a typical food and medicine agricultural product, which has the effects of nourishing the kidney and benefitting the lungs. The quality and price of Tiegun yam are affected by its origin, and counterfeiting and adulteration are common. Therefore, it is necessary to establish a method to identify the origin and index component contents of Tiegun yam. Hyperspectral imaging combined with chemometrics was used, for the first time, to explore and implement the identification of origin and index component contents of Tiegun yam. The origin identification models were established by partial least squares-discriminant analysis (PLS-DA), support vector machine (SVM), and random forest (RF) using full wavelength and feature wavelength. Compared with other models, MSC-PLS-DA is the best model, and the accuracy of the training set and prediction set is 100% and 98.40%. Partial least squares regression (PLSR), random forest (RF), and support vector regression (SVR) models were used to predict the contents of starch, polysaccharide, and protein in Tiegun yam powder. The optimal residual predictive deviation (RPD) values of starch, polysaccharide, and protein prediction models selected in this study were 5.21, 3.21, and 2.94, respectively. The characteristic wavelength extracted by the successive projections algorithm (SPA) method can achieve similar results as the full-wavelength model. These results confirmed the application of hyperspectral imaging (HSI) in the identification of the origin and the rapid nondestructive prediction of starch, polysaccharide, and protein contents of Tiegun yam powder. Therefore, the HSI combined with the chemometric method was available for conveniently and accurately determining the origin and index component contents of Tiegun yam, which can expect to be an attractive alternative method for identifying the origin of other food.

1. Introduction

Yam is the fleshy underground tuber of *Dioscorea*. In China, yam is a typical food and medicine agricultural product. It is not only a common vegetable, often used for fresh eating, fresh stir-fry, steaming, making vermicelli, and potato chips, but also a traditional Chinese medicine which can invigorate the spleen and stomach, benefit the lung, generate saliva, and benefit the kidney [1]. Yam contains many nutrient metabolites which include starch, polysaccharide, and protein [2]. Because of its nutritional content, it is widely used in traditional Chinese medicine to treat chronic diseases such as indigestion [3]. China is an important center of yam cultivation, with more than 90 varieties [4]. Tiegun yam (*D. opposite* Thunb.) is mainly produced in Wen County, Henan Province of China, has the best quality, and is considered to be the representative of Chinese yam (*Dioscorea opposite* Thunb.). In recent years, Tiegun yam has been planted in most areas of China, such as Inner Mongolia, Shaanxi, Jiangsu, Shandong, Hebei, and other places, and there have been many problems with fake Tiegun yam from other origins in the market. The fake Tiegun yam not only damages the interests of consumers but also has a great difference from that of the Wen County Tiegun yam in nutrient substance. Starch is the most abundant ingredient, which also affects the taste of Tiegun yam, while polysaccharides and proteins are the main pharmacological components. The content of these three components has also become an important index to evaluate the quality of Tiegun Yam. Therefore, it is necessary to establish an analytical method to identify the origin and index component contents of Tiegun yam.

At present, the traditional nutrient content evaluation methods include enzymatic hydrolysis [5] and the underwater weight method [6] to determine the starch content. The content of polysaccharides was determined by inductively coupled plasma mass spectrometry (ICP-MS) [7], spectrophotometry [8], and liquid chromatography-tandem mass spectrometry (LC-MS) [9], and the content of protein was determined by the Kjeldner method [10], spectrophotometry, and combustion [11]. However, most of the traditional content determination methods are destructive, time consuming, and environment polluting with few other shortcomings and only applies to small samples and cannot meet the requirements of online monitoring [12]. In addition, the traditional methods of origin identification include liquid chromatography-mass spectrometry [13], gas chromatography-mass spectrometry [14], molecular biology techniques [15], stable isotope [16], and chemical fingerprinting [17]. However, traditional chemical methods have high accuracy but also have some disadvantages, such as high detection cost, difficult operation, and time consuming. Therefore, the establishment of a rapid and accurate detection method has an urgent market demand.

Hyperspectral imaging (HSI) technology is a nondestructive detection method that integrates image information and spectral information [18]. Compared with traditional spectral analysis technology, HSI can not only obtain two-dimensional spatial and one-dimensional spectral information that corresponds to internal and external features [5] but also collect data from multiple samples simultaneously. Compared to single-point measurement technologies, HSI is capable of performing spatial substance content analysis [19]. It has been widely used in the rapid identification of corn [20], sorghum [21], wolfberry [22], chrysanthemum [23], and other samples due to its characteristics of simultaneously obtaining spectral and spatial information. At the same time, HSI is also used for detecting the content of various substances, such as starch content in sorghum detection [19], corn grain oil content [24], total content of flavonoids in the cherry prediction [12], analysis of protein content in rice [11], and prediction of total flavonoids and polysaccharides in Anoectochilus formosanus [25]. All these studies have achieved satisfactory results. However, as far as we know, no studies have been published on using HIS to identify the origin and determine the contents of starch, polysaccharide, and protein in Tiegun yam.

In this study, we discussed the feasibility of the HSI method to identify the origin and determine the contents of starch, polysaccharide, and protein in Tiegun yam. An efficient and accurate method based on the graph segmentation algorithm was developed to achieve the rapid automatic identification and information extraction of the hyperspectral information of Tiegun yam powder samples. In addition, it combined with the chemometric method that can establish different models to realize the rapid identification of different regions of Tiegun yam and the accurate prediction index of its composition.

2. Materials and Methods

2.1. Sample Preparation. Tiegun yam samples were collected from late October to November 2019, from six producing areas, including Inner Mongolia (NM) ($n = 7$), Shaanxi (SX) ($n = 6$), Jiangsu (JS) ($n = 6$), Shandong (SD) ($n = 11$), Hebei (HB) ($n = 13$), and Henan (HN) ($n = 17$) provinces, and finally a total of 60 batches of Tiegun yam samples was collected. All batches of samples were purchased from the local medicinal herbs market, and each batch consisted of 10–20 Tiegun yam. The specific information of the sample is shown in Table 1. The collected Tiegun yam samples were cleaned, peeled, and cut into 10 cm sections and dried for approximately 36 hours in an oven at 50°C. Finally, all dried Tiegun yam samples were ground into powder and sifted through 50 mesh. The powder was sealed in a polyethylene bag and stored at 4°C.

2.2. Hyperspectral Imaging Systems. The HSI system consisted of an imaging spectrograph, a high-performance charged couple device (CCD) camera, a pair of 150 W halogen lamps (150 W/12 V, H-LAM Norsk Elektro Optikk, Norway), a mobile platform (Standa Translation Stage, Lithuania) driven by a stepper motor, and a computer with data acquisition and analysis software (HySpex Ground, Norsk Elektro Optikk, Norway). The imaging spectrograph consisted of SN0605 VNIR (H-V16, Norsk Elektro Optikk, Norway) and N3124 SWIR (H-S16, Norsk Elektro Optikk, Norway).

2.3. Hyperspectral Data Acquisition. First, 60 samples were randomly selected from different batches of Tiegun yam powder from each producing area, with a total of 360 samples. 15 g of each Tiegun yam samples powder sample was selected as a hyperspectral sample and loaded into a Petri dish with a diameter of 5 cm. The criterion is that the bottom of the Petri dish should not be seen when the powder is laid flat.

When the sample was collected, the distance between the spectrometer lens and the sample was 25 cm, the platform moving speed was 1.5 mm/s, the integration time of SN0605 VNIR lens was 3500 μs, the frame time was 18000, and the spectral range was 410–990 nm. The integration time of the N3124 SWIR lens was 4500 μs, the frame time was 46928, and the spectral range was 950–2500 nm. The spectral

Table 1: Detailed information of Tiegun yam.

Number	Origin	Quantity of sample
NM	Dengkou County, Bayanchuer city, Inner Mongolia, China	60
SX	Dali County, Weinan city, Shaanxi province, China	60
JS	Fengxian County, Xuzhou city, Jiangsu province, China	60
SD	Dingtao District, Heze city, Shandong province, China	60
HB	Li County, Baoding city, Hebei province, China	60
HN	Wen County, Jiaozuo city, Henan province, China	60

resolution of both VNIR and SWIR lenses was 6 nm. The samples were arranged on a black horizontal moving platform according to the matrix, and the Teflon whiteboard was placed at the end of the sample row to collect the hyperspectral images. In order to reduce the influence of natural light on the experiment, the whole experiment was conducted in a dark room. Finally, the average surface spectral data of the powder in each Petri dish were used as a region of interest (ROI).

2.4. Hyperspectral Image Processing. In order to eliminate the influence of instruments and environment on the sample data, the raw hyperspectral image data were corrected by software (HySpex RAD, Norsk Elektro Optikk, Norway), followed by black-and-white plate correction. Black-and-white plate correction is a common method in hyperspectral image data processing, which is used to eliminate the influence of air and surrounding environment on spectral images, so as to obtain the relative reflectance of the spectrum. This method is used to calculate the relative reflectivity of samples, whiteboards, and blackboards, and the calculation formula is as follows:

$$R = \frac{R\mathrm{raw} - R\mathrm{d}}{R\mathrm{w} - R\mathrm{d}}, \tag{1}$$

where R is the corrected reflectivity image, Rraw is the original reflectivity image, Rw is the whiteboard reference image, which is obtained by Teflon whiteboard (reflectivity is close to 1), and Rd is the blackboard reference image, which is obtained by covering the lens cap (reflectivity is close to 0).

2.5. Reference Measurement of Nutrient Substances Content

2.5.1. Evaluation of Starch Content. The soluble sugar and starch in the samples were separated by 80% ethanol, and the starch was hydrolyzed into glucose by acid hydrolysis. Glucose content was determined by anthrone colorimetry, and starch content was calculated [26]. Glucose standard solution of 1, 0.8, 0.4, 0.2, 0.1, and 0.05 mg/mL was prepared. The standard curve $Y = 2.9468x + 0.2768$ ($R^2 = 0.997$) was established with glucose concentration as abscissa and ΔA ($\Delta A = A - A$ blank) as abscissa. The 0.01 g Tiegun yam powder sample was weighed, and the test solution was configured. The absorbance value A was measured at 620 nm with a microplate reader. The abovementioned determination was completed with a total starch content determination kit (BC0700, Solarbio, Beijing, China). The formula for calculating starch content is as follows:

$$M1 = \frac{x * V * F}{1.11 * W}, \tag{2}$$

where $M1$ stands for starch content (mg/g), x is the calculated concentration of starch based on the standard curve (mg/mL), W is the sample mass (g), F stands for sample dilution ratio, and V is the volume after extraction (ml).

2.5.2. Evaluation of Polysaccharide Content. Total polysaccharides were extracted by the water extraction and alcohol precipitation method, and the content of total polysaccharides was determined by the phenol-sulfuric acid method [27, 28]. Standard solution of 0.4, 0.2, 0.1, 0.05, 0.025, and 0.0125 mg/mL was prepared, and the standard curve $Y = 6.7386x + 0.2257$ ($R^2 = 0.998$) was established with concentration as abscissa and ΔA ($\Delta A = A - A$ blank). The 0.025 g Tiegun yam powder sample was weighed, and the test solution was configured. The absorbance value A was measured at 490 nm with a microplate tester. The abovementioned determination was completed with a total polysaccharide content determination kit (YX-W-ZDT, HEPENGBIO, Shanghai, China). The formula for calculating the polysaccharide content is as follows:

$$M2 = \frac{5 * Y}{W}, \tag{3}$$

where $M2$ stands for polysaccharide content (mg/g), Y is the calculated concentration of polysaccharide based on the standard curve (mg/mL), and W stands for sample mass (g).

2.5.3. Evaluation of Protein Content. Protein concentration was detected by the Bradford method. The standard protein solution of 0.0625, 0.125, 0.25, 0.5, 0.75, 1, and 1.5 mg/mL was configured to establish the standard curve $Y = 0.5676x + 1.4197$ ($R^2 = 0.994$). The 0.025 g Tiegun yam powder sample was weighed with the test solution, and the absorbance was measured at 595 nm. The Bradford protein Assay kit (P0006C, Beyotime Biotechnology, Shanghai, China) was used to detect protein concentration. The formula for calculating protein concentration is as follows:

$$M3 = \frac{Y}{W}, \tag{4}$$

where Y stands for protein concentration (mg/mL) and W stands for weight of the sample (g).

2.6. Statistical and Chemometrics Analysis

2.6.1. Statistical Analysis. The mean value and standard deviation of starch, polysaccharide, and protein contents of Tiegun yam from different origins were calculated. The contents of starch, polysaccharide, and protein of Tiegun yam from different producing areas were compared. One-way analysis of variance (ANOVA) ($P < 0.05$) was used to analyze whether there were significant differences in the contents of polysaccharides, starch, and proteins in Tiegun yam from different origins.

2.6.2. Data Preprocessing. A total of 360 hyperspectral data were obtained from the extraction of hyperspectral regions of interest. The pretreatment of spectral data can reduce errors caused by baseline changes such as background, noise, and other physical factors and can improve the prediction ability and stability of the model. In this study, five methods including multiple scattering corrections (MSCs), first derivative (D1), second derivative (D2), SG smoothing (SG), and standard normal variable transformation (SNV) were used to preprocess spectral data to improve the accuracy and stability of the discrimination model.

2.6.3. Chemometric Method. Three different classification models, including partial least squares discriminant analysis (PLS-DA), support vector machine (SVM), and random forest (RF), were established to identify the origin of Tiegun yam. Similarly, three different classification models including partial least square regression (PLSR), support vector regression (SVR), and random forest (RF) were established.

PLSR model, a classical linear regression algorithm, can consider both matrices x (spectral data) and y (chemical index), to find the maximal correlation between the new variables of X and Y [29, 30]. PLS-DA is a supervised classification algorithm adapted from PLSR. The optimal number of 10–12 important potential variables in different prediction groups was obtained by using the leave-one cross-validation method.

The SVM model, which aims to obtain the best hyperplane by selecting the hyperplane passing through the maximum possible gap between points of different categories, was used with a nonlinear radial basis function to reduce the training complexity. In this research, the SVM model was constructed based on the radial basis function, and the optimal combination of two important parameters, namely, the penalty factor ($C = 12000$) and the kernel parameter ($\gamma = 100$), was determined by a grid-search method [31]. Support vector regression (SVR) is an important application branch of the support vector machine (SVM).

RF is an integration algorithm based on a classified regression tree, which builds multiple regression trees by constructing multiple training sets with the putback samples. The number of trees in this study is 50.

The successive projections algorithm (SPA) was used to select the characteristic wavelength of the classification model. SPA is a forward variable selection algorithm that minimizes the space collinearity of vector quantity. Its advantage lies in extracting several characteristic wavelengths of the whole band and eliminating redundant information in the original spectral matrix. Finally, the characteristic variable modeling results are compared with the full-band modeling results.

2.6.4. Model Evaluation. The performance of the classification model was evaluated based on the classification discrimination accuracy and confusion matrix. The confusion matrix is a method to evaluate the prediction results of the classification model in data analysis. The specific evaluation indexes include accuracy, sensitivity, and specificity [32]. These precision indexes reflect the accuracy of model classification from different aspects.

$$\begin{aligned} \text{Accuracy} &= \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP} + \mathrm{TN}}, \\ \text{Sensitivity} &= \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \\ \text{Specificity} &= \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}, \end{aligned} \tag{5}$$

where TP is the number of true positive samples, TN is the number of true negative samples, FP is the number of false positive samples, and FN is the number of false negative samples.

The prediction effect of pretreatment methods combined with regression models was evaluated based on residual predictive deviation (RPD) and curve correlation coefficient R^2 values. Usually, the R^2 value from 0.61 to 0.80 and the RPD value ranging from 2.0 to 2.5 indicate that the model can be used for prediction. R^2 value between 0.81 and 0.90 and RPD value between 2.5 and 3.0 demonstrate high model performance. The model has an excellent prediction performance with an R^2 value higher than 0.90 and an RPD value higher than 3.0.

2.7. Data Division. The samples were randomly divided into training sets and prediction sets in a ratio of 7 : 3 for subsequent modeling and analysis. 240 (2/3 samples) and 120 (1/3 samples) samples were randomly assigned to establish prediction models for starch, polysaccharide, and protein. The predictive set content range should be included in the training set content range.

2.8. Software and Program. The image correction tool used in this study is RAD correction software. ROI was collected by ENVI 5.3 software (Harris Geospatial Solutions Inc., CO, USA). ANOVA was conducted on SPSS software (22.0 version, IBM Inc., Chicago, IL, USA). Data analysis, such as spectral data preprocessing and classification model construction, was realized by Matlab 2020a (MathWorks, USA) software, and scripts were written by our research group.

Table 2: Statistical values of starch, polysaccharide, and protein contents in Tiegun yam powder for both calibration and prediction sets (mg/g).

Content	Sets	Range	Mean	SD
Starch	Training	377.4–638.3	489.5	60.23
	Prediction	391.4–603.8	490.5	60.33
Polysaccharide	Training	8.06–36.01	22.41	8.13
	Prediction	8.06–35.20	23.57	11.04
Protein	Training	5.09–51.25	23.68	10.05
	Prediction	5.44–36.59	22.37	9.62

3. Results and Discussion

3.1. Statistical Analysis. The content ranges, mean, and standard deviation (SD) of starch, polysaccharide, and protein of Tiegun yam powder classified into a training set and prediction set are shown in Table 2. The contents of starch, polysaccharide, and protein training set ranged from (377.4 to 638.3) mg/g, (8.06 to 36.01) mg/g, and (5.09 to 51.25) mg/g and (391.4 to 603.8) mg/g, (8.06 to 35.20) mg/g, and (5.44 to 36.59) mg/g for the prediction set. The predictive set content range should be included in the training set content range. Meanwhile, at the same time, ANOVA showed that there were significant differences in starch, polysaccharide, and protein contents of Tiegun yam beans from 6 origins ($P < 0.05$). The starch content of Tiegun yam in JS was the highest, while that in SD was the lowest. The other four producing areas have little difference in starch content. The content of polysaccharide of Tiegun yam in NM, SX, and JS was lower than that in SD, HB, and HN. The protein content of Tiegun yam was very low in NM, and there was little difference in other producing areas. Wen County of Henan province, as an authentic production area of Tiegun yam, had a high content of all these three nutrients. In other words, starch, polysaccharide, and protein content was significantly affected by the origin. The place of origin can be used as a grouping basis for classification modeling of hyperspectral Tiegun yam.

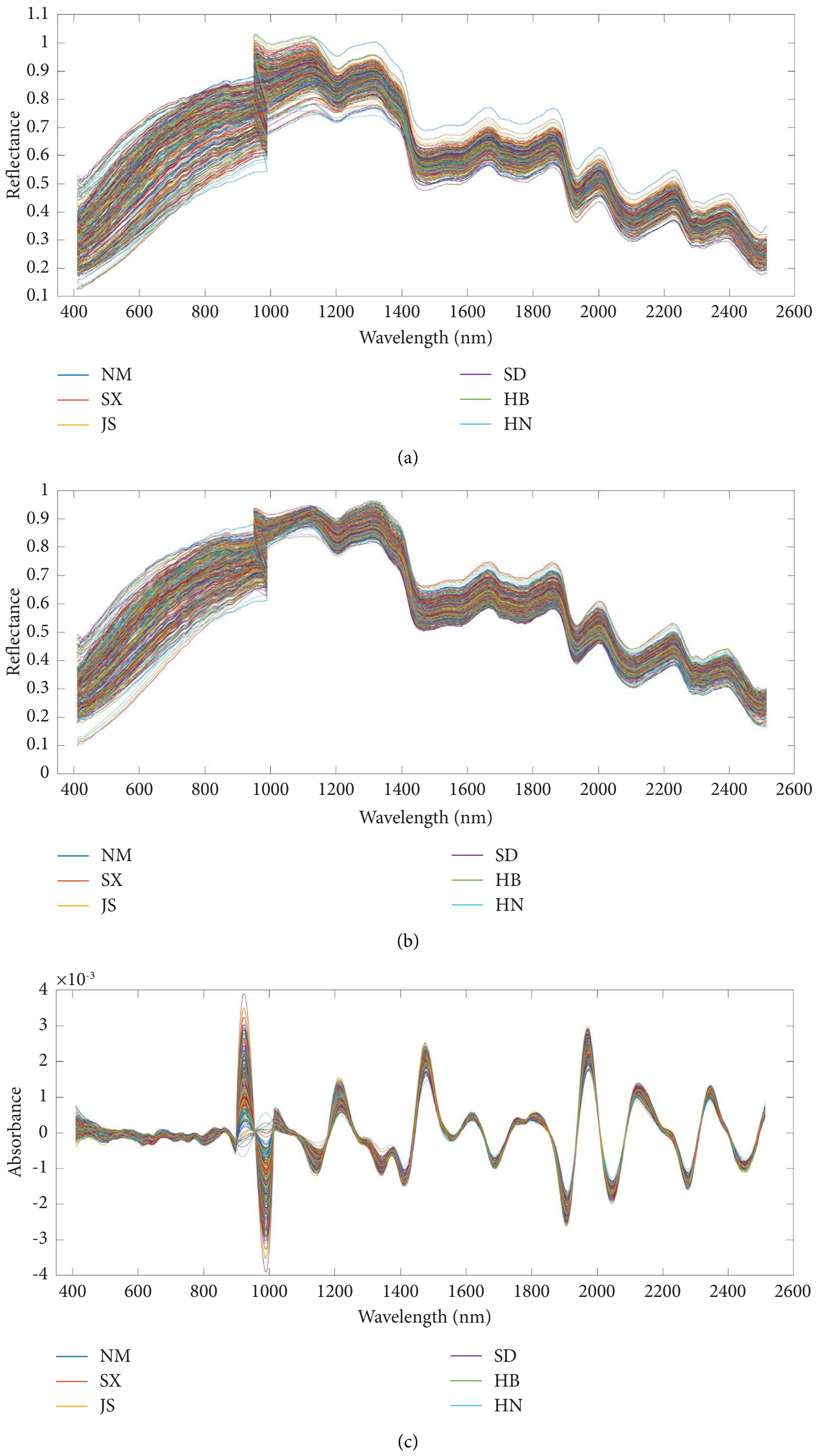
3.2. Original Spectral Curve Analysis. The spectral curves of Tiegun yam powder samples from different origins (Figure 1) have similar variation trends in VNIR and SWIR bands, and the mean values of spectral data have obvious differences in visible near-infrared (VNIR) bands, which may be caused by the significant differences in chemical composition content of Tiegun yam powder samples from different origins as shown in Table 2. This is related to different plant base sources, environmental conditions, and planting methods. However, the overall spectral characteristics of the short wave near-infrared (SWIR) band are similar with little difference.

The absorption peaks near 980 nm, 1450 nm, and 1855 nm were mainly attributed to the moisture [33]. However, because 980 nm is where the VNIR band ends, the signature is not obvious. The absorption peak near 1210 nm corresponded to the second stretching overtone of C-H. This absorption peak was mainly attributed to carbohydrates and fats [34]. The absorption peak at 1290 nm and 1471 nm was formed by the in-plane bending of C-H. The absorption peak at 1648 nm was formed due to the effect of amide groups [35]. The absorption peak at 1792 nm indicated the anhydride group. The absorption peak at 2069 nm was formed due to the combined effect of stretching and bending of O-H. The absorption peak at 2101 and 2190 nm was the characteristic absorption peaks of the protein. The absorption peak at 2101 nm might be possibly associated with the carboxyl group. The absorption peak at 2190 nm indicated the combined absorption peak of C-H and C-O [11]. Compared to VNIR, the wavelengths in SWIR could fully reflect the vibration of molecular bonds in different compounds.

3.3. Results of the Origin Identification Model of Tiegun Yam. The original spectral data of Tiegun yam powder samples from different origins were preprocessed by MSC, D1, D2, SG, and SNV, and the training set/prediction set was divided into input variables to calculate the accuracy of PLS-DA, RF, and SVM classification and identification methods (Table 3). For the PLS-DA model, MSC preprocessing can improve the accuracy of the model training set and prediction set. The prediction set accuracy of raw data-PLS-DA was 96.00%, and that of MSC-PLS-DA was 98.40%, with an improvement of 2.40%. The SVM model has high precision in the training set and low precision in the prediction set, and there is a big gap between them. The SVM model may not be suitable for the origin identification data of Tiegun yam. The accuracy of the prediction set of the D2-RF model is 83.33%, which is 14.44% higher than the raw data, but lower than that of the MSC-PLS-DA model. Therefore, MSC-PLS-DA is the optimal model for the origin identification of Tiegun yam, and this model is used for spectral modeling after feature wavelength selection and model evaluation.

As shown in Figure 2, there are 55 feature wavelengths selected based on SPA. MSC preprocessing and PLS-DA algorithm are used for modeling, and the accuracy of the training set is 99.13% and the prediction set is 97.71%. The results show that the extraction of characteristic wavelength modeling can achieve almost the same results as full-wavelength modeling.

In classification problems, a confusion matrix is a visual evaluation criterion to describe the real category attributes of sample data and to predict the performance of algorithms. The behavior of the confounding matrix is a true label, listed as the predictive label. The bottom line shows the percentage of predicted correct or incorrect classification, that is, sensitivity and error rate. The right-most column shows the percentage of all examples that fall into each category that are correctly and incorrectly classified, that is, the precision and false negative rates. After MSC treatment of the full-band spectrum of Tiegun yam powder samples, the confusion matrix generated by the PLS-DA classification model prediction results of samples from different sources is shown in Figure 3(a), and the sensitivity and precision of the classification and identification models of each origin are all above 95. The confusion matrix generated by the PLS-DA classification model after SPA screening characteristic



(a)

(b)

(c)

Figure 1: Continued.

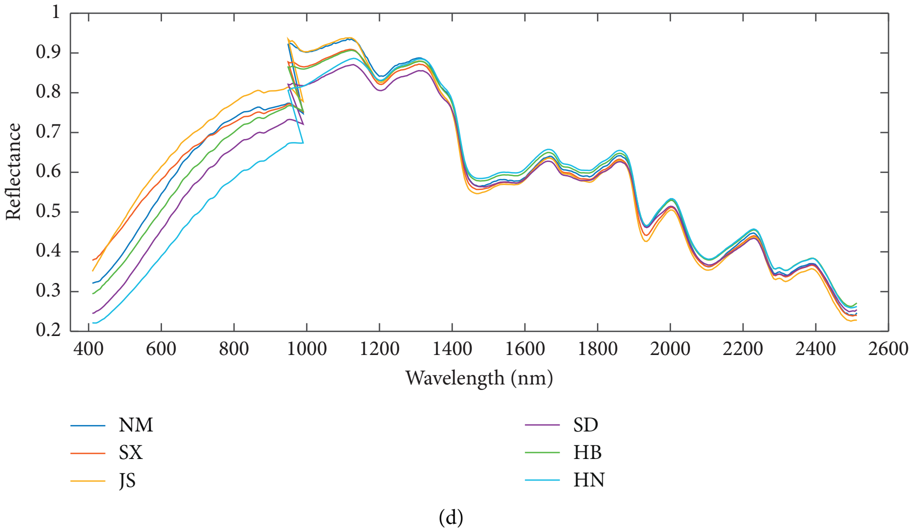


Figure 1: Raw spectra (a), spectra after pretreated by MSC (b), second derivative (c), and mean reflectance spectra (d) of Tiegun yam powder samples from different origins.

Table 3: Pairwise combination classification accuracy of the preprocessing method and classification model of Tiegun yam powder samples from different origins.

Models	Preprocessing	Accuracy (%)	
		Training set	Prediction set
PLS-DA	Raw data	99.57	96.00
	MSC	100.00	98.40
	D1	100.00	96.00
	D2	97.82	95.42
	SG	98.72	95.20
	SNV	99.15	96.80
SVM	Raw data	98.81	61.11
	MSC	97.22	61.11
	D1	98.81	25.93
	D2	34.52	27.78
	SG	98.41	59.26
	SNV	100.00	24.07
RF	Raw data	97.14	68.89
	MSC	99.05	76.67
	D1	99.05	52.22
	D2	100.00	83.33
	SG	98.09	77.78
	SNV	97.14	68.89

variables of spectral data is shown in Figure 3(b). The precision and sensitivity of different producing areas are both above 92%, which is not different from full-wavelength identification, thus showing a good performance.

3.4. Starch, Polysaccharide, and Protein Content Prediction. The prediction results of Tiegun yam powder showed that the three models had better prediction ability for starch content but worse prediction ability for polysaccharide and protein content than starch. Spectral data preprocessing is an important step in chemometrics modeling. Its purpose is to reduce the error caused by background, noise, and other physical factors so as to improve the prediction ability and stability of the model. Tamburini et al. [36] have also reported that MSC, SNV, D1, and D2 can improve the accuracy of PLSR models. In contrast, the Caporaso study [37] showed that the accuracy of the model was not improved when the MSC, SNV, and D1 or D2 were applied.The prediction results of different pretreatment combined with three models for starch, polysaccharide and protein are shown in Tables S1, S2, and S3. MSC, D2, SG smooth, and SNV pretreatment methods can all improve the accuracy of the model, but the improved accuracy varies significantly according to different components and models. The first derivative does not apply to the prediction model. As shown in Table 4, R^2 values of the training set and the prediction set

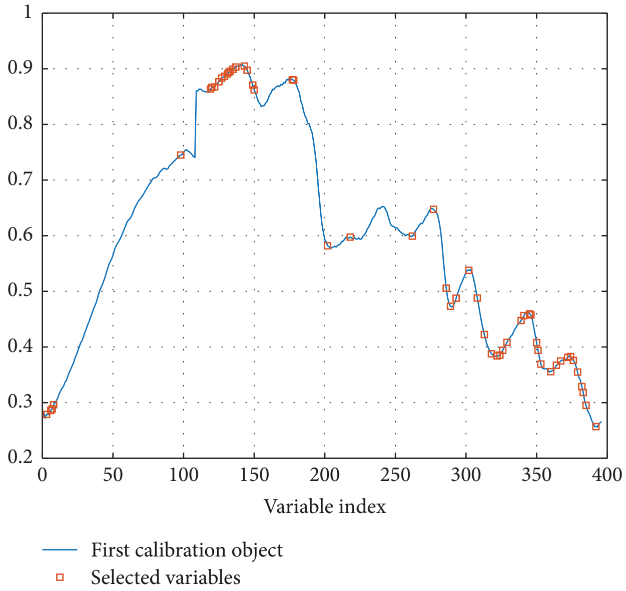
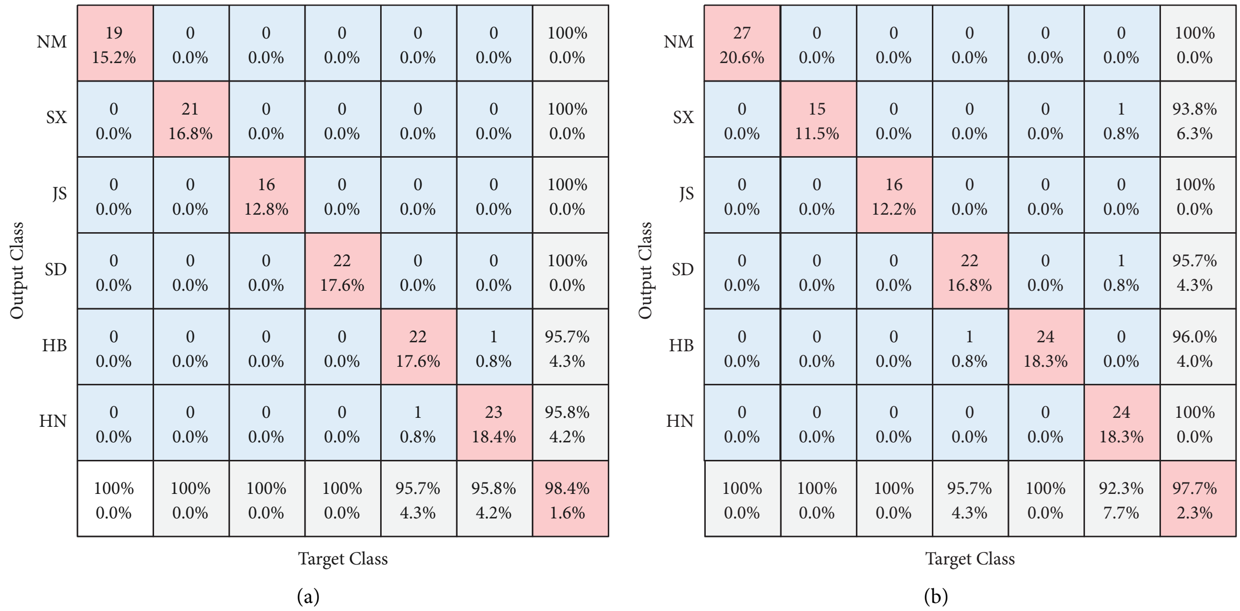


Figure 2: Screening results of SPA characteristic variables.



(a)

NM	19 15.2%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
SX	0 0.0%	21 16.8%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
JS	0 0.0%	0 0.0%	16 12.8%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
SD	0 0.0%	0 0.0%	0 0.0%	22 17.6%	0 0.0%	0 0.0%	100% 0.0%
HB	0 0.0%	0 0.0%	0 0.0%	0 0.0%	22 17.6%	1 0.8%	95.7% 4.3%
HN	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 0.8%	23 18.4%	95.8% 4.2%
	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	95.7% 4.3%	95.8% 4.2%	98.4% 1.6%

(b)

NM	27 20.6%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
SX	0 0.0%	15 11.5%	0 0.0%	0 0.0%	0 0.0%	1 0.8%	93.8% 6.3%
JS	0 0.0%	0 0.0%	16 12.2%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
SD	0 0.0%	0 0.0%	0 0.0%	22 16.8%	0 0.0%	1 0.8%	95.7% 4.3%
HB	0 0.0%	0 0.0%	0 0.0%	1 0.8%	24 18.3%	0 0.0%	96.0% 4.0%
HN	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	24 18.3%	100% 0.0%
	100% 0.0%	100% 0.0%	100% 0.0%	95.7% 4.3%	100% 0.0%	92.3% 7.7%	97.7% 2.3%

Figure 3: The hyperspectral raw data (a) and hyperspectral data after characteristic wavelength selection (b) results of origin identification prediction set of Tiegun yam powder samples.

Table 4: Prediction results of spectral data of Tiegun yam powder samples after different pretreatments.

Content	Models	Training set		Prediction set		
		R^2	RMSEC (mg/g)	R^2	RMSEP (mg/g)	RPD
Starch	PLSR	0.9598	11.0908	0.9270	13.6931	3.22
	SVR	0.9948	3.8399	0.9636	10.7103	5.21
	RF	0.9896	8.4654	0.9677	14.4532	3.45
Polysaccharide	PLSR	0.8684	2.6837	0.8425	2.9637	2.52
	SVR	0.9996	0.1397	0.9275	2.0717	3.21
	RF	0.9818	1.6933	0.9012	3.5702	1.45
Protein	PLSR	0.8933	2.4077	0.8610	2.6676	2.61
	SVR	0.9943	0.5219	0.8939	2.6464	2.94
	RF	0.9770	1.7388	0.9292	3.5203	1.63

should be significantly different, and the optimal pretreatment method should be selected for different models based on the size of the prediction model of starch RPD value. The optimal prediction model of starch, polysaccharide, and protein was D2-SVR. Among the three models, SVR has the best prediction result, followed by PLSR, and RF has the lowest prediction result. The D2-SVR prediction model of starch had a higher R^2p value (0.9636) and RPD value (5.21) and lower RMSE value (RMSEC = 3.8399 mg/g; RMSEP = 10.7103 mg/g). For the polysaccharide prediction model, the parameters of the optimal model were $R^2p = 0.9275$, RPD = 3.21, RMSEC = 0.1397 mg/g, and RMSEP = 2.0717 mg/g. For the protein prediction model, the parameters of the optimal model were $R^2p = 0.8939$, RPD = 2.94, RMSEC = 0.5219 mg/g, and RMSEP = 2.6464 mg/g, indicating that the prediction models have good accuracy and stability. Therefore, appropriate spectral pretreatment is needed in the starch prediction model to improve the regression model. The D2-SVR prediction model results of the three components are shown in Figure 4.

The characteristic variables were selected based on the SPA method. 39 characteristic wavelengths were selected for the starch regression model, 23 characteristic variables for the polysaccharide regression model, and 48 characteristic variables for the protein model. The modeling results of the model selected above are shown in Table 5. D2-SVM was used to model hyperspectral data selected by characteristic wavelength. RPD values of starch, polysaccharide, and protein models were 4.45, 2.05, and 2.18, respectively. The results show that the characteristic wavelength modeling of protein extraction can obtain almost the same results as that of the full-wavelength modeling. The selection of SPA characteristic wavelengths revealed the important spectral regions for predicting the index components of Tiegun yam powder. In addition, modeling results similar to those of full-wavelength can be obtained, which greatly reduces the difficulty of model data processing and reduces the time of model operation.

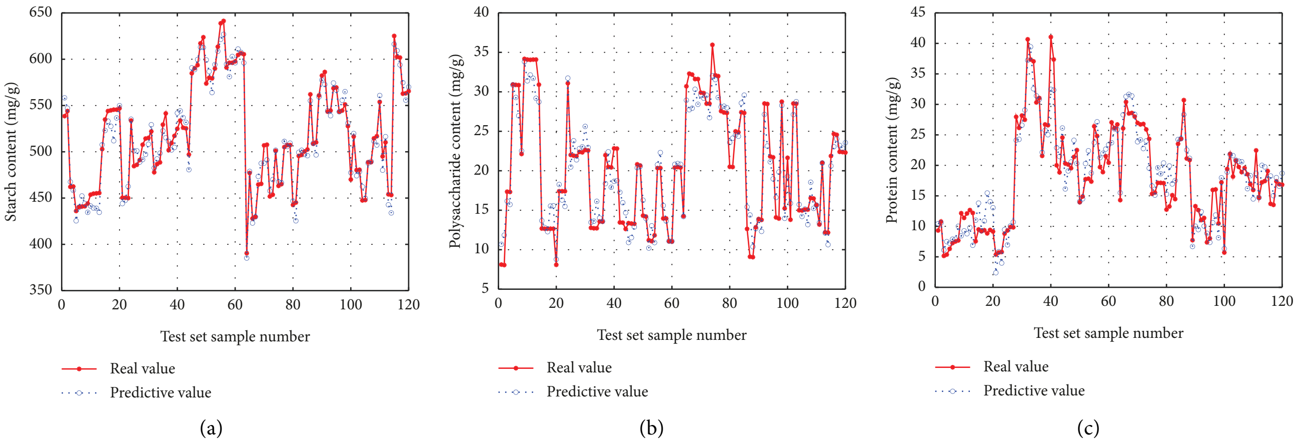


Figure 4: D2-SVR prediction model results for starch (a), polysaccharide (b), and protein (c).

Table 5: Prediction results of selected data by SPA of Tiegun yam powder samples.

Content	Number	Training set		Prediction set		
		R^2	RMSEC (mg/g)	R^2	RMSEP (mg/g)	RPD
Starch	39	0.9943	4.0396	0.9521	12.0895	4.45
Polysaccharide	23	0.9992	0.2161	0.8145	3.5468	2.05
Protein	48	0.9608	1.4008	0.8259	3.3886	2.18

4. Conclusions

This study indicated that it was feasible to use hyperspectral technology combined with chemometric pretreatment and modeling to fast and nondestructive identification of the origin and nondestructive detection of starch, polysaccharide, and protein contents of the Tiegun yam powder samples. Some spectral pretreatment methods (MSC and SNV) can improve the accuracy of hyperspectral data, while others (first-order derivation) are not suitable for the content regression model. MSC combined with PLS-DA is the best combination for the discriminant model, and the accuracy of the training set and the prediction set is over 98%. D2-SVR was the best pretreatment method for starch, polysaccharide, and protein prediction models with relatively high R^2P and RPD values. The characteristic wavelength extracted by the SPA method can achieve similar results to the full-wavelength model, which can greatly reduce the complexity of the model and can reduce the operation time of the model. This study demonstrates the great potential of using hyperspectral images to quickly and nondestructively determine the indicator components of samples, which will be helpful for further prediction of other chemical components in Tiegun yam or applied to other materials with the homology of medicine and food.

Data Availability

The data used to support the findings of this study are included in the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Yue Zhang conceptualised the study, curated the data, wrote of the manuscript, and reviewed and edited the data. Yuan Li performed the investigation. Cong Zhou curated and investigated the data. Junhui Zhou provided the resources and supervised the study. Tiegui Nan supervised the study. Jian Yang provided the resources, developed the methodology, conceptualised and supervised the study, and acquired the funding. Luqi Huang provided the resources, conceptualised and supervised the study, and acquired the funding.

Acknowledgments

This work was supported by the National Key R&D Program of China (Grant no. 2020YFC1712700), the scientific and technological innovation project of the China Academy of

Chinese Medical Sciences (Grant nos. CI2021A04005 and CI2021A01809), the China Agriculture Research System of MOF and MARA (Grant no. CARS-21), and the Shandong Provincial Key Research and Development Program (a Major Technological Innovation Project) (Grant no. 2021CXGC010508).

Supplementary Materials

The supplementary materials included prediction results of Starch, polysaccharide, and protein from different pretreatment spectral data of Tiegun yam powder samples in Result 3.4, which are shown in table S1, S2, and S3. (*Supplementary Materials*)

References

[1] C. G. Lyu, J. Yang, T. L. Wang et al., “A field trials-based authentication study of conventionally and organically grown Chinese yams using light stable isotopes and multi-elemental analysis combined with machine learning algorithms,” *Food Chemistry*, vol. 343, Article ID 128506, 2021.

[2] L. An, Y. L. Yuan, J. W. Ma et al., “NMR-based metabolomics approach to investigate the distribution characteristics of metabolites in Dioscorea opposita Thunb. cv. Tiegun,” *Food Chemistry*, vol. 298, Article ID 125063, 2019.

[3] S. Guo, X. Y. Zhao, Y. Ma, Y. B. Wang, and D. Wang, “Fingerprints and changes analysis of volatile compounds in fresh-cut yam during yellowing process by using HS-GC-IMS,” *Food Chemistry*, vol. 369, Article ID 130939, 2022.

[4] S. Y. Zhou, G. L. Huang, and G. Y. Chen, “Extraction, structural analysis, derivatization and antioxidant activity of polysaccharide from Chinese yam,” *Food Chemistry*, vol. 361, Article ID 130089, 2021.

[5] F. X. Wang, C. G. Wang, S. Y. Song, S. S. Xie, and F. L. Kang, “Study on starch content detection and visualization of potato based on hyperspectral imaging,” *Food Sciences and Nutrition*, vol. 9, no. 8, pp. 4420–4430, 2021.

[6] P. Meise, S. Seddig, R. Uptmoor, F. Ordon, and A. Schum, “Assessment of yield and yield components of starch potato cultivars (Solanum tuberosum L.) under nitrogen deficiency and drought stress conditions,” *Potato Research*, vol. 62, no. 2, pp. 193–220, 2019.

[7] D. D. Xu, W. Zheng, Y. Q. Zhang, Q. P. Gao, M. X. Wang, and Y. Gao, “A method for determining polysaccharide content in biological samples,” *International Journal of Biological Macromolecules*, vol. 107, pp. 843–847, 2018.

[8] M. A. Galvão, M. R. Ferreira, B. M. Nunes, A. S. Santana, K. P. Randau, and L. A. Soares, “Validation of a spectrophotometric methodology for the quantification of polysaccharides from roots of Operculina macrocarpa (jalapa),” *Revista Brasileira de Farmacognosia*, vol. 24, no. 6, pp. 683–690, 2014.

[9] K. Raymond, J. M. Lacey, G. Dimitar et al., “Mucopolysaccharides quantitation in serum by liquid chromatography-tandem mass spectrometry,” *Molecular Genetics and Metabolism*, vol. 123, p. S123, 2018.

[10] S. Pakfetrat, S. Amiri, M. Radi, E. Abedi, and L. Torri, “The influence of green tea extract as the steeping solution on nutritional and microbial characteristics of germinated wheat,” *Food Chemistry*, vol. 332, Article ID 127288, 2020.

[11] C. Y. Ma, Z. S. Ren, Z. H. Zhang, J. Du, C. Q. Jin, and X. Yin, “Development of simplified models for nondestructive testing of rice (with husk) protein content using hyperspectral imaging technology,” *Vibrational Spectroscopy*, vol. 114, Article ID 103230, 2021.

[12] B. Wang, J. L. He, S. J. Zhang, and L. L. Li, “Nondestructive prediction and visualization of total flavonoids content in Cerasus Humilis fruit during storage periods based on hyperspectral imaging technique,” *Journal of Food Process Engineering*, vol. 44, no. 10, 2021.

[13] G. Y. Deng, S. W. Guo, F. Zaman, T. Y. Li, and Y. Q. Huang, “Recent advances in animal origin identification of gelatin-based products using liquid chromatography-mass spectrometry methods: a mini review,” *Reviews in Analytical Chemistry*, vol. 39, no. 1, pp. 260–271, 2020.

[14] A. M. Li, S. L. Duan, Y. T. Dang et al., “Origin identification of Chinese Maca using electronic nose coupled with GC-MS,” *Scientific Reports*, vol. 9, no. 1, Article ID 12216, 2019.

[15] H. Nakanishi, K. Yoneyama, M. Hara, A. Takada, and K. Saito, “The origin identification method for crude drugs derived from arthropods and annelids using molecular biological techniques,” *Journal of Natural Medicines*, vol. 74, no. 1, pp. 275–281, 2020.

[16] J. Zhang, Z. Q. Tian, Y. Q. Ma et al., “Origin identification of the sauce-flavor Chinese baijiu by organic acids, trace elements, and the stable carbon isotope ratio,” *Journal of Food Quality*, vol. 2019, pp. 1–7, 2019.

[17] H. W. Gu, X. L. Yin, T. Q. Peng et al., “Geographical origin identification and chemical markers screening of Chinese green tea using two-dimensional fingerprints technique coupled with multivariate chemometric methods,” *Food Control*, vol. 135, Article ID 108795, 2022.

[18] G. Elmasry, M. Kamruzzaman, D. W. Sun, and P. Allen, “Principles and applications of hyperspectral imaging in quality evaluation of agro-food products: a review,” *Critical Reviews in Food Science and Nutrition*, vol. 52, no. 11, pp. 999–1023, 2012.

[19] H. P. Huang, X. J. Hu, J. P. Tian et al., “Rapid and nondestructive prediction of amylose and amylopectin contents in sorghum based on hyperspectral imaging,” *Food Chemistry*, vol. 359, Article ID 129954, 2021.

[20] X. L. Bai, C. Zhang, Q. L. Xiao, Y. He, and Y. D. Bao, “Application of near-infrared hyperspectral imaging to identify a variety of silage maize seeds and common maize seeds,” *RSC Advances*, vol. 10, no. 20, pp. 11707–11715, 2020.

[21] Z. Z. Bai, X. J. Hu, J. P. Tian, P. Chen, H. B. Luo, and D. Huang, “Rapid and nondestructive detection of sorghum adulteration using optimization algorithms and hyperspectral imaging,” *Food Chemistry*, vol. 331, Article ID 127290, 2020.

[22] C. Zhang, W. Y. Wu, L. Zhou, H. Cheng, X. Q. Ye, and Y. He, “Developing deep learning based regression approaches for determination of chemical compositions in dry black goji berries (Lycium ruthenicum Murr.) using near-infrared hyperspectral imaging,” *Food Chemistry*, vol. 319, Article ID 126536, 2020.

[23] J. He, L. D. Chen, B. Q. Chu, and C. Zhang, “Determination of total polysaccharides andtotal flavonoids in Chrysanthemum morifolium using near-infrared hyperspectral imaging and multivariate analysis,” *Molecules*, vol. 23, no. 9, p. 2395, 2018.

[24] L. Zhang, Y. Q. Wang, Y. G. Wei, and D An, “Near-infrared hyperspectral imaging technology combined with deep convolutional generative adversarial network to predict oil content of single maize kernel,” *Food Chemistry*, vol. 370, Article ID 131047, 2022.

[25] X. Chu, R. Li, H. Y. Wei et al., “Determination of total flavonoid and polysaccharide content in Anoectochilus

formosanus in response to different light qualities using hyperspectral imaging," *Infrared Physics & Technology*, vol. 122, Article ID 104098, 2022.

[26] I. da Silva Lindemann, C. Lambrecht Dittgen, C. de Souza Batista et al., "Rice and common bean blends: effect of cooking on in vitro starch digestibility and phenolics profile," *Food Chemistry*, vol. 340, Article ID 127908, 2021.

[27] H. Q. Zhao, Z. B. Wang, Y. P. Sun, C. J. Yang, B. Y. Yang, and H. X. Kuang, "Advances in isolation, identification and bioactivity of Fritillaria polysaccharides," *Chinese Traditional Patent Medicine*, vol. 44, pp. 505–510, 2020.

[28] H. Lin, S. H. Gui, B. B. Yu, X. H. Que, and J. Q. Zhu, "Analysis of polysaccharides and monosaccharides from Rehmannia glutinosa by different processing technology and effects on ovarian granulosa cells," *Chinese Traditional Patent Medicine*, vol. 41, pp. 2958–2963, 2019.

[29] I. Baek, H. Lee, B. K. Cho, C. Mo, D. E. Chan, and M. S. Kim, "Shortwave infrared hyperspectral imaging system coupled with multivariable method for TVB-N measurement in pork," *Food Control*, vol. 124, Article ID 107854, 2021.

[30] W. D. Zhang, A. L. Cao, P. Y. Shi, and L. Y. Cai, "Rapid evaluation of freshness of largemouth bass under different thawing methods using hyperspectral imaging," *Food Control*, vol. 125, Article ID 108023, 2021.

[31] A. Borin, M. F. Ferrão, C. Mello, D. A. Maretto, and R. J. Poppi, "Least-squares support vector machines and near infrared spectroscopy for quantification of common adulterants in powdered milk," *Analytica Chimica Acta*, vol. 579, no. 1, pp. 25–32, 2006.

[32] W. Lan, S. Wang, Y. Wu et al., "A novel fluorescence sensing strategy based on nanoparticles combined with spectral splicing and chemometrics for the recognition of *Citrus reticulata* 'Chachi' and its storage year," *Journal of the Science of Food and Agriculture*, vol. 100, no. 11, pp. 4199–4207, 2020.

[33] J. Ma and D. W. Sun, "Prediction of monounsaturated and polyunsaturated fatty acids of various processed pork meats using improved hyperspectral imaging technique," *Food Chemistry*, vol. 321, Article ID 126695, 2020.

[34] L. Yang, H. Q. Gao, L. W. Meng et al., "Nondestructive measurement of pectin polysaccharides using hyperspectral imaging in mulberry fruit," *Food Chemistry*, vol. 334, Article ID 127614, 2021.

[35] X. Li, F. Feng, R. Z. Gao et al., "Application of near infrared reflectance (NIR) spectroscopy to identify potential PSE meat," *Journal of the Science of Food and Agriculture*, vol. 96, no. 9, pp. 3148–3156, 2016.

[36] E. Tamburini, E. Mamolini, M. De Bastiani, M. G. Marchetti, and W. R. Seitz, "Quantitative determination of Fusarium proliferatum concentration in intact garlic cloves using near-infrared spectroscopy," *Sensors*, vol. 16, p. 1099, 2016.

[37] N. Caporaso, M. B. Whitworth, and I. D. Fisk, "Total lipid prediction in single intact cocoa beans by hyperspectral chemical imaging," *Food Chemistry*, vol. 344, Article ID 128663, 2021.