Journal of Food Composition and Analysis 116 (2023) 105048

Contents lists available at ScienceDirect

Journal of Food Composition and Analysis

journal homepage: www.elsevier.com/locate/jfca

Application of visible-near-infrared hyperspectral imaging technology coupled with wavelength selection algorithm for rapid determination of moisture content of soybean seeds

Zhen Guo [a], Jing Zhang [a], Chengye Ma [a], Xiang Yin [a], Yemin Guo [a,b,c,*], Xia Sun [a,b,c], Chengqian Jin [a,d,*]

[a] School of Agricultural Engineering and Food Science, Shandong University of Technology, Zibo, Shandong 255049, China
[b] Shandong Provincial Engineering Research Center of Vegetable Safety and Quality Traceability, Zibo, Shandong 255049, China
[c] Zibo City Key Laboratory of Agricultural Product Safety Trace ability, Zibo, Shandong 255049, China
[d] Nanjing Research Institute for Agricultural Mechanization, Ministry of Agriculture and Rural Affairs, Nanjing, Jiangsu 210014, China

ARTICLE INFO

Keywords:
Hyperspectral imaging
Wavelength selection
Visible-near-infrared
Moisture content
Food composition
Food analysis

ABSTRACT

Moisture content is a crucial factor affecting the quality of soybean seeds. However, the determination of moisture content of soybean seeds is time-consuming and expensive. In this study, visible-near-infrared hyperspectral imaging technology (400–1000 nm) coupled with wavelength selection algorithm was applied to determine the moisture content of soybean seeds. Hyperspectral images of 96 soybean samples were obtained, and the sample set partitioning based on joint x-y distance algorithm was used to divide the calibration and prediction sets after removing outliers. Then, partial least squares regression (PLSR) models based on the original and preprocessing spectra were established, and the prediction effect of the original spectra was better than that of the preprocessing spectra. Five wavelength selection algorithms were used to select feature wavelengths to optimize the models further. Each wavelength selection algorithm was run 100 times independently to investigate its stability. The PLSR models were established based on the results of wavelength selection, and the prediction effects of all models were statistically analyzed. Results showed that the combination of interval variable iterative space shrinkage approach and successive projections algorithm (IVISSA-SPA) based on the original spectra was the most suitable model for the determination of moisture content of soybean seeds. The prediction accuracies of the IVISSA-SPA model were $R_P^2 = 0.9713 \pm 0.0044$, RMSEP $= 0.307 \pm 0.021$ and RPD $= 6.058 \pm 0.344$ in 100 independent experiments. Results indicated that visible-near-infrared hyperspectral imaging coupled with wavelength selection algorithm provided a rapid method for determining the moisture content of soybean seeds.

1. Introduction

Soybean is an important cash crop in the world and is one of the crucial feed sources for the breeding industry (Kusumaningrum et al., 2018). In addition, soybean plays a key role in extracting vegetable oil, plant protein, and soybean products (Han et al., 2014). Moisture content affects the quality and storage period of soybean seeds (Ziegler et al., 2021). The moisture content of soybean seeds used for sale, storage, and processing in China should not exceed 12 %. In addition, moisture content affects the vigor of soybean seeds during the breeding process (Finch-Savage and Bassel, 2016), and the control of moisture content is an important link to ensuring seed quality. Therefore, evaluating the moisture content of soybean seeds accurately is crucial. Conventional methods in moisture content determination include the oven drying method (Butts et al., 2014) and the electronic moisture analyzer (Mireei et al., 2016). The principle of the oven drying method is to calculate the moisture content by comparing the changes of sample weight before and after drying. However, this method is time-consuming and destructive. Compared with the oven drying method, the electronic moisture analyzer is simpler to use and more portable. However, the limitation of

* Corresponding authors at: School of Agricultural Engineering and Food Science, Shandong University of Technology, Zibo, Shandong 255049, China
E-mail addresses: 806559062@qq.com (Z. Guo), 1820787057@qq.com (J. Zhang), mcycn2002@sdut.edu.cn (C. Ma), 252751968@qq.com (X. Yin), gym@sdut.edu.cn (Y. Guo), sunxia2151@sina.com (X. Sun), 412114402@qq.com (C. Jin).

https://doi.org/10.1016/j.jfca.2022.105048
Received 27 June 2022; Received in revised form 12 October 2022; Accepted 16 November 2022
Available online 21 November 2022
0889-1575/

this method lies in its large error. A rapid and non-destructive determination of moisture content in soybean seeds is thus crucial.

Hyperspectral imaging technology is an emerging rapid and non-destructive detection method. Hyperspectral images contain image information at a certain wavelength and the spectral information of each pixel in the image (ElMasry and Nakauchi, 2016; Mo et al., 2017b). Visible-near-infrared refers to the spectral range including visible and near-infrared; the spectral range in this study is 400–1000 nm. Visible-near-infrared hyperspectral imaging technology has been widely used in the detection of the external features of samples, such as biological contaminants in fresh-cut lettuce (Mo et al., 2017a), freshness of organic beef (Crichton et al., 2017), and muscle myopathy in poultry meat (Jiang et al., 2019), and in the detection of internal features of samples, such as soluble solid content of plums (Li et al., 2018), pH of Kyoho grape (Xu et al., 2022), and polyphenols and caffeine in matcha (Ouyang et al., 2021). The principle of this technology is that light is scattered and absorbed when it is irradiated on the surface of the sample. The scattering of light is influenced by cell structure and tissue density, and the absorption of light is more related to the chemical composition of the sample (Li et al., 2018). Therefore, the spectra contain the composition information of the sample, and spectral analysis coupled with chemometrics can achieve the detection of the chemical composition of the sample (Yun et al., 2019; Li et al., 2020). Hyperspectral imaging technology is mainly applied in soybean variety identification (Zhu et al., 2019), soybean seed viability detection (Li et al., 2019), and isoflavones (Kezhu et al., 2014) and fatty acid content detection (Fu et al., 2021). However, no relevant reports on the application of visible-near-infrared hyperspectral imaging technology in the detection of moisture content exist. Rabanera et al. (2021) used hyperspectral imaging technology to determine the moisture content of peanut kernels. However, their research selected wavelengths only used the highest absolute weighted regression coefficient method. Liu et al. (2020) used competitive adaptive reweighted sampling (CARS) to select feature wavelengths when determining the starch content of a single kernel. Xu et al. (2019) used CARS, iteratively retaining informative variables, and random frog to select feature wavelengths to determine the moisture content of cucumber seeds. In the above report, the wavelength selection algorithm was used in the process of moisture content determination of samples, but the stability of the algorithm was ignored.

Hyperspectral images contain hundreds of wavelengths. Given the correlation between adjacent wavelengths, the hyperspectral images carry substantial redundant and collinear information. Useless information may affect the stability of the model and reduce the calculation speed (Wu and Sun, 2013). Therefore, the dimension of spectra must be reduced, and feature wavelengths that can improve the stability and prediction performance of the model must be selected (Liu et al., 2014).

The wavelength selection algorithm can be divided into wavelength interval selection and wavelength point selection (Sun et al., 2020). The wavelength interval selection algorithm selects several groups of continuous wavelength interval combinations as the variables. The established model has a good interpretation. The wavelength point selection algorithm selects wavelengths that are discretely distributed, which have the advantage of a better prediction effect. The combination of different wavelength selection algorithms can combine their advantages to achieve improved results. The former algorithm removes useless information wavelengths and retains key wavelengths. The latter algorithm needs to select important wavelengths accurately while reserving fewer wavelengths to achieve improved modeling results (Yu et al., 2020). Therefore, this study adopted CARS, the variable combination population analysis (VCPA), and the successive projections algorithm (SPA) as wavelength point selection algorithms and the interval variable iterative space shrinkage approach (IVISSA) as wavelength interval selection algorithm to select feature wavelengths; all of them were run 100 times to verify its stability. Finally, the combination of IVISSA and SPA (IVISSA-SPA) had improved prediction effects.

The research objectives were as follows: (1) Investigate the feasibility of rapid determination of moisture content in soybean seeds by visible-near-infrared hyperspectral imaging technology. (2) Discuss the influence of preprocessing methods on the prediction effect and analyze the reasons. (3) Verify the stability of the five wavelength selection algorithms by establishing feature wavelength models. (4) Provide a simple model based on the high-frequency feature wavelengths of IVISSA-SPA.

2. Materials and methods

2.1. Sample preparation and moisture content determination

The samples consist of 96 varieties from the market in Zibo, China, including “Zhonghuang 37″, “Heinong 84″, “Shennong 8″, and “Dongdou 1133″, among others. All soybean samples were labeled according to varieties. The moisture content of soybean was determined by the direct drying method from the Chinese national standard (a reference to GB5009.3–2016).

2.2. Visible-near-infrared hyperspectral imaging system

Hyperspectral images were acquired in reflectance mode using a laboratory visible-near-infrared hyperspectral imaging system (spectral range 400–1000 nm) (Isuzu Optics Corp., Taiwan, China). The visible-near-infrared hyperspectral imaging system consists of a line-scan spectrograph (400–1000 nm) (Specim., Oulu, Finland), a 2560 × 2160 CCD camera (Andor., Belfast, Ireland), a camera lens (Specim., Oulu, Finland), two lamps and a light source (Illumination Technologies Inc., New York, USA), a mobile platform (Isuzu Optics Corp., Taiwan, China), and a computer used to process spectral data (DELL., Round Rock, USA). To eliminate the influence of ambient light, all the components were integrated into a dark box (Isuzu Optics Corp., Taiwan, China). Fig. 1 shows the principal components of the visible-near-infrared hyperspectral imaging system.

2.3. Hyperspectral image acquisition and correction

Soybean seeds were placed in 96 culture dishes (φ9 cm × 1 cm) according to varieties, and all of the culture dishes were placed on the

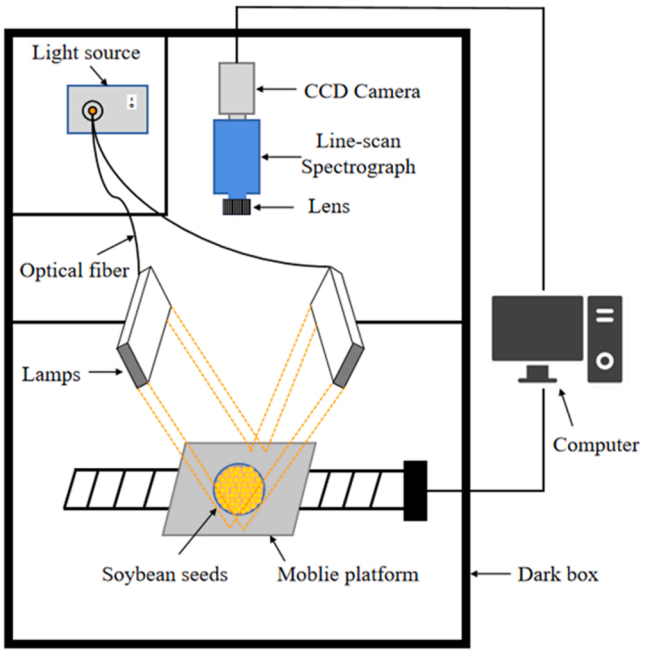

Fig. 1. Visible-near-infrared hyperspectral imaging system.

mobile platform to collect hyperspectral images in sequence. Fig. 2 shows the soybean seeds. To eliminate the influence of dark current and ambient noise on the hyperspectral image, dark and white reference calibration images were used to correct the original hyperspectral image. The whiteboard (Specim., Oulu, Finland) with a reflectivity of 0.99 was placed in the sample acquisition area, and the collected image was recorded as I_w as the white reference calibration image. Then, the lens cover of the CCD camera was covered, and the collected image was recorded as I_d as the black reference calibration image. Finally, the relevant correction formula was as follows:

$$R_T = \frac{I - I_d}{I_w - I_d} \quad (1)$$

Where, R_T is the corrected sample image and I is the original sample image.

During sample scanning, the exposure time was 2.9 ms, the speed of the mobile platform was 15.34 mm/s, and the object distance was 400 mm.

2.4. Identification of regions of interest and spectral extraction

The hyperspectral imaging analyzer (Isuzu Optics Corp., Taiwan, China) was used to select regions of interest. The tool for selecting regions of interest was used to select a circular area with a radius of 650 pixels, which was large enough to represent the spectral information of each soybean variety. Three regions of interest were extracted from each image. Then, the spectral data of each pixel point were averaged at each wavelength to obtain a spectral curve (Sun et al., 2019). Finally, the spectra extracted from the three regions of interest were averaged to represent the samples.

2.5. Multivariate calibration methods

In the process of spectral acquisition, environment and instrument would cause data errors; thus, the abnormal values were eliminated. Monte Carlo–partial least squares (MCPLS) (Guo et al., 2012) had the advantage of detecting spectral outliers, physical and chemical reference outliers simultaneously. The specific method was to take all the sample sets as a correction set and establish a model. The minimum root mean square errors of cross-validation was the best principal component score. Then, the mean and standard deviation of the prediction error was determined for each sample. Finally, the scatter diagram was drawn with the mean as the abscissa and the standard deviation as the ordinate. In addition, 2.5 times of the standard deviation and mean as the limit values and those greater than the limit values were abnormal samples. Fig. 3 shows that the mean value or standard deviation of the prediction error of samples 71, 10, 29, 72, 23, and 20 were large, and these samples were eliminated in turn. After samples 71, 72 and 29 were eliminated, the modeling R_C^2 value increased from 0.9383 to 0.9652. After excluding samples 10, 23 and 20, the modeling R_C^2 value decreased from 0.9652 to 0.9502, indicating that these three samples were not abnormal values. Therefore, 93 samples were reserved for modeling and prediction after excluding outliers.

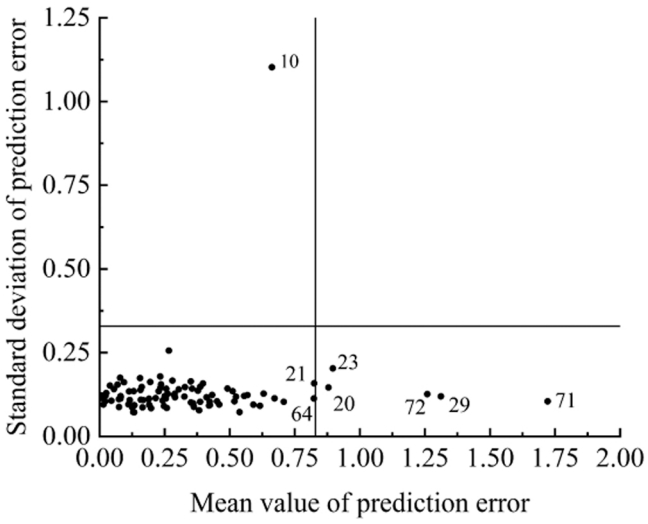

Fig. 3. Monte Carlo-partial least squares detection.

The samples set partitioning based on joint x–y distance (Sun et al., 2021; Wang et al., 2021a) algorithm selected 70 samples as the calibration set, and the remaining 23 samples comprised the prediction set. Table 1 shows that the moisture content range of the calibration set covered the prediction set, indicating that the sample set was reasonably divided.

2.6. Chemometric analysis

Partial least squares regression (PLSR) (Wold et al., 2001) was established to correlate the spectra with the moisture content of soybean seeds. PLSR could indirectly describe the relationship between independent variables and dependent variables by constructing a linear regression model, which solved the problem of high linear correlation in each spectral variable set effectively.

2.7. Model evaluation

Model evaluation was a crucial part of the research, and the performance of the model was usually measured by coefficient of determination (R^2) of cross-validation (R_{CV}^2), calibration (R_C^2), and prediction (R_P^2), respectively, by root mean square errors (RMSE) of cross-validation (RMSECV), calibration (RMSEC), and prediction (RMSEP), respectively, and by residual prediction deviation (RPD) of prediction. Generally, a model with improved predictive performance should have higher R^2 and RPD, and lower RMSE (Sun et al., 2019).

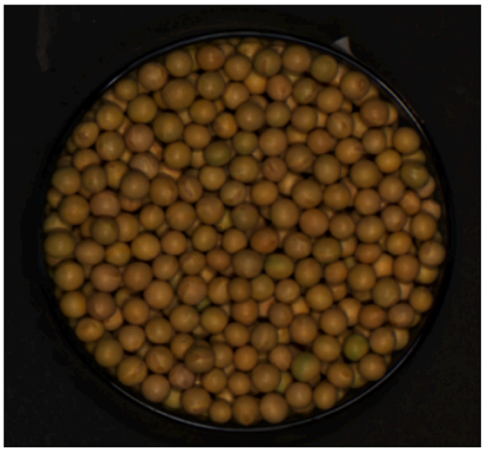

Fig. 2. Soybean seeds.

Table 1
Moisture contents (g/100 g) of soybean seeds.

Indexes	Calibration set	Prediction set
Number of soybean varieties	70	23
Minimum	6.12	6.13
Maximum	11.54	11.03
Mean	8.51	8.08
Standard deviation	1.93	1.85
Range	5.42	4.90

2.8. Feature wavelength selection

Hyperspectral images contain substantial redundant information. To improve the running speed of the model and reduce the dimension of the spectra, CARS, VCPA, IVISSA, SPA, and IVISSA-SPA were used to select the feature wavelengths.

CARS (Li et al., 2009) was a wavelength selection algorithm that took the regression coefficient as the wavelength importance index. It could effectively reduce the influence of collinear variables on the model and remove useless variables. VCPA (Yun et al., 2015; Yang et al., 2017) was a wavelength selection algorithm based on Darwin's theory of "survival of the fittest." The algorithm mainly applied exponential decay function (EDF), binary matrix sampling, and model population analysis to select the optimal subset of wavelengths from the wavelength space. IVISSA (Song et al., 2016) was a selected optimal wavelength algorithm based on the model population analysis. Global search and local search were used in IVISSA to optimize the position, width, and combination of spectral intervals iteratively and intelligently. SPA (Araújo et al., 2001) was a forward selection algorithm for optimal wavelengths. It could select the least redundant variable to solve the collinearity problem.

3. Results and discussion

3.1. Spectrum characteristics analyses

The original spectra acquired by the hyperspectral imaging system were in the range of 400–965 nm. Fig. 4 shows the mean spectral curves extracted from the regions of interest of 93 soybean samples. Fig. 4 showed that 93 spectral curves had the same trends. The reflectance of spectral slightly decreased in the range of 400–410 nm, and then it showed an upward trend from 410 nm and reached a maximum at 855 nm, followed by a small downward trend. An evident absorption valley was located at 920–935 nm. This absorption valley was probably related to the second overtone O–H and N–H stretches and the third overtone C–H stretches and was close to the weak absorption valley of moisture at 910 nm. An absorption peak was observed at 950 nm, followed by an absorption valley at 960 nm which was caused by the second overtone O–H stretches and was related to moisture.

3.2. Modeling analyses based on full spectra

The pretreatment methods were used to preprocess the original spectra, including Savitzky–Golay smoothing (SG), normalization, baseline, standard normal variate (SNV), detrending, multiplicative scatter correction (MSC) and orthogonal signal correction (OSC). The PLSR model was established based on the calibration set samples using the quantitative relationship between the spectra and moisture content of soybean seed (Wang et al., 2021b). Table 2 shows that the PLSR model established by the original spectra had the highest RPD, which was 5.394. As shown in Fig. 5, the RMSEC and the RMSECV values of the model were 0.285 and 0.362, respectively. The R_P^2 and RMSEP values of the model were 0.9642 and 0.343, respectively. Compared with seven preprocessing spectral models, the PLSR model based on the original spectra had the best prediction effect probably because preprocessing reduced spectral noise but removed some useful spectral information. Therefore, the original spectral model was used in the subsequent study. The spectra in the range of 400–965 nm contained numerous wavelengths, which were not suitable for improving the robustness and running speed of the model. Thus, wavelength selection algorithms were used to select feature wavelengths from the original spectra to optimize the model further.

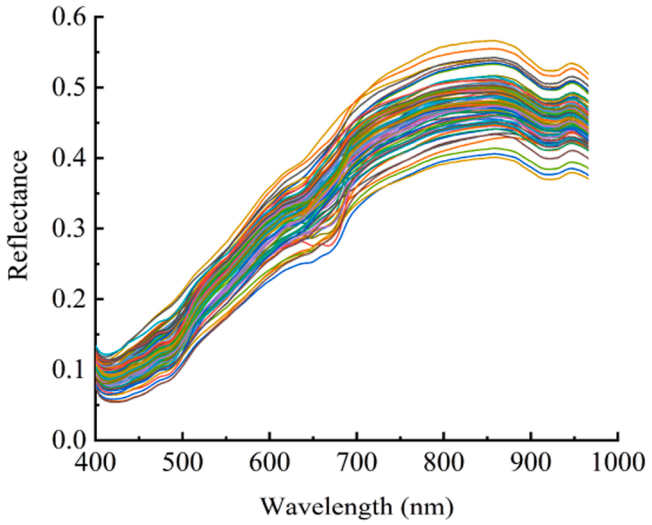

Fig. 4. Reflectance curves of spectral.

3.3. Selection of feature wavelengths

During the process of running the wavelength selection algorithm, the different parameter settings led to different results. After several tests, the optimal parameters of each algorithm were determined. Meanwhile, the stability of the wavelength selection algorithm was discussed in this study. Therefore, uniform parameters were used for each algorithm in 100 independent experiments.

3.3.1. CARS

For CARS, the number of Monte-Carlo sampling (MCS) was set to 500 and the model was evaluated using five-fold cross-validation. Fig. 6(a) shows the variation trend of detection wavelength variables with the sampling runs. When the number of sampling runs was 1 –20 times, the number of wavelength variables decreased sharply from 890 to approximately 50, and then the decline began to slow down. Fig. 6(b) shows the variation trend of RMSECV in the sampling runs. The RMSECV value was lowest when the number of times was 295. Fig. 6(c) shows that each curve recorded the coefficient path of each wavelength at different sampling runs. The position corresponding to * in Fig. 6(c) was 295 MCS runs, in which the RMSECV was the minimum and the number of selected feature wavelengths was also the minimum.

The result of feature wavelengths selected by CARS was not stable each time due to the randomness of MCS. Thus, CARS conducted 100 independent experiments. The mean and standard deviation of the data from the calibration and prediction sets are shown in Table 3. The number of wavelengths retained by CARS was between 20 and 32, which had decreased effectively compared with the full spectra. From Fig. 10 (a), the high-frequency wavelengths were mostly distributed at 401, 403, 407, 428, 474, 475, 519, 556, 858, 884, 889, 917, 942, 958, 959, 960, 961, and 963 nm. Among them, 958, 959, 960, 961, 963 nm were near the absorption valley at 960 nm, which also confirmed that the spectral absorption valley at 960 nm was related to moisture.

3.3.2. VCPA

For VCPA, the number of EDF running was set to 50, and the number of binary matrix sampling running was set to 1000; the model was evaluated using five-fold cross-validation, and the ratio of the optimal subset was 0.1. Fig. 7 shows the variation trend of RMSECV during EDF operation. Overall, RMSECV presented a downward trend, with the repeated operation of the EDF. Under these circumstances, the wavelengths with minimal correlation with moisture content were deleted, and the remaining wavelengths were added to the optimal subset. Finally, the combination of the minimum wavelength variables of RMSECV was selected after EDF bunching.

The result of feature wavelengths selected by VCPA was not stable each time due to the randomness of the EDF. Thus, the VCPA was also run 100 times to obtain statistical results. As shown in Fig. 10 (b), the

Table 2
Prediction results of the PLSR model based on the original and preprocessing spectra.

Spectra type	Calibration set		Cross-validation set		Prediction set		RPD
	R_C^2	RMSEC (%)	R_{CV}^2	RMSECV (%)	R_P^2	RMSEP (%)	
Original spectra	0.9652	0.285	0.9464	0.362	0.9642	0.343	5.394
SG	0.9644	0.288	0.9461	0.367	0.9627	0.350	5.286
Normalization	0.9685	0.270	0.9539	0.339	0.9633	0.347	5.331
Baseline	0.9653	0.284	0.9432	0.373	0.9597	0.364	5.082
SNV	0.9654	0.271	0.9520	0.343	0.9617	0.355	5.211
Detrending	0.9764	0.235	0.9325	0.406	0.9376	0.453	4.084
MSC	0.9665	0.279	0.9531	0.340	0.9621	0.349	5.301
OSC	0.9561	0.320	0.9487	0.353	0.9316	0.474	3.903

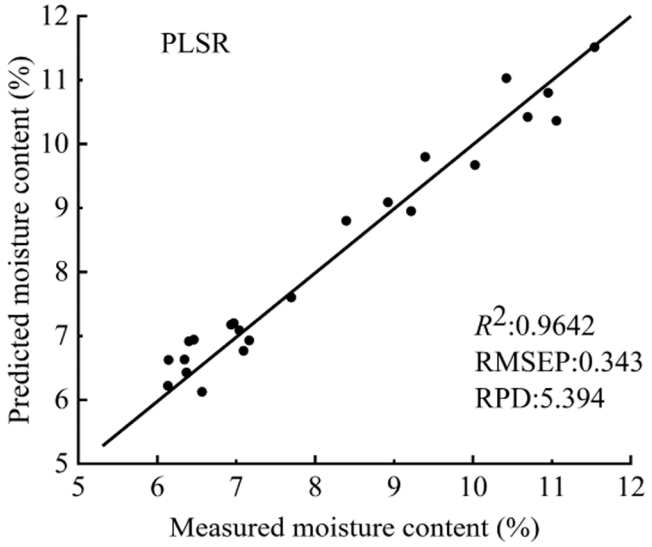

Fig. 5. Effect of PLSR prediction model based on the original spectra.

high-frequency wavelengths were mostly distributed at 532, 557, 858, 884, 889, 917, 948, 958, 959, and 960 nm. VCPA adopted the EDF method to eliminate wavelength variables quickly, but the number of reserved wavelength variables was usually small that some valid information variables might be also eliminated.

3.3.3. IVISSA

For IVISSA, the number of potential variables was set to 4, the model was evaluated using five-fold cross-validation, and the number of binary matrix sampling running was set to 1000. Fig. 8 shows the variation trend of the RMSECV during iterations. Compared with VCPA, the RMSECV value of IVISSA changed more gently and showed a downward trend in the iterative process. After 72 iterations of IVISSA, the RMSECV value decreased to 0.344, and 167 wavelengths were retained.

The selection of feature wavelengths by IVISSA also had certain randomness; thus, the algorithm was run 100 times, and the frequency of statistical wavelengths is shown in Fig. 10 (c). The high-frequency wavelengths were mostly distributed at 532, 557, 858, 884, 889, 917, 948, 958, 959, and 960 nm, which showed a distinct peak. As shown in Table 3, IVISSA optimized approximately 155 feature wavelengths, accounting for approximately 17.4 % of all wavelengths. To improve the robustness and running speed of the model, it was essential to further optimize the feature wavelengths.

3.3.4. SPA

For SPA, the number of potential variables was set from 10 to 20, and the model was evaluated using five-fold cross-validation. The variable variation trend of RMSECV in the process of SPA selecting feature wavelengths is shown in Fig. 9. Specifically, a sharp drop in RMSECV occurred when the number of variables included in the model was less

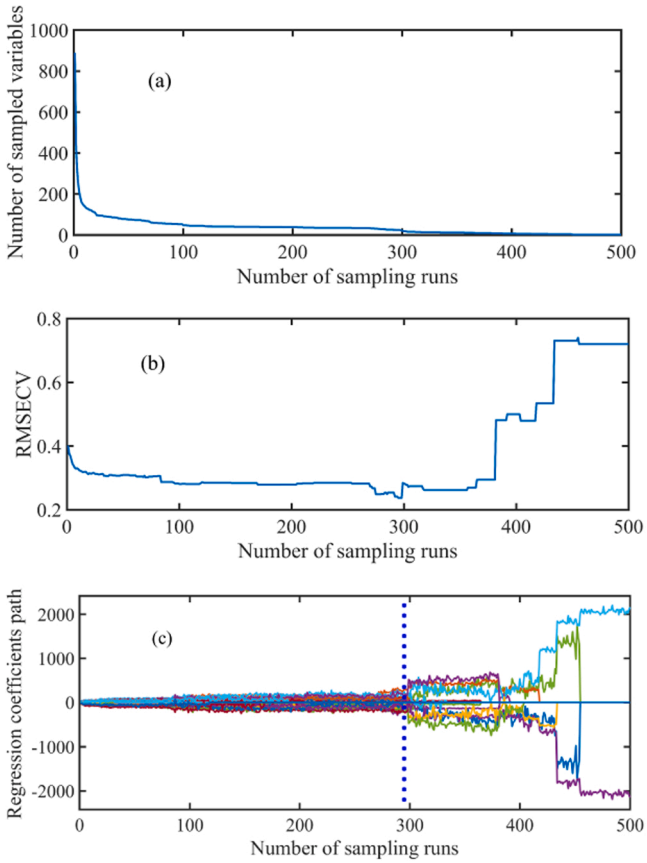

Fig. 6. Wavelengths selection process of CARS. (a) Variation trend of the number of variables with the number of samples, (b) RMSECV and (c) The change process of regression coefficient of each variable with sampling times. The blue line represents the position with the lowest RMSECV.

than 2, and it was followed by a slight fall. Finally, 14 wavelengths were retained, and the RMSECV value was 0.331. After running SPA 100 times, the feature wavelengths selected were consistent because SPA extracted the maximum wavelengths of the projection vector, whereas the projection vector of wavelengths was unchanged. Distribution positions and frequency of feature wavelengths are shown in Fig. 10 (d). The 14 feature wavelengths selected were as follows: 401, 438, 491, 512, 549, 599, 641, 669, 688, 732, 771, 862, 923, and 961 nm.

3.3.5. IVISSA-SPA

SPA was used to select the feature wavelengths after using IVISSA. SPA was selected for secondary selection because its prediction effect was better than those of VCPA and CARS, and the feature wavelengths selected by SPA were deterministic, which could combine the advantages of SPA and VISSA. This feature was also an innovation point of this

Table 3
Prediction results of PLSR model based on different wavelength selection algorithms.

Wavelength Selection algorithm	The number of wavelengths	Calibration set		Cross-validation set		Prediction set		RPD
		R_C^2	RMSEC (%)	R_{CV}^2	RMSECV (%)	R_P^2	RMSEP (%)	
None	890	0.9652	0.285	0.9464	0.362	0.9642	0.343	5.394
CARS	26 ± 6	0.9818 ± 0.0086	0.198 ± 0.023	0.9728 ± 0.0024	0.256 ± 0.010	0.9550 ± 0.0081	0.384 ± 0.035	4.866 ± 0.470
VCPA	10 ± 1	0.9766 ± 0.0027	0.233 ± 0.013	0.9687 ± 0.0103	0.269 ± 0.014	0.9531 ± 0.0321	0.378 ± 0.039	4.938 ± 0.486
IVISSA	155 ± 22	0.9670 ± 0.0057	0.278 ± 0.032	0.9462 ± 0.0517	0.328 ± 0.013	0.9677 ± 0.0060	0.325 ± 0.025	5.710 ± 0.328
SPA	14	0.9672	0.286	0.9527	0.347	0.9702	0.312	6.026
IVISSA-SPA	13 ± 2	0.9633 ± 0.0043	0.293 ± 0.016	0.9465 ± 0.0314	0.346 ± 0.025	0.9713 ± 0.0044	0.307 ± 0.021	6.058 ± 0.344

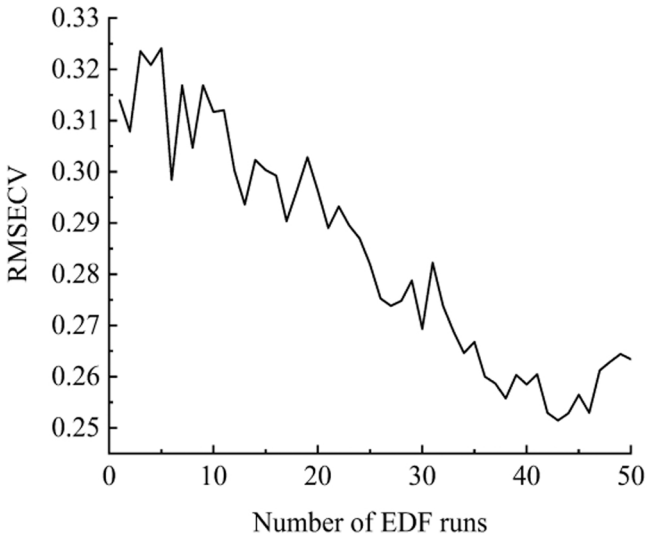

Fig. 7. Wavelength selection by VCPA.

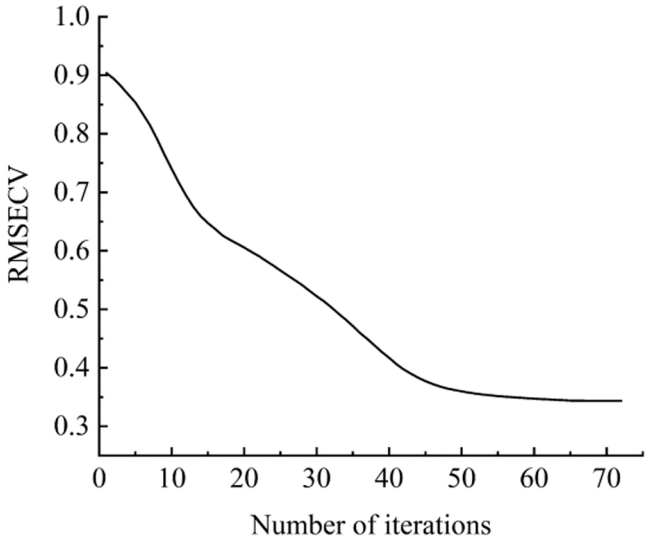

Fig. 8. Wavelength selection by IVISSA.

study.

IVISSA-SPA was also run 100 times, and the prediction results are shown in Table 3. After secondary selection by SPA, the number of feature wavelengths decreased from approximately 155 to approximately 13. The frequency diagram of feature wavelengths selected by IVISSA-SPA within 100 times is shown in Fig. 10 (e). The frequency of

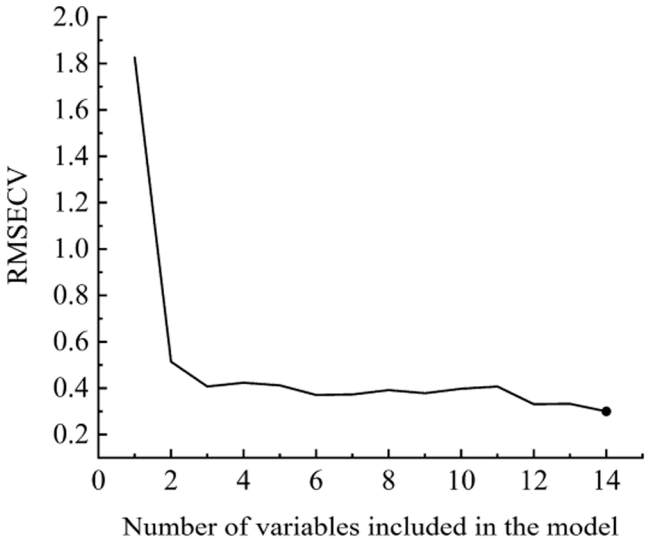

Fig. 9. Wavelength selection by SPA.

feature wavelengths selected by IVISSA-SPA and IVISSA had the same distribution, but the redundant wavelengths within the range of 890–965 nm were effectively removed. From Fig. 10 (e), the high-frequency wavelengths were mostly distributed at 516, 581, 631, 689, 794, 890, 907, 920, 923, 932, 934, 962, 963, 964, and 965 nm.

3.4. Comparison of models

As shown in Table 3, the wavelength selection algorithms could effectively reduce the spectral dimension. Among them, IVISSA, CARS, SPA, VCPA and IVISSA-SPA retained approximately 155, 26, 14, 13 and 10 wavelengths, respectively. The RPD values of CARS and VCPA were lower than that of the original spectra, which might be due to the small number of feature wavelengths retained by CARS and VCPA, and some effective information was removed in the process of wavelength selection. Comparing and analyzing the location of high-frequency wavelength distribution selected by each algorithm from Fig. 10, it revealed that each algorithm selected high-frequency wavelengths at 920–935 nm and 960 nm, which was consistent with the results of spectral characteristics analyses.

Fig. 11 shows the R_P^2 and RMSEP values of 100 PLSR models established by each algorithm, and the box plot intuitively reflects the prediction effect of each algorithm. In general, the R_P^2 values of CARS, IVISSA, VCPA, SPA and IVISSA-SPA showed an upward step distribution in turn, whereas the RMSEP values showed a downward step distribution in turn. The prediction effect of SPA and IVISSA-SPA was remarkably better than that of CARS and VCPA. The mean values of RMSEC and RMSECV of IVISSA-SPA were 0.293 ± 0.016 and 0.346 ± 0.025,

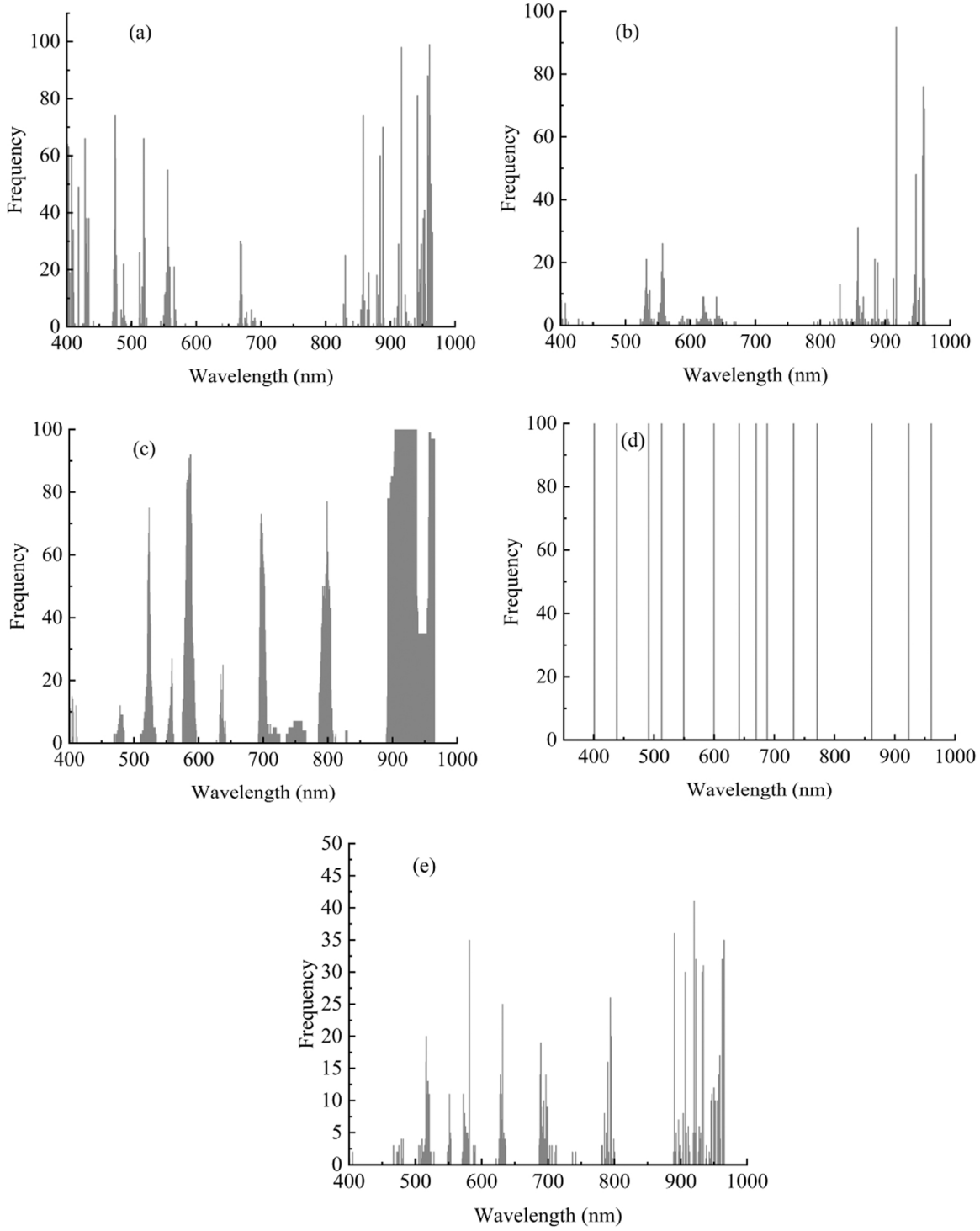

Fig. 10. Results of five wavelength selection algorithms running 100 times independently. (a), (b), (c), (d) and (e) represent the results of CARS, VCPA, IVISSA, SPA and IVISSA-SPA, respectively.

respectively; the R_P^2 value of IVISSA-SPA was 0.9713 ± 0.0044, and the RMSEP value was 0.307 ± 0.021; the mean value of R_P^2 was higher, the mean value of RMSEP was lower, and the standard deviation of R_P^2 and RMSEP was the smallest, indicating that IVISSA-SPA was relatively stable, and the number of selected feature wavelengths was small. Meanwhile, Table 3 shows the RPD value of each algorithm. The RPD value of IVISSA-SPA was higher, which was 6.058 ± 0.344, indicating that the PLSR model established by IVISSA-SPA had a better prediction effect on the moisture content of soybean seeds.

The feature wavelengths selected by IVISSA-SPA were sorted according to frequency, and the top 15 high-frequency wavelengths were 920, 890, 581, 965, 923, 962, 964, 934, 907, 932, 963, 794, 631, 516, and 689 nm. PLSR models were established for the first 8, 9, 10, 11, 12, 13, 14 and 15 high-frequency wavelengths, respectively. PLSR model prediction results based on different numbers of high-frequency wavelengths are shown in Fig. 12. When the number of high-frequency wavelengths was 12, the PLSR model had the highest R_P^2 value and the lowest RMSEP value, which were 0.9717 and 0.305, respectively. Under

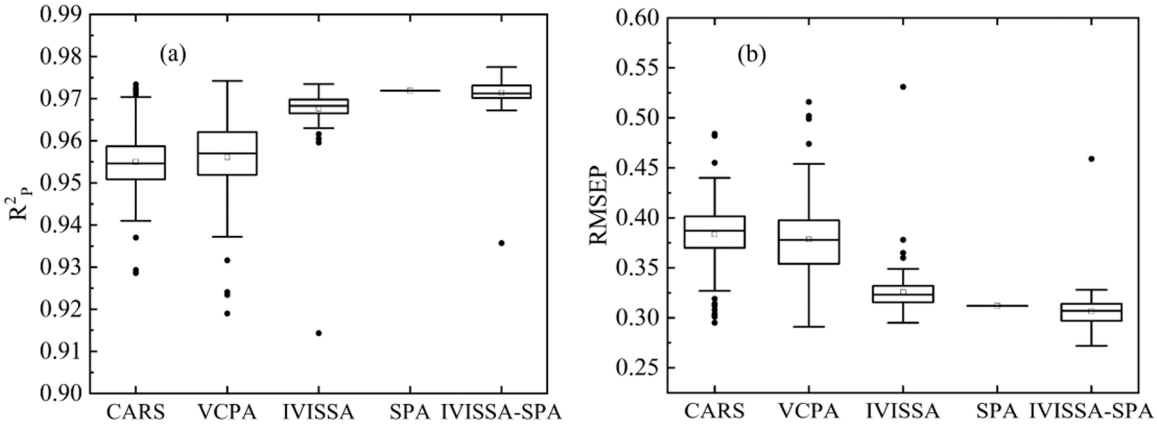

Fig. 11. Box plot of the predicted effect of the model running 100 times independently, (a) and (b) represent the box plot of R_P^2 and RMSEP, respectively.

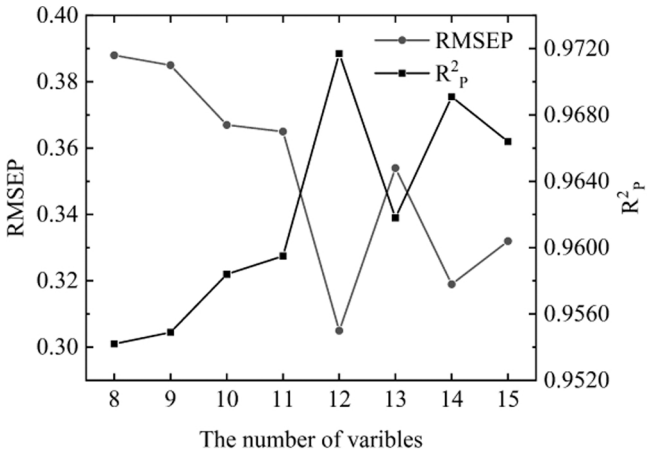

Fig. 12. Results of IVISSA-SPA-PLSR based on different number of variables.

these circumstances, the value of RPD was 6.066, indicating that the top 12 high-frequency wavelengths could have a powerful correlation with the moisture content of soybean seeds. Moreover, 920, 923, 932, and 934 nm were mainly related to the second overtone O–H and N–H stretches and the third overtone C–H stretches at 920–935 nm. In addition, 962, 964, and 963 nm were mainly related to the second overtone O–H stretches at 960 nm.

Compared with other studies, Huang et al. (2014) determined the moisture content of vegetable soybean using visible-near-infrared hyperspectral imaging, and the PLSR model based on the full spectra had a better effect; the R_P^2 was 0.971, and RMSEP was 4.7 %, which were close to the results of this study, but feature wavelength selection was not reported. Sun et al. (2019) determined peanut kernel moisture content in the range of 400–1000 nm, the value of R_P^2 was 0.9363, the value of RMSEP was 0.7021 %, and the value of RPD was 3.988, indicating that this study had a good prediction effect on the moisture content of soybean seeds. The 15 feature wavelengths extracted by SPA were concentrated at 930–1000 nm, and 964 nm was retained, which also indicated that 960 nm was strongly correlated with moisture. Xu et al. (2019) determined the moisture content of cucumber seeds using visible-near-infrared hyperspectral imaging and short-wave infrared hyperspectral imaging, which also showed the feasibility of visible-near-infrared hyperspectral imaging determination of crop seeds. However, their research results showed that the spectral detection in the 1050–2500 nm region was more promising than the moisture content detection in the 400–1000 nm region. In conclusion, visible-near-infrared hyperspectral imaging was a suitable method for determining the moisture content of soybean seeds.

4. Conclusions

This study demonstrated the feasibility of visible-near-infrared hyperspectral imaging technology to determine the moisture content of 96 varieties of soybean seeds. Different algorithms were used to select feature wavelengths, which provided a new simplified stable model for determining the moisture content of soybean seeds. The PLSR model was established for the full spectra, and the original spectra without preprocessing had a better prediction effect. The hybrid wavelength selection algorithm IVISSA combined with SPA had the best performance in wavelength selection. It organically combined the advantages of the two algorithms and improved the model prediction effect while greatly reducing the number of feature wavelengths. In the process of selecting feature wavelengths, CARS, VCPA and IVISSA all had uncertain selection results. Therefore, this study used five wavelength selection algorithms to run 100 independent experiments and then established PLSR models and measured their stability. The results indicated that the distribution of high-frequency feature wavelengths selected by different algorithms was regular, and most of them were concentrated in the vicinity of 920–935 nm and 960 nm, which were related to the spectral absorption bands. The relationship between the feature wavelength selection algorithm and the moisture content of soybean seeds were established, thus providing ideas and references for the determination of soybean in the future.

Ethical Approval

This article has no any study with human participants or animals by any of the authors.

Declaration of Competing Interest

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

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 31772068), Shandong Provincial Science and Technology Achievement Transfer Subsidy (Lu-Yu Science and Technology Cooperation) (LYXZ10) and Zibo City School City Integration Development Project (No. 2019ZBXC090).

References

Araújo, M.C.U., Saldanha, T.C.B., Galvão, R.K.H., Yoneyama, T., Chame, H.C., Visani, V., 2001. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. Chemom. Intell. Lab. Syst. 57 (2), 65–73.

Butts, C.L., Lamb, M.C., Sorensen, R.B., Chen, S., 2014. Oven drying times for moisture content determination of single peanut kernels. Trans. ASABE 57 (2), 10438.

Crichton, S.O.J., Kirchner, S.M., Porley, V., Retz, S., von Gersdorff, G., Hensel, O., Weygandt, M., Sturm, B., 2017. Classification of organic beef freshness using VNIR hyperspectral imaging. Meat Sci. 129, 20–27.

ElMasry, G.M., Nakauchi, S., 2016. Image analysis operations applied to hyperspectral images for non-invasive sensing of food quality – a comprehensive review. Biosyst. Eng. 142, 53–82.

Finch-Savage, W.E., Bassel, G.W., 2016. Seed vigour and crop establishment: extending performance beyond adaptation. J. Exp. Bot. 67 (3), 567–591.

Fu, D., Zhou, J., Scaboo, A.M., Niu, X., 2021. Nondestructive phenotyping fatty acid trait of single soybean seeds using reflective hyperspectral imagery. J. Food Process Eng. 44 (8), 13759.

Guo, W.L., Du, Y.P., Zhou, Y.C., Yang, S., Lu, J.H., Zhao, H.Y., Wang, Y., Teng, L.R., 2012. At-line monitoring of key parameters of nisin fermentation by near infrared spectroscopy, chemometric modeling and model improvement. World J. Microbiol. Biotechnol. 28 (3), 993–1002.

Han, S.I., Chae, J.H., Bilyeu, K., Shannon, J.G., Lee, J.D., 2014. Non-destructive determination of high oleic acid content in single soybean seeds by near infrared reflectance spectroscopy. J. Am. Oil Chemists' Soc. 91 (2), 229–234.

Huang, M., Wang, Q., Zhang, M., Zhu, Q., 2014. Prediction of color and moisture content for vegetable soybean during drying using hyperspectral imaging technology. J. Food Eng. 128, 24–30.

Jiang, H., Yoon, S.C., Zhuang, H., Wang, W., Li, Y., Yang, Y., 2019. Integration of spectral and textural features of visible and near-infrared hyperspectral imaging for differentiating between normal and white striping broiler breast meat. Spectrochim. Acta - Part A: Mol. Biomol. Spectrosc. 213 (128-126).

Kezhu, T., Yuhua, C., Weixian, S., Xiaoda, C., 2014. Detection of isoflavones content in soybean based on hyperspectral imaging technology. Sens. Transducers 169 (4), 55–60.

Kusumaningrum, D., Lee, H., Lohumi, S., Mo, C., Kim, M.S., Cho, B.K., 2018. Non-destructive technique for determining the viability of soybean (Glycine max) seeds using FT-NIR spectroscopy. J. Sci. Food Agric. 98 (5), 1734–1742.

Li, B., Cobo-Medina, M., Lecourt, J., Harrison, N.B., Harrison, R.J., Cross, J. v, 2018. Application of hyperspectral imaging for nondestructive measurement of plum quality attributes. Postharvest Biol. Technol. 141, 8–15.

Li, H., Liang, Y., Xu, Q., Cao, D., 2009. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. Anal. Chim. Acta 648 (1), 77–84.

Li, X., Zhang, L., Zhang, Y., Wang, D., Wang, X., Yu, L., Zhang, W., Li, P., 2020. Review of NIR spectroscopy methods for nondestructive quality analysis of oilseeds and edible oils. Trends Food Sci. Technol. 101, 172–181.

Li, Y., Sun, J., Wu, X., Chen, Q., Lu, B., Dai, C., 2019. Detection of viability of soybean seed based on fluorescence hyperspectra and CARS-SVM-AdaBoost model. J. Food Process. Preserv. 43 (12), 14238.

Liu, C., Huang, W., Yang, G., Wang, Q., Li, J., Chen, L., 2020. Determination of starch content in single kernel using near-infrared hyperspectral images from two sides of corn seeds. Infrared Phys. Technol. 110, 103462.

Liu, D., Sun, D.W., Zeng, X.A., 2014. Recent advances in wavelength selection techniques for hyperspectral image processing in the food industry. Food Bioprocess Technol. 7 (2), 307–323.

Mireei, S.A., Bagheri, R., Sadeghi, M., Shahraki, A., 2016. Developing an electronic portable device based on dielectric power spectroscopy for non-destructive prediction of date moisture content. Sens. Actuators, A: Phys. 247, 289–297.

Mo, C., Kim, G., Kim, M.S., Lim, J., Lee, S.H., Lee, H.S., Cho, B.K., 2017a. Discrimination methods for biological contaminants in fresh-cut lettuce based on VNIR and NIR hyperspectral imaging. Infrared Phys. Technol. 85, 1–12.

Mo, C., Kim, M.S., Kim, G., Lim, J., Delwiche, S.R., Chao, K., Lee, H., Cho, B.K., 2017b. Spatial assessment of soluble solid contents on apple slices using hyperspectral imaging. Biosyst. Eng. 159, 10–21.

Ouyang, Q., Wang, L., Park, B., Kang, R., Chen, Q., 2021. Simultaneous quantification of chemical constituents in matcha with visible-near infrared hyperspectral imaging technology. Food Chem. 350, 129141.

Rabanera, J.D., Guzman, J.D., Yaptenco, K.F., 2021. Rapid and Non-destructive measurement of moisture content of peanut (Arachis hypogaea L.) kernel using a near-infrared hyperspectral imaging technique. J. Food Meas. Charact. 15 (4), 3069–3078.

Song, X., Huang, Y., Yan, H., Xiong, Y., Min, S., 2016. A novel algorithm for spectral interval combination optimization. Anal. Chim. Acta 948, 19–29.

Sun, J., Shi, X., Zhang, H., Xia, L., Guo, Y., Sun, X., 2019. Detection of moisture content in peanut kernels using hyperspectral imaging technology coupled with chemometrics. J. Food Process Eng. 42 (7), 13263.

Sun, Y., Yuan, M., Liu, X., Su, M., Wang, L., Zeng, Y., Zang, H., Nie, L., 2021. A sample selection method specific to unknown test samples for calibration and validation sets based on spectra similarity. Spectrochim. Acta - Part A: Mol. Biomol. Spectrosc. 258, 119870.

Sun, Z.B., Wang, T.Z., Liu, X.Y., Zou, X.B., Liang, L.M., Li, J.K., Niu, Z., Gao, Y.L., 2020. Detection of prepared steaks freshness using hyperspectral technology combined with wavelengths selection methods combination strategy. Guang Pu Xue Yu Guang Pu Fen. Xi/Spectrosc. Spectr. Anal. 40 (10), 3224–3229.

Wang, Q., Wu, G., Pian, F., Shan, P., Li, Z., Ma, Z., 2021a. Simultaneous detection of glucose, triglycerides, and total cholesterol in whole blood by Fourier-Transform Raman spectroscopy. Spectrochim. Acta - Part A: Mol. Biomol. Spectrosc. 260, 119906.

Wang, Z., Fan, S., Wu, J., Zhang, C., Xu, F., Yang, X., Li, J., 2021b. Application of long-wave near infrared hyperspectral imaging for determination of moisture content of single maize seed. Spectrochim. Acta - Part A: Mol. Biomol. Spectrosc. 254, 119666.

Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemom. Intell. Lab. Syst. 58 (2), 109–130.

Wu, D., Sun, D.W., 2013. Advanced applications of hyperspectral imaging technology for food quality and safety analysis and assessment: a review - part I: Fundamentals. Innov. Food Sci. Emerg. Technol. 19, 1–14.

Xu, M., Sun, J., Yao, K., Cai, Q., Shen, J., Tian, Y., Zhou, X., 2022. Developing deep learning based regression approaches for prediction of firmness and pH in Kyoho grape using Vis/NIR hyperspectral imaging. Infrared Phys. Technol. 120, 104003.

Xu, Y., Zhang, H., Zhang, C., Wu, P., Li, J., Xia, Y., Fan, S., 2019. Rapid prediction and visualization of moisture content in single cucumber (Cucumis sativus L.) seed using hyperspectral imaging technology. Infrared Phys. Technol. 102, 103034.

Yang, D., He, D., Lu, A., Ren, D., Wang, J., 2017. Combination of spectral and textural information of hyperspectral imaging for the prediction of the moisture content and storage time of cooked beef. Infrared Phys. Technol. 83, 206–216.

Yu, H.D., Yun, Y.H., Zhang, W., Chen, H., Liu, D., Zhong, Q., Chen, W., Chen, W., 2020. Three-step hybrid strategy towards efficiently selecting variables in multivariate calibration of near-infrared spectra. Spectrochim. Acta - Part A: Mol. Biomol. Spectrosc. 224, 117376.

Yun, Y.H., Wang, W.T., Deng, B.C., Lai, G.B., Liu, X. bo, Ren, D.B., Liang, Y.Z., Fan, W., Xu, Q.S., 2015. Using variable combination population analysis for variable selection in multivariate calibration. Anal. Chim. Acta 862, 14–23.

Yun, Y.H., Li, H.D., Deng, B.C., Cao, D.S., 2019. An overview of variable selection methods in multivariate analysis of near-infrared spectra. Trends Anal. Chem. 113, 102–115.

Zhu, S., Chao, M., Zhang, J., Xu, X., Song, P., Zhang, J., Huang, Z., 2019. Identification of soybean seed varieties based on hyperspectral imaging technology. Sensors 19 (23), 5225.

Ziegler, V., Paraginski, R.T., Ferreira, C.D., 2021. Grain storage systems and effects of moisture, temperature and time on grain quality – a review. J. Stored Prod. Res. 91, 101770.