NONDESTRUCTIVE RAPID IDENTIFICATION OF SOYBEAN VARIETIES USING HYPERSPECTRAL IMAGING TECHNOLOGY

L. Wang, L. Pang, L. Yan, J. Zhang*

School of Technology at Beijing Forestry University,
Beijing, China; e-mail: joyzhangjm@163.com

Hyperspectral imaging technology was used to classify four types of soybean varieties. The reflectance spectra of four varieties of soybeans were extracted from hyperspectral images covering wavelengths from 400 to 1000 nm. Firstly, exploratory principal component analysis and linear discriminant analysis (LDA) were carried out to infer the separability of soybean spectral data. Secondly, the spectral data were preprocessed using multiplicative scattering correction (MSC), Savitzky–Golay (SG) smoothing, and MSC and SG smoothing together. Finally, classification models based on LDA, support vector machine (SVM), and k nearest neighbor (KNN) were established based on the full wavelengths or feature wavelengths. MSC and SG smoothing joint preprocessing of the spectral data was applied to establish the SVM classification model based on the full wavelengths, which returned a classification accuracy of 95.19%. Random forest was used to select the feature wavelengths from the full wavelengths to establish the LDA classification model, and the classification accuracy reached 82.69%. The results showed that the hyperspectral imaging technique combined with SVM, KNN, and LDA algorithms can be used to classify different soybean varieties in a fast and nondestructive way.

Keywords: *soybean seeds, hyperspectral imaging, variety classification, machine learning.*

НЕРАЗРУШАЮЩАЯ ЭКСПРЕСС-ИДЕНТИФИКАЦИЯ СОРТОВ СОИ С ИСПОЛЬЗОВАНИЕМ ТЕХНОЛОГИИ ГИПЕРСПЕКТРАЛЬНОЙ ВИЗУАЛИЗАЦИИ

L. Wang, L. Pang, L. Yan, J. Zhang*

УДК 535.3:582.739

Технологическая школа Пекинского университета лесного хозяйства,
Пекин, Китай; e-mail: joyzhangjm@163.com

(Поступила 19 ноября 2020)

Для классификации сортов сои использована технология гиперспектральной визуализации. Получены спектры отражения четырех сортов сои из гиперспектральных изображений в диапазоне 400—1000 нм. Метод главных компонент и линейный дискриминантный анализ (LDA) позволили сделать вывод о разделимости спектральных данных сои. Спектральные данные предварительно обрабатывались с использованием мультипликативной коррекции рассеяния (MSC), сглаживания Савицкого—Голея (SG), а также одновременно MSC и SG. Модели классификации, основанные на LDA, методах опорных векторов (SVM) и k-ближайших соседей (KNN), созданы на основе полных или характерных длин волн. Совместная предварительная обработка спектральных данных MSC и SG применена для создания модели классификации SVM, основанной на полных длинах волн, которая показала точность классификации 95.19 %. Метод случайного леса использован для выбора признаков среди всех длин волн для создания модели классификации LDA с точностью 82.69 %. Показано, что метод гиперспектральной визуализации в сочетании с алгоритмами SVM, KNN и LDA может использоваться для быстрой и неразрушающей классификации различных сортов сои.

Ключевые слова: *семена сои, гиперспектральная съемка, классификация сортов, машинное обучение.*

Introduction. Soybeans are important, globally traded agricultural products that are frequently exchanged due to the high demand coming from a variety of countries [1]. Different types of soybean varieties contain different compositions of amino acids, organic acids, and sugars. Furthermore, there is a gap between seed germination ability and vitality, which also affects the nutritional values, germination rates, and yields of soybeans [2, 3]. Therefore, the ability to identify soybean varieties is of great significance.

At present, seed variety identification is mainly based on the morphological characteristics of seeds. However, for seeds with no obvious distinguishing characteristics, this manual detection method is time consuming and has a large degree of observer error. To address the identification error, genetic marker technology can be used to accurately identify different kinds of seeds [4–6]. However, this approach damages seeds during the treatment process and requires professional, highly trained, technicians. Furthermore, the process of identification using genetic markers is complex and inefficient, so it is difficult to realize the large-scale batch detection of seeds. Therefore, a method that can quickly and nondestructively detect soybean varieties is needed.

Hyperspectral imaging (HSI) is an emerging rapid nondestructive technology that has the ability to simultaneously acquire spectra and spatial information. It has been widely used for identifying varieties of seeds in other fields and has achieved good results. Guo et al. [7] proposed a model-updating algorithm for maize seed variety recognition based on hyperspectral imaging using a pre-labeling method. The average classification accuracies were improved by 8.9, 35.8, and 9.6%. Feng et al. [8] used hyperspectral imaging to detect genetically modified maize kernels and their non-genetically modified parents. Their results demonstrated that clear differences between genetically modified and non- genetically modified maize kernels can be easily visualized using their nondestructive determination method. To date, there have been many studies on the use of hyperspectral imaging technology to classify seed varieties [9–11]. Zhu et al. [12] used near-infrared (NIR) hyperspectral imaging to classify three soybean cultivars and established a convolutive neural network using their average spectra and pixel level spectra; the classification accuracy surpassed 90%. Furthermore, near infrared hyperspectral imaging (NIR-HSI) has been used to identify the vigor of rice seeds in which the full spectrum and selected feature wavelengths were used to obtain a reliable classification performance (94.38% accuracy) [13].

Most previous studies have established classification models based on large data sets and achieved accurate results, but relatively few studies have attempted to establish classification models of soybean varieties with small sample sizes. The main purpose of this study was to explore the feasibility of using hyperspectral imaging technology to classify different soybean varieties. The specific objectives were: (1) to explore the feasibility of using principal component analysis (PCA) and linear discriminant analysis (LDA) to visualize different soybean varieties; (2) to qualitatively evaluate the detection ability of soybean varieties; (3) to analyze and compare the influence of the Savitzky–Golay (SG) smoothing and the multiplicative scatter correction (MSC) on soybean spectral data; (4) to establish a soybean variety classification model based on full wavelengths and feature wavelengths; and (5) to compare soybean classification performance of three classification models using support vector machine (SVM), linear discriminant analysis (LDA) and k-nearest neighbor (KNN).

Materials and methods. *Sample preparation of soybean seeds.* Four soybean varieties were used in this experiment, including Wan Dou 28 (WD28), Zhong Huang 55 (ZH55), Zhong Huang 13 (ZH13), and Zhong Huang 41 (ZH41), all of which were purchased from the seed market in Beijing (China). To the naked eye there were no remarkable differences in appearances of these four soybeans. For each type of soybean, 204 intact soybeans were selected. Groups of 34 soybeans of the same variety were placed on sampling plates and hyperspectral imaging was performed on each group of soybeans separately.

Hyperspectral image acquisition and correction. The hyperspectral imaging system used in this experiment consisted of the hyperspectral imager, halogen lamp, mobile platform, and computer. The hyperspectral imager was the SOC710VP hyperspectral imager manufactured by Surface Optics Corporation (USA). Two 150 W halogen (OSRAM GCA) lamps were used as light sources. The conveying platform was driven by a stepper motor for accurate control. Transmission of motor control signals and spectral image information was conducted through USB serial communication. In this study, hyperspectral information of soybeans was obtained in the spectral range of 400–1000 nm. Table 1 shows the main performance parameters of the hyperspectral imager.

Because the hyperspectral image acquisition process is easily affected by the external environment, the acquisition process was conducted entirely in a dark box. To obtain a deformable and clear hyperspectral image, the distance between the sample and the camera lens, speed of the moving platform, and exposure

time were set to 18.0 cm, 18 mm/s, and 4ms, respectively. The obtained raw hyperspectral images, which image light intensity, needed to be corrected to the reflected hyperspectral image. The formula for image correction is as follows:

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

where I is the corrected image, I_r is the original image, I_w is the white reference image obtained using a white Teflon board with a high reflectivity (close to 100%), and I_d is obtained through a fully covered camera lens [12].

TABLE 1. Performance Parameters of SOC 710-vp Hyperspectral Imager

Spectral range, nm	400–1000	Lens type	C-Mount
Spectral resolution, nm	4.69	Weight, kg	2.95 (6.5 lbs)
Band	128	Size, cm	9.5×16.8×22 cm
Dynamic range, Bit	12/16	Power source	12-VDC/100-240 VAC (50-60Hz)
Pixels per line	696	Speed	30 rows/s, 23.2 s/cube

Spectral extraction and data preprocessing. After obtaining hyperspectral images, threshold segmentation, image filling, and denoising processes were used to eliminate any background influence; then connected regions were obtained to mark the centroid of each seed sample. The center of the circle was the centroid of the sample, and a circle with a radius of 10 pixels was selected as the region of interest (ROI). The spectral average of pixel points in the circle was calculated according to the equation

$$\bar{I} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} I_{ij}}{m}, \quad (2)$$

where m is the number of pixels in the circle region; n is the number of spectral bands of hyperspectral images, 128 in this project; I_{ij} is the spectral value of the ith pixel in the jth band; and $\bar{I}$ is the average spectral value of the circle region.

The pixel spectra within each soybean seed ROI were averaged to obtain a total of 816 spectral curves. During the process of collecting hyperspectral image information, random noise will affect the spectrum of the sample [14, 15]. The starting position of the spectral line showed high levels of noise. Therefore, the spectral information of the intermediate bands from 450–1000 nm was selected from the bands from 400–1000 nm for further analysis. At present, the most commonly used spectral information preprocessing methods include the standard normal variable (SNV) transformation, multiplicative scatter correction (MSC) [16], Savitzky–Golay (SG) smoothing [17], and wavelet transform (WT) [18]. In this study, the MSC and SG smoothing algorithms were used to preprocess the spectral data, and then analyzed and compared to identify the optimal performance of the model.

Results and discussions. *Spectral profiles.* The extracted spectral information of the four soybean varieties are presented in the spectral curves in Fig. 1a. Figure 1b showed the average of the spectra for each soybean variety. It can be seen from the spectral curves of the four varieties that they were basically the same, with similar peaks and valleys, but there were differences in reflectance.

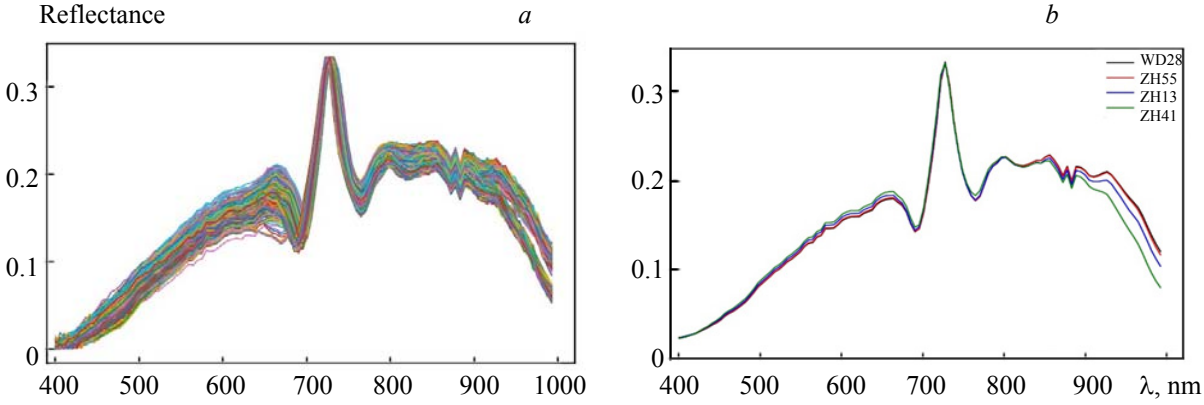

Fig. 1. Spectrum of four soybean varieties: a) spectra of all samples; b) average spectra of all samples.

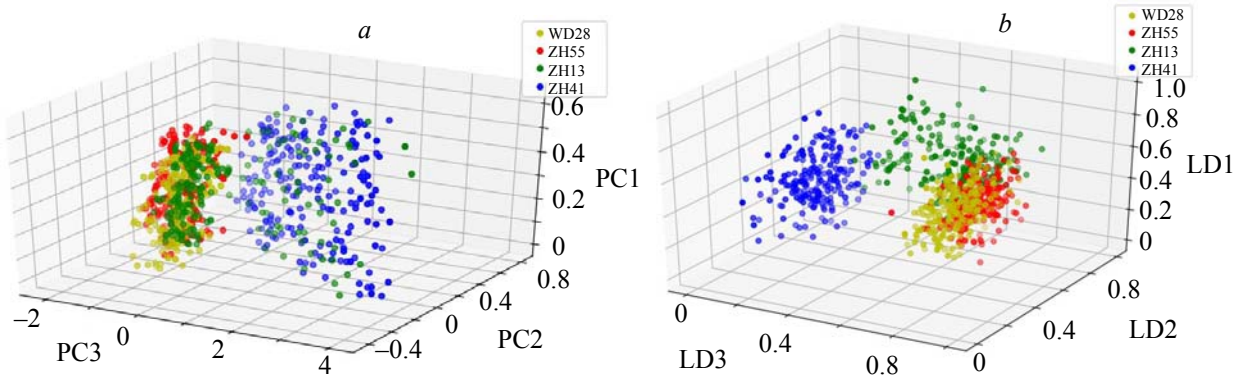

Fig. 2. Exploratory classification scatter plot of four different varieties of soybeans: a) three-dimensional scatter plot performed by principal component analysis (PCA); b) three-dimensional scatter plot performed by linear discrimination analysis (LDA).

Exploratory classification analysis. To explore the distinguishing characteristics of the soybean varieties, 150 soybean seeds were randomly selected from each soybean variety, and the key information from the full spectrum was extracted through principal component analysis. The results are shown in Fig. 2. The contributions of the first three principal components were 84.78, 6.86, and 5.24%, respectively, which together explained 96.88% of the information contained in the original spectrum. Therefore, the first three principal components were selected for analysis. It can be seen in Fig. 2a that there were intersecting areas between the WD28 and ZH55 soybean varieties and the ZH13 and ZH41 soybean varieties, and there were no obvious boundaries between classes, which was generally consistent with the spectral curve analysis. The above indicated that the principal component analysis failed to provide a reliable basis for soybean classification.

To better explore the separability of the four soybean varieties, this study introduced the supervised dimension reduction algorithm LDA. The LDA algorithm makes intra-class distances as small as possible and makes the distances between classes as large as possible, so it is easier to explore the separability of soybean varieties [19, 20]. The first three principal components, with the largest contributions, were also used to make a three-dimensional scatter plot. As shown in Fig. 2b, the results showed that LDA was able to make the boundaries between soybean varieties more obvious, especially between WD13 and WD41. Compared

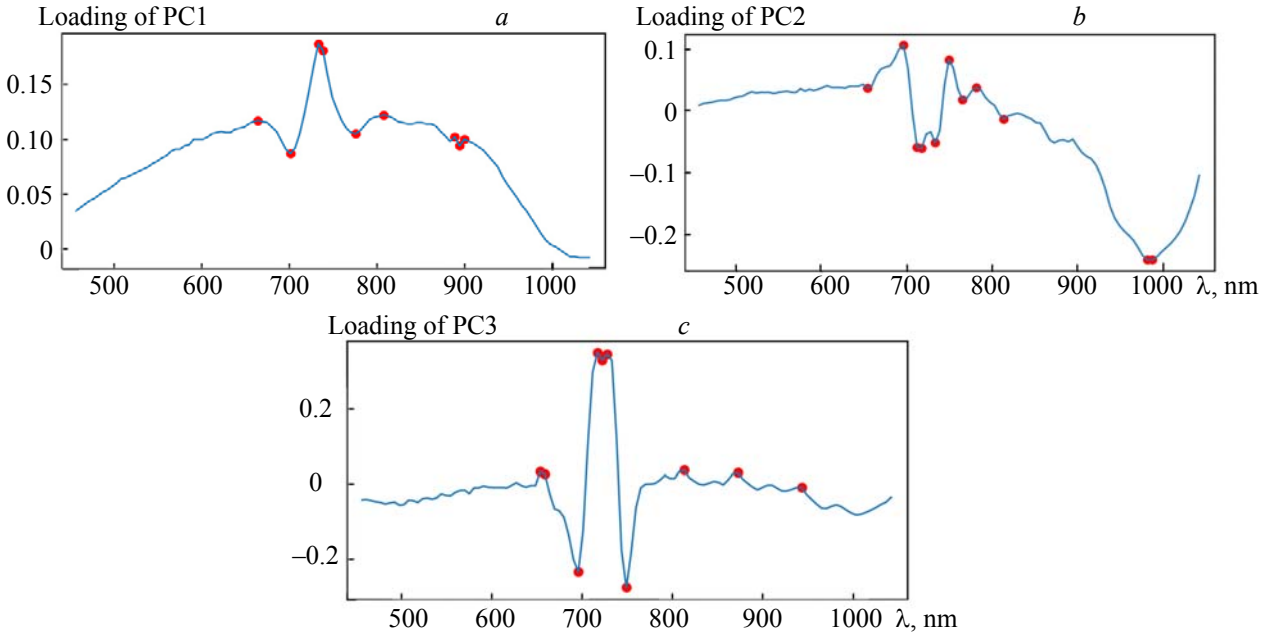

Fig. 3. Principal component analysis extracted feature wavelengths from the spectrum: a) weight coefficient distribution of PC1; b) weight coefficient distribution of PC2; c) weight coefficient distribution of PC3.

with PCA, LDA had better classification ability, which established a solid foundation for soybean variety classification models.

Classification results and analysis of discriminant models based on the full wavelengths. Taking the overall accuracy of soybean classification and recognition as the evaluation index for the model, soybean variety discrimination models using LDA, SVM, and KNN and based on full wavelengths were established. Each soybean variety data set was divided into training sets and test sets at a ratio of 3:1. The training set and test set of the four soybean varieties were used as the training set and test set for the discriminant models. After repeating the cross-validation five times, the average accuracy of the five outputs was used as the final accuracy output. This study also compared how the three data preprocessing methods (SG, MSC, and MSC+SG) affected the three discriminant models. As shown in Table 2, for the original data set, LDA was best at classification with an accuracy of 88.21% for the test set, followed by SVM and KNN models, in that order. After SG preprocessing, the classification ability of the SVM model was better than both the LDA and KNN models, with test set accuracy of 86.92%. After MSC preprocessing, the output of the SVM and KNN models both improved, but the output accuracy of the LDA model decreased. After MSC+SG preprocessing, the accuracy of the SVM model for the training set and test set reached 100 and 95.19%, respectively, but the classification performance by the other models decreased, although in the test set the LDA model was slightly better than KNN model.

TABLE 2. Classification Performance of Models Based on Full Wavelengths and Different Preprocessing Methods

Classification model	Spectral data set	Training set, %	Test set, %
SVM	Raw	92.38	84.28
	Raw+SG	93.91	86.92
	Raw+MSC	93.43	88.21
	Raw+SG+MSC	100.00	95.19
KNN	Raw	85.10	71.63
	Raw+SG	84.94	71.63
	Raw+MSC	91.99	80.29
	Raw+SG+MSC	90.54	81.73
LDA	Raw	92.47	88.21
	Raw+SG	90.31	86.29
	Raw+MSC	87.34	82.92
	Raw+SG+MSC	86.38	82.48

Superposition pretreatment greatly improved the performance of the SVM classification model, and the accuracy of the KNN classification model was also improved. However, the classifications by the LDA classification model were slightly less accurate after preprocessing. This comparison of three types of discriminant models found that data preprocessing did not improve the output accuracy of all model types in the same way and that spectral data sets produced different results from different models depending on the different preprocessing methods.

Extraction of feature wavelengths. To reduce overfitting, reduce the number of irrelevant features, improve the generalization ability of the models, make the model easier to understand, and accelerate the training speed of the model, PCA Loading, random forest (RF), and the genetic algorithm (GA) were used for feature selection based on 117 bands of the full spectrum.

The contributions of the first three principal components explained more than 95% of the information in the original spectrum. Therefore, the loading of the first three PCs was used to identify important wavelengths. It can be seen that the loading of the PC1 curve was consistent with the original spectral curve (Fig. 3). The wavelengths at the peak and trough of the curve were the preferred choice for feature selection [11]. The PCA loadings of the first three PCs were used, and 24 essential feature wavelengths were selected.

RF used bootstrap resampling technology to randomly extract some samples from the original training set to generate a new training set for training decision trees. That process was repeated many times to generate multiple decision trees, and the classification results were decided by voting according to each decision tree's results. While training the decision tree, the contribution of each feature can be calculated, and important features can be selected by establishing the threshold of contribution rate. As shown in Fig. 4, 15 im-

portant feature wavelengths were selected from the full wavelengths using random forest feature selection method.

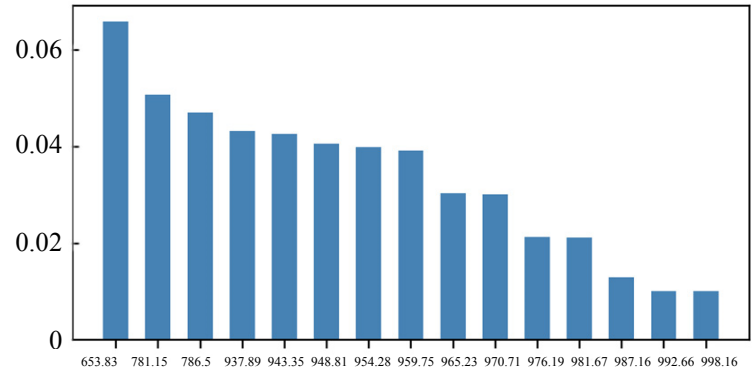

Fig. 4. Random forest algorithm extracts feature wavelengths from the full wavelengths.

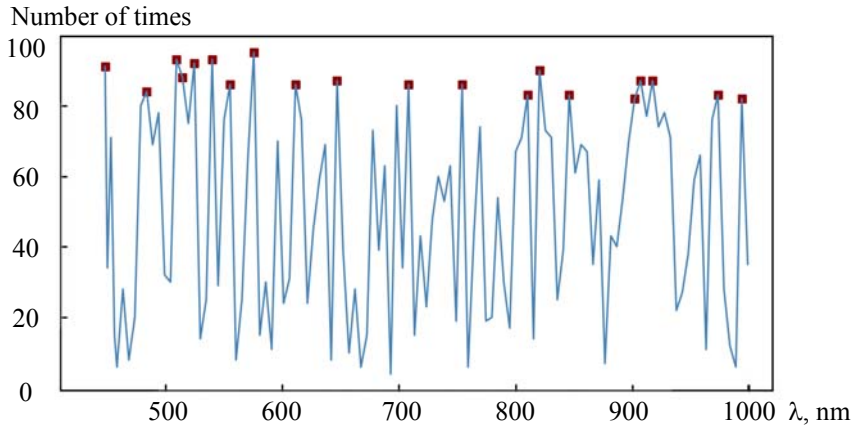

Fig. 5. In 100 iterations of the genetic algorithm, the feature wavelengths were those selected more than 80 times.

TABLE 3. Feature Wavelengths Selected by the Three Feature Selection Methods

Method	Number	Feature wavelengths, nm
PCA Loading	24	653.83, 659.09, 664.36, 696.03, 701.32, 711.92, 717.22, 722.53, 727.84, 733.15, 738.47, 749.12, 765.12, 775.81, 781.15, 807.95, 813.32, 872.65, 888.91, 894.33, 899.77, 943.35, 981.68, 987.17
RF	15	653.83, 781.15, 786.50, 937.89, 943.35, 948.81, 954.28, 959.75, 965.23, 970.71, 976.19, 981.67, 987.16, 992.66, 998.16
GA	20	451.54, 497.71, 523.50, 528.66, 539.01, 554.56, 570.14, 590.97, 627.57, 664.35, 727.84, 775.80 834.84, 845.62, 872.64, 932.42, 937.88, 948.81, 981.68, 987.17

The selection of feature wavelengths was carried out using GA [18]. The 117 bands were mapped to the genes of the chromosomes using the binary coding method. First, the population was initialized. A population contains multiple chromosomes, each of which contains 117 genes. The accuracy of the decision tree was used as the fitness to calculate the fitness value of each individual. The roulette selection method was used for individual selection, and individuals with high fitness value were more likely to pass on genes to the next generation. Then the chromosomes were crossed and mutated. Each iteration can create the optimal individual and the genes contained in the individual, and the selected feature wavelengths can be determined

by the genes. After analyzing the genes on the chromosomes after 100 iterations, it was found that 20 genes were selected more than 80 times, so the feature wavelengths represented by the genes were selected, as shown in Fig. 5. The feature wavelengths were selected according to three feature selection methods, as shown in Table 3.

Classification results and analysis of discriminant models based on feature wavelengths. By extracting feature wavelengths from 117 wavelength bands without preprocessing, the SVM, KNN, and LDA models were established to classify different soybean varieties. The overall classification accuracy was used as the evaluation index to measure the performance of different feature extraction methods in classifying soybeans. The results, presented in Table 4, show that RF was the best of the three feature extraction methods. By comparing the test sets, when PCA loading and GA were used for the extraction of feature wavelengths, the LDA model was most accurate, followed by SVM and KNN, in that order. When RF was to extract feature wavelengths, the LDA model was most accurate again, the classification accuracy of the KNN model was higher than that of the SVM model; the reason may be that the number of support vectors was reduced to 15, which hindered the SVM model's ability to deal with multi-classification problems. The KNN model had a certain classification advantage in dealing with the overlapping sample set, which depended on the limited adjacent samples. In comparing feature selection methods, we see that, although the number of features selected by GA was greater than those selected by RF, the classification accuracy of the genetic algorithm was slightly lower than that of the random forest algorithm. The reason may be that the wavelengths selected more than 80 times in 100 iterations were used as feature wavelengths, which may have caused information loss. Without preprocessing, classification models based on feature wavelengths have lower in classification accuracies than classification models based on full spectra, but they are capable of more rapid and efficient identification of soybean varieties.

TABLE 4. Classification Performance of Models Based on Feature Wavelengths

Feature extracting method	SVM		KNN		LDA	
	Cal, %	Pre, %	Cal, %	Pre, %	Cal, %	Pre, %
PCA Loading	83.17	76.44	74.62	66.83	80.77	77.88
RF	85.82	80.10	85.90	81.25	83.89	82.69
GA	80.60	75.48	75.81	70.38	79.17	76.04

Conclusions. This study was based on hyperspectral imaging technology and used machine learning algorithms to classify four soybean varieties. PCA and LDA were used for exploratory analysis. LDA was shown to have a good recognition effect and was able to differentiate between WD28, ZH55, ZH13, and ZH41. The models were then assessed using data that were preprocessed with MSC, SG, and MSC+SG, and compared to the original spectral data. The results showed that after MSC+SG preprocessing, the performance of the SVM model was significantly improved with an accuracy of over 95%. However, not all the preprocessing methods had positive effects on the classification performance of all models. The feature wavelengths of the original spectra were extracted by the PCA Loading, RF, and GA algorithms. Then a classification model was built by selecting important features, and the influence of using the full wavelengths or the feature wavelengths on the performance of the three classification models was examined. The results showed that the classification model based on feature wavelengths needed to be improved. However, according to the classification results based on the RF feature selection algorithm, extracting some features for rapid detection of soybean varieties has great potential in detection speed and economic cost. In addition, it is necessary to further study the internal relationship between soybean spectra and soybean composition, characterize more soybean varieties, and establish and expand the experimental sample database.

Acknowledgments. This work was supported by National Natural Science Foundation of China (Grant No. 31770769), the National Key Research and Development Program of China (No. 2017YFC0504403), and the Fundamental Research Funds for the Central Universities (No. 2015ZCQ-GX-03).

REFERENCES

1. Baek Insuck, Kusumaningrum Dewi, Kandpal Lalit Mohan, Lohumi Santosh, Mo Changyeun, Kim S. Moon, Cho Byoung-Kwan, *Sensors*, **19**, No. 2 (2019), https://doi.org/10.3390/s19020271.
2. M. T. McCarville, C. C. Marett, M. P. Mullaney, G. D. Gebhart, G. L. Tylka, *Plant Health Prog*., **18**, 146–155 (2017), https://doi.org/10.1094/PHP-RS-16-0062.
3. K. M. Maria John, S. Natarajan, D. L. Luthria, *Food Chem*., **211**, 347–355 (2016), https://doi.org/10.1016/j.foodchem.2016.05.055.
4. S. Ye, Y. Wang, D. Huang, J. Li, Y. Gong, L. Xu, L. Liu, *Sci. Horticult.*, **155**, 92–96 (2013), https://doi.org/10.1016/j.scienta.2013.03.016.
5. S. Vanisri, R. Durga, G. Swathi, M. Jamal, M. Sreedhar, N. R. Kumar, E. Ramprasad, Y. Raviteja, *Agric. Res*. (**2018**), https://doi.org/10.1007/s40003-018-0324-8.
6. K. Moorthy, P. Babu, M. Sreedhar, et al., *J. Seed Sci. Technol*., **39**, No. 2, 282–292 (2011), https://doi.org/10.1007/s10681-007-9630-0.
7. D. Guo, Q. Zhu, M. Huang, et al., *Comput. Electron. Agric*., **142**, 1–8 (2017), https://doi.org/10.1016/j.compag.2017.08.015.
8. X. Feng, et al., *Sensors* (Basel), **17**, No. 8, 1894 (2017), https://doi.org/10.3390/s17081894.
9. M. Huang, J. Tang, B. Yang, Q. Zhu, *Comput. Electron. Agric*., **122**, 139–145 (2016), https://doi.org/10.1016/j.compag.2016.01.029.
10. S. Jia, et al., *J. Cereal Sci*., **63**, 21–26 (2015), https://doi.org/10.1016/j.jcs.2014.07.003.
11. Y. Zhao, et al., *Molecules*, **23**, No. 6 (2018), https://doi.org/10.3390/molecules23061352.
12. Susu Zhu, Lei Zhou, Chu Zhan, Yidan Ba, Baohua Wu, Hangjian Chu, Yue Y, Yong He, Lei Feng, *Sensors*, **19**, 4065 (2019), https://doi.org/10.3390/s19194065.
13. X. He, X. Feng, D. Sun, et al., *Molecules*, **24**, No. 12, 2227 (2019), https://doi.org/10.3390/molecules24122227.
14. N. Wu, Y. Zhang, R. Na, C. Mi, S. Zhu, Y. He, C. Zhang, *RSC Adv*., 12635–12644 (2019), https://doi.org/10.1039/c8ra10335f.
15. L. Ravikanth, C. B. Singh, D. S. Jayas, N. D. White, *Biosyst. Eng*., **135**, 73–86 (2015), https://doi.org/10.1016/j.biosystemseng.2015.04.007.
16. T. Isaksson, T. Næs, *Appl. Spectrosc*., **42**, No. 7, 1273–1284 (1988), http://doi.org/10.1366/0003702884429869.
17. A. Savitzky, M. J. E. Golay, *Anal. Chem*., **36**, 1627–1639 (1964), http://dx.doi.org/10.1021/ac60214a047.
18. J. Yang, V. Honavar, *IEEE Intell. Systems Their Appl*., **13**, No. 2, 44–49 (1998), https://doi.org/10.1109/5254.671091.
19. Y. Zhao, S. Zhu, C. Zhang, X. Feng, L. Feng, Y. He, *RSC Adv*., 1337–1345 (2018), https://doi.org/10.1039/C7RA05954J.
20. X. Feng, C. Peng, Y. Chen, X. Liu, X. Feng, Y. He, *Sci. Rep*., 15934 (2017), https://doi.org/10.1038/s41598-017-16254-z.