APPLICATION OF HYPERSPECTRAL IMAGING TO IDENTIFY PINE SEED VARIETIES[**]

Jianing Ma, Lei Pang, Yuemeng Guo, Jinghua Wang, Jingjing Ma, Fang He, Lei Yan[*]

School of Technology, Beijing Forestry University, Beijing, China;
e-mail: mark_yanlei@bjfu.edu.cn

Seed variety purity is the main indicator of seed quality, which affects crop yield and product quality. In the present study, a new method for the identification of pine nut varieties based on hyperspectral imaging and convolutional neural networks LeNet-5 was established so as to avoid the hybridization of different varieties of pine nuts, improve the identification efficiency and reduce the cost of identification. Images of 128 wavelengths in the 370–1042 nm range were acquired by hyperspectral imaging. The spectrum and image of each seed were obtained by means of black-and-white correction and region segmentation of the original image. Twenty characteristic wavelengths were extracted from the first three principal components (PCs) of principal component analysis (PCA). A support vector machine (SVM) spectral recognition model based on full wavelengths and characteristic wavelengths was established. For different species of pine seeds, the classification accuracies of the prediction set in the aforementioned datasets were 97.7 and 93.1%, respectively. The seed images of 20 characteristic wavelengths were input into LeNet-5 to improve the network structure and the number of convolution channels. The improved LeNet-5 performed better with over 99% accuracy. Such results show that the convolutional neural network is of considerable significance for fast and nondestructive identification of pine seed varieties.

Keywords: *pine seed, variety identification, hyperspectral imaging, support vector machine model, principal component analysis, convolutional neural networks.*

ИДЕНТИФИКАЦИЯ СОРТОВ СЕМЯН СОСНЫ С ПОМОЩЬЮ ГИПЕРСПЕКТРАЛЬНОЙ ВИЗУАЛИЗАЦИИ

Jianing Ma, L. Pang, Y. Guo, J. Wang, Jingjing Ma, F. He, L. Yan[*]

УДК 535.317.1:582.475.4

Пекинский университет лесного хозяйства, Пекин, Китай;
e-mail: mark_yanlei@bjfu.edu.cn

(Поступила 23 июня 2022)

Разработан метод идентификации сортов кедровых орехов на основе гиперспектральной визуализации и сверточной нейронной сети LeNet-5 с целью избежать гибридизации разных сортов кедровых орехов, повысить эффективность идентификации и снизить ее стоимость. Изображения 128 длин волн в диапазоне 370–1042 нм получены с помощью гиперспектральной визуализации. Спектр и изображение каждого семени получены с помощью черно-белой коррекции и сегментации области исходного изображения. Двадцать характеристических длин волн получены из первых трех главных компонент при анализе методом главных компонент. Создана модель спектрального распознавания на основе машины опорных векторов для наборов всех длин волн и характеристических длин волн. Для разных видов семян сосны точность классификации в указанных наборах данных составила 97.7 и 93.1 %. Исходные изображения 20 характерных длин волн введены в LeNet-5 для улучшения структуры сети и каналов свертки. Точность улучшенного LeNet-5 <99 %.

Ключевые слова: *семена сосны, идентификация сортов, гиперспектральная визуализация, модель машины опорных векторов, метод главных компонент, сверточная нейронная сеть.*

[**] Full text is published in JAS V. 90, No. 4 (http://springer.com/journal/10812) and in electronic version of ZhPS V. 90, No. 4 (http://www.elibrary.ru/title_about.asp?id=7318; sales@elibrary.ru).

Introduction. Among the existing gymnosperms, Pinaceae is the group with the most species, the widest distribution, and the largest forest area and wood stock. There is a wealth of information on phylogeny involving anatomy [1], wood, cells [2, 3], phytochemistry [4] and biochemistry [5], with different species of pines having different properties and uses [6]. Surveys have shown that *P. thunbergii Parlatore* has wind resistance and is highly adaptable to harsh weather, being mainly used for green viewing. The branches of *P. massoniana Lamb.* are rich in turpentine and can be utilized as fuelwood. The ancient Egyptians added *cedar* oil to cosmetics for beauty purposes and as an insect repellent. *P. tabuliformis Carriere* is an environmentally friendly tree species with a strong adsorption capacity for PM2.5 (airborne particulate matter (PM) less than 2.5 μm in diameter). Swamp pine has strong adaptability and has been extensively adopted in construction, fiberboard, paper and other industries. Widely used in vitality detection [7–9], disease detection [10], and internal component detection [11], hyperspectral imaging technology is a fusion of spectroscopy technology and image technology, which can simultaneously acquire the spatial information of the measured object and the spectral information of each pixel in the image. Compared with traditional technology, hyperspectral imaging technology will promote rapid, accurate and convenient nondestructive seed identification methods. With the inevitable trend of agricultural automation and intensification, such technology will be of considerable significance in seed purity identification.

At present, there are numerous existing studies on the detection and identification of seed varieties [12, 13]. In terms of accurately evaluating seed quality grades, the purity identification of seed varieties is the main basis for such evaluation [14, 15]. The traditional methods of seed recognition primarily include grain shape identification, seedling identification, field planting identification, electrophoretic analysis of biochemical indicators, molecular marker detection, and others [16]. Although such testing methods are accurate and intuitive, a variety of deficiencies remain, such as damage to seeds, long identification time and strong dependence on personnel [17, 18]. The current seed industry development requires a fast and efficient method for seed variety testing, which is of considerable significance for seed variety identification and classification. Convolutional neural networks (CNNs) are one of the most popular methods for image classification [19, 20]. In a large number of studies, convolutional neural networks have been adopted for detection and prediction of hyperspectral images. With regard to crops such as corn [18, 21], soybeans [22–24], and others, CCNs also have a wide range of uses.

The purpose of the present study was to investigate the feasibility of hyperspectral imaging for identifying different pine seed species. The details are as follows: qualitative analysis of different pine seed varieties by PCA; establishing the SVM recognition model based on full wavelengths and characteristic wavelengths, and comparing the advantages and disadvantages; comparing the results from the improved Lenet-6 network after inputting the image data obtained from the 20 characteristic wavelengths selected based on PCA.

Materials and methods. Five pine seeds with similar appearance and different primary uses were selected, namely *P. thunbergii Parlatore*, *P. massoniana Lamb.*, *Cedrus deodara, P. elliottii*, and *P. tabuliformis Carriere*. After purchasing samples in the market, 600 seeds of uniform size were randomly selected from each variety as data collection samples, totaling 3000 pine seeds.

The hyperspectral data acquisition systems used in the present study included hyperspectral imagers, industrial cameras, light sources, mobile platforms, and computers. The hyperspectral imager SOC 710-VP manufactured by Polytec (Germany) was used to obtain reflected light from the seeds. Two halogen lamps were fixed on both sides of the acquisition platform to provide stable, continuous illumination. The specific parameters of the equipment are described in Table 1. Sixty seeds were placed at equal intervals on the black cardboard each time. After shooting was completed, data were stored in the computer, and the aforementioned process was regarded as the data collection process. Seed collection for each variety required 10 replicates.

Owing to the uneven distribution of light intensity of the light source, a dark current exists in the sensor, the light intensity is weak, and the obtained hyperspectral image has large noise. As such, black and white correction is required:

$$I = \frac{I_0 - B}{W - B}, \tag{1}$$

where I is the corrected spectral data; I_0 is the spectral data before correction; B is the black reference value which is obtained by completely shading the lens; W is the white reference value (the maximum reflectance), which is obtained from a standard tetrafluoroethylene whiteboard with a reflectivity of close to 100%.

TABLE 1. Technical Specifications

Hyperspectral imaging spectrometer model	Spectral range	Spectral channels	Spectral resolution	Dynamic range
SOC710-VP	400–1000 nm	128	4.69 nm	12/16-Bit
Schneider Lens model	focal length	F/# range	Max. sensor size	rec. working distance range
Xenoplan 2.8/50	50 mm	F/2.8 F/32	24mm	131 mm ... ∞
Halogen lamp model	power	maximum voltage	wavelength range	wavelength accuracy
SLS CL-150	150 W	250 V	350–2000 nm	±0.1 nm

The steps of image region segmentation in the study were as follows:

1. Obtain the image threshold by means of the Otsu maximum inter-class variance method [25].
2. Enter a threshold to convert the grayscale image to a binary image.
3. Use morphological operations to denoise to obtain connected regions.
4. Draw the binary image after the opening operation.
5. Extract the region of interest (ROI) of a single seed after drawing the binary image.

Standard normal variable transformation (SNV) is a common spectral preprocessing method that is widely used by researchers [26—28] and is primarily adopted to eliminate the effects of solid particle size, surface scattering, and optical path variation on the NIR diffuse reflectance spectrum [29]. The spectral SNV is calculated as follows:

$$x_{\mathrm{SNV}} = (x - \bar{x}) \Big/ \sqrt{\sum_{k=1}^{m} \left(x_k - \bar{x}\right)^2 \Big/ (m-1)} \ . \tag{2}$$

Principal component analysis (PCA) is one of the most widely used data dimensionality reduction algorithms, and allows for the original features with strong correlations to be mapped to a new set of features [30]. The mapped feature variables are linear combinations of the original matrices, and the variables are linearly uncorrelated. The amount of spectral data obtained by the hyperspectral imaging system is considerably large, and contains a large amount of redundant information, which indirectly affects the accuracy of model identification. Therefore, PCA needs to be used to extract characteristic wavelengths, thereby reducing the amount of input data and improving the performance of the model [31].

Support vector machine (SVM) [32] is a generalized linear classifier that performs binary classification of data in a supervised learning fashion that can simply represent complex nonlinear patterns [33]. The basic concept involves solving the separating hyperplane that correctly divides the training dataset and has the largest geometric separation. Through such division of the hyperplane, the generalization ability to unseen samples is not only the strongest, but will also have the least impact on the sample when locally perturbed, thus the most robust classification results will be produced. The key to SVM is the kernel function. The low-dimensional space vector set is usually difficult to divide, and the function of the high-dimensional space can be obtained by selecting the appropriate kernel function [34]. Gaussian radial basis kernel function [35] was the kernel function used in the present study. SVM is described as a quadratic optimization problem [36]:

$$\min_{w,b,\xi} \left(\frac{1}{2} \|w\|^2 + c \sum_{i=1}^{N} \xi_i \right) \ , \tag{3}$$

where w is the optimal solution; b is the bias parameter; $c > 0$ is the penalty parameter of the error term; and ξ_i is the slack variable that is related to prediction errors in SVM.

LeNet is one of the most representative experimental systems in early convolutional neural networks. Lecun *et al.* [37] first recognized handwritten characters according to LeNet-5, and the artificial neurons respond to a surrounding area of a portion of the coverage, performing well for large image processing [38]. LeNet-5 primarily has two convolutional layers, two pooling layers, and three fully connected layers. In the present study, network improvements were made on LeNet-5.

In the present study, the quality of the model was evaluated in terms of the number of parameters, accuracy and loss function. Accuracy is a commonly used indicator for evaluating the quality of classification models, representing the ratio of the number of samples correctly classified by the classification model to the total number of samples for a given test set. Accuracy presents the model’s overall predictions and can be calculated as follows:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}, \qquad (4)$$

where TP is true positive; TN is true negative; FP is false positive; and FN is false negative.

Results and discussion. *Hyperspectral analysis of pine seeds.* In the obtained hyperspectral images, each seed needs to be segmented to obtain a single seed, which is convenient for subsequent calculation of the spectral reflectance of a single seed. The spectrum curve of each seed was calculated from ROI, and the average spectra of the five types of seeds were drawn (Fig. 1a). According to Fig. 1b, the spectral curves of different kinds of pine seeds exhibited similar trends and had peaks and valleys at the same position. Such findings show that the characteristics were similar. Compared with other categories, *P. massoniana Lamb.* had the highest reflectivity. However, the spectra of different varieties of seeds differed only in reflectivity. Consequently, there were difficulties in subjectively discriminating seed types based on spectral curves.

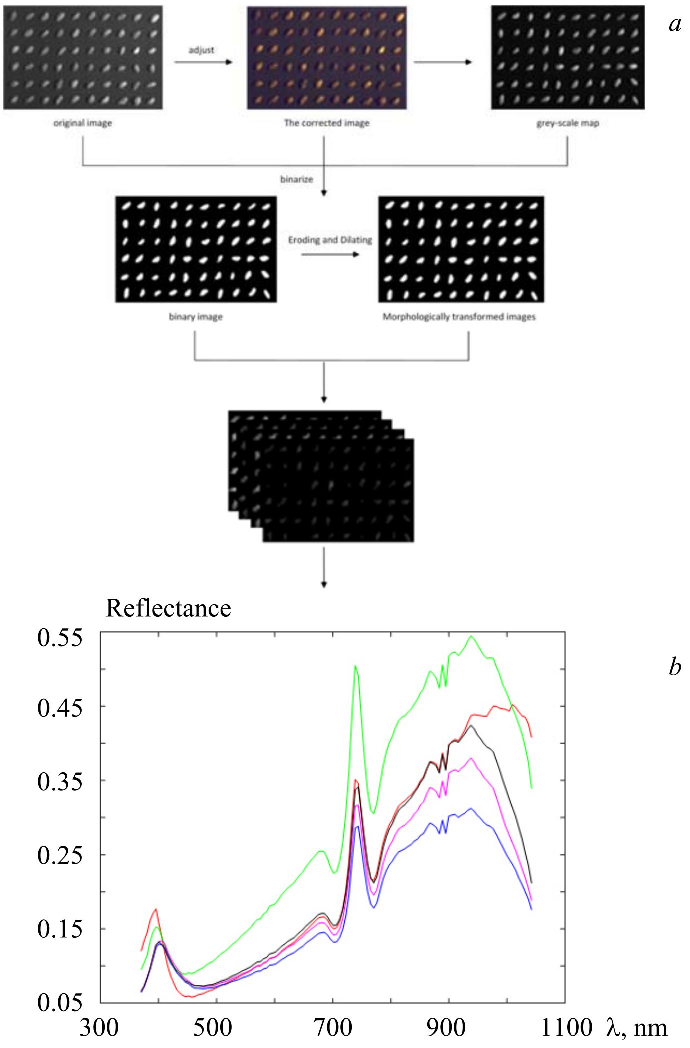

Fig. 1. Hyperspectral analysis of pine seeds: (a) Image segmentation and spectral extraction process; (b) Average spectrum of different species of pine seeds (the green curve represents *P. massoniana Lamb.*; the red curve represents *P. thunbergii Parlatore*; the black curve represents *P. tabuliformis Carriere*; the pink curve represents *P. elliottii*; the blue curve represents *C. deodara*).

Principal component analysis. To investigate the fractional spread of different principal components, PCA was performed for five pine seed varieties, with the transformed results being shown in Fig. 2a. An observation can be made that the cumulative contribution rate of the three principal components reached 98.8% (84.14, 11.64, and 2.64%, respectively), and such results can represent most of the hyperspectral information. Since each principal component was a linear combination of average spectral reflectance in 128 wavelengths, to determine the main wavelength of information expression in each principal component, the loadings of the three principal components were plotted in Fig. 2b. According to the loadings map of each principal component, the position where each peak was located was selected as the characteristic wavelength. After eliminating repeating wavelengths, a total of 20 wavelengths were identified as the optimal wavelengths (Table 2).

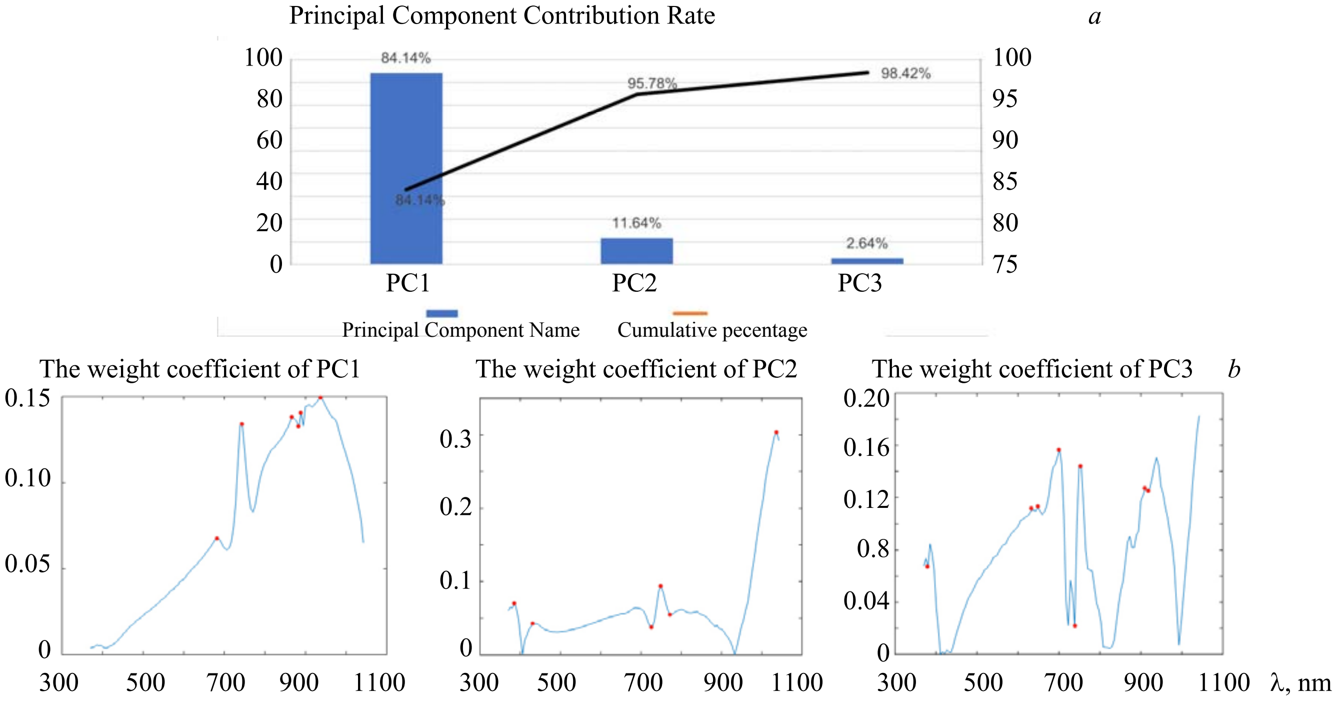

Fig. 2. Characteristic wavelength extraction results of PCA: (a) Principal component contribution of spectral reflectivity; (b) The first three principal component loadings of PCA.

TABLE 2. Characteristic Wavelength Selection Results

Principal component number	Characteristic wavelengths, nm
PC1	685.46, 743.79, 872.65, 867.24, 888.91, 937.89
PC2	385.39, 431.12, 727.84, 749.12, 781.15, 1309.76
PC3	375.26, 632.82, 648.57, 701.32, 738.47, 754.45, 905.2, 910.64

SVM classification performance. Spectral data of 600 samples of each pine species were obtained, and there was a total of 3000 samples. The samples were randomly divided into the train set and the test set in a ratio of 4:1, with the training sample size totaling 2,400, and the test sample size totaling 600. The SVM spectral classification model was established with the full wavelengths (128 features) and the characteristic wavelengths (20 features) as input. Using RBF as a kernel function, each model training determined the optimal penalty factor (c) and the kernel function radius (g) by means of cross-validation. In the classification model based on spectral data, the accuracy rate based on the full-wavelength model train set was 100%, and the accuracy rate based on the characteristic wavelength model was 97.2%. The classification accuracy rate based on the full-wavelength test set was 97.7%. Based on the characteristic wavelengths, the classification accuracy of the test set was 93.1%.

The penalty factor is a weight used for the weight loss and classification interval. For classification problems, the larger the penalty factor, the more important the loss. When a particularly large penalty factor is selected, if there are wrongly classified samples, the penalty will be considerably large, which will lead to a hard interval effect. Figure 3 shows the classification results on the test set in different spectral datasets. An observation can be made that when the penalty factor was larger, the classification result was better.

LeNet-5 classification performance. The difference between the classification accuracy based on full-band and characteristic wavelengths was within 5%. To avoid unnecessary calculations, referring to the characteristic wavelengths extracted by the PCA, the images corresponding to the 20 characteristic wavelengths were chosen for training of the CNN. The train set and test set were divided in the same way. A total of 40 000 images were used for model building, and 20,000 images constituted a test set. The LeNet-5 network structure was constructed, in which the learning rate was 0.1; the learning rate decay rate was 0.99; and the regularization coefficient was 0.0001. The network structure and parameters of Lenet-5 were improved by changing the number of output filters and network structure of the convolutional layer (adding the dropout layer). The numbers in brackets represent the dimensions of the output spaces of the first and second convolutions, respectively (that is, the number of output filters). The following results could be obtained (Table 3).

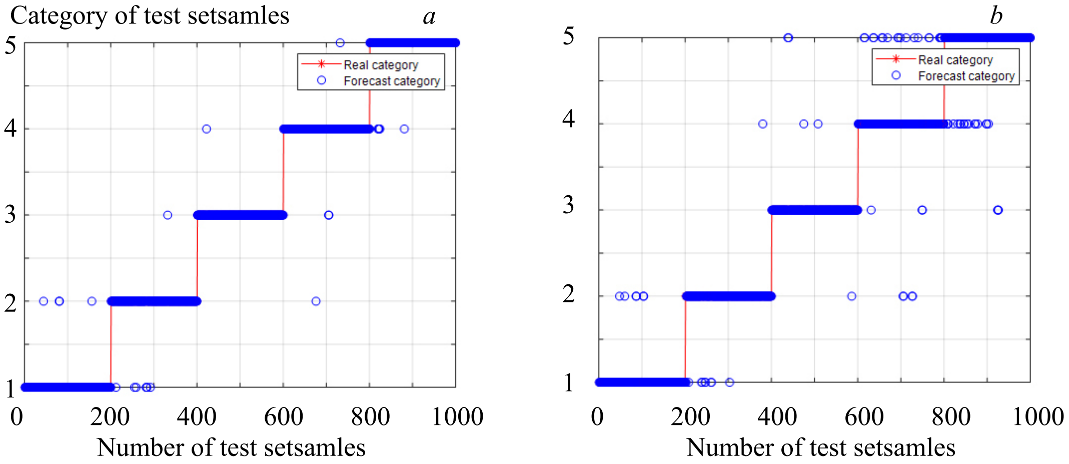

Fig. 3. SVM model classification results: (a) Full wavelengths (c = 2, g = 1); (b) Characteristic wavelengths (c = 0.5, g = 1).

TABLE 3. Network Model Running Results

Network model	Parameters	Training accuracy, %	Testing accuracy, %	Training loss	Testing loss	Epoch
LeNet-5 (32-64)	703.101	94.89	93.06	0.1523	0.2002	200
LeNet-5 (28-10)	62.007	98.12	99.06	0.0524	0.0324	96
LeNet-5 (16-6)	35.587	99.62	99.44	0.0161	0.0239	71
LeNet-5 (6-16)	35.587	99.70	99.58	0.0102	0.0149	157
LeNet-5 (16-6+dropout)	35.587	96.52	99.72	0.1024	0.0153	119
LeNet-5 (6-16+dropout)	89.097	95.99	99.92	0.1130	0.0182	113

An observation can be made from Table 3 that the number of convolution output filters had a significant influence on the convolution speed. Comparing the training speeds of the four groups of convolution output filters with different numbers, the experimental data set used a network with a small number of convolution output filters to meet the requirements, with fast training speed and high efficiency.

After adding a dropout layer after the flatten layer and the first dense layer with a dropout ratio of 0.1, the experimental results show that in the improved network model, the time (the number of iterations) for the training set and the test set to reach the fitting varied significantly. Notably, the fitted loss function curve and accuracy curve were significantly smoother than before.

After swapping the number of output filters in the two convolutional layers, the results show that the duration of each iteration varied significantly. Taking LeNet-5(16-6) and LeNet-5(6-16)) as examples to compare the results, the former had a larger number of convolutional output filters in the first layer than in the second layer LeNet-5(16-6), while the latter had a larger number of second-layer convolutional output filters than the first-layer convolutional output filters LeNet-5(6-16). The number of iterations required for the former to reach the fit was significantly less than the latter, indicating that reasonable selection of the number of output filters of each convolutional layer is considerably significant.

Conclusions. The research object was pine seeds of different varieties, and hyperspectral imaging technology was used to realize the rapid and nondestructive identification of seed varieties. PCA was adopted to qualitatively analyze and extract characteristic wavelengths. In the SVM classification model, 128 spectral features in the full wavelengths and 20 spectral features in the characteristic wavelengths were well represented. The recognition accuracies of the test set were 97.7 and 93.1%, respectively. The CNN training results were even better, and after many iterations, the accuracy rate was over 99%. Among the network models, the improved Lenet-5 performed better. As such, in convolutional neural networks, the selection of the number of convolutional channels and the design of the network structure are of considerable significance. The results of the present study provide evidence that convolutional neural networks perform well in the application of seed variety recognition, and verify the superiority of convolutional neural networks in processing a large amount of image data. In subsequent experiments, more types of seeds will be included, and images of different wavelengths will be compared with different network models.

Acknowledgments. This research was supported by the National Natural Science Foundation of China (Grant No. 31770769).

REFERENCES

1. R. Florin, *Biol. Rev.*, **29**, No. 4, 367–389 (2010).
2. M. Hizume, *Natural Sci.*, **8**, 1–108 (1988).
3. L. Linchu, *Guangxi Plants*, 324–328 (1988).
4. J. G. Niemann, *Acta Bot. Neerl.*, **28**, 73–88 (1979).
5. R. A. Price, J. Olsen-Stojkovich, J. M. Lowenstein, *Syst. Bot.*, **12**, 91–97 (1987).
6. A. Guri, P. Kefalas, V. Roussis, *Phytother. Res.*, **20**, 263–266 (2006).
7. G. Dong, J. Guo, C. Wang, Z. L. Chen, D. Z. Zhu, *Spectrosc. Spectr. Anal.*, **35**, 3369 (2015).
8. J. Zhang, L. M. Dai, F. Cheng, *Molecules*, **24**, **25** (2019).
9. P. Yuan, L. Pang, L. M. Wang, L. Yan, *Int. Food Res. J.*, **29**, No. 2, 397–405 (2022).
10. I. Baek, M. S. Kim, B. K. Cho, C. Mo, J. Y. Barnaby, A. M. McClung, M. Oh, *Appl. Sci.-Basel.*, **9**, No. 5, 1027 (2019).
11. D. W. Sun, H. Y. Cen, H. Y. Weng, L. Wan, A. Abdalla, A. I. El-Manawy, Y. M. Zhu, N. Zhao, H. W. Fu, J. Tang, X. L. Li, H. K. Zheng, Q. Y. Shu, F. Liu, Y. He, *Plant Methods*, **15**, 16 (2019).
12. Y. D. Bao, C. X. Mi, N. Wu, F. Liu, Y. He, *Appl. Sci.-Basel.*, **9**, No. 19, 4119 (2019).
13. S. S. Zhu, L. Zhou, P. Gao, Y. D. Bao, Y. He, L. Feng, *Molecules*, **24**, 17 (2019).
14. M. B. McDonald, *Seed Sci. Res.*, **8**, 265–275 (1998).
15. J. S. C. Smith, J. C. Register, *Seed Sci. Res.*, **8**, 285–293 (1998).
16. Z. Chaoliang, W. Bin, *Crop Mag.*, **1**, 13–16 (1998).
17. Q. Zhu, Z. Feng, M. Huang, X. Zhu, *Transact. Chin. Soc. Agric. Eng.*, **28**, 271–276 (2012).
18. S. Javanmardi, S. H. M. Ashtiani, F. J. Verbeek, A. Martynenko, *J. Stored Prod. Res.*, **92** (2021).
19. A. Krizhevsky, I. Sutskever, G. E. Hinton, *Commun. ACM*, **60**, 84–90 (2017).
20. Y. LeCun, Y. Bengio, G. Hinton, *Nature*, **521**, 436–444 (2015).
21. L. B. Wang, J. Y. Liu, J. Zhang, J. Wang, X. F. Fan, *Front. Plant Sci.*, **13**, 168–174 (2022).
22. W. Lu, R. T. Du, P. S. Niu, G. N. Xing, H. Luo, Y. M. Deng, L. Shu, *Front. Plant Sci.*, **12**, 791256 (2022).
23. H. Li, L. Zhang, H. Sun, Z. H. Rao, H. Y. Ji, *J. Food Proc. Eng.*, **44**, e13767 (2021).
24. G. Y. Zhao, L. Z. Quan, H. L. Li, H. Q. Feng, S. W. Li, S. H. Zhang, R. Q. Liu, *Comput. Electron. Agric.*, **187**, 106230 (2021).
25. N. Otsu, *IEEE Trans. Syst. Man Cybern.*, **9**, 62–66 (1979).
26. A. Femenias, M. B. Bainotti, F. Gatius, A. J. Ramos, S. Marin, *Food Res. Int.*, **139**, 109925 (2021).
27. B. Hasanzadeh, Y. Abbaspour-Gilandeh, A. Soltani-Nazarloo, M. Hernandez-Hernandez, I. Gallardo-Bernal, J. L. Hernandez-Hernandez, *Horticulturae*, **8**, 598 (2022).
28. M. Gabrielli, V. Lancon-Verdier, P. Picouet, C. Maury, *Chemosensors*, **9**, 71 (2021).
29. R. J. Barnes, M. S. Dhanoa, Susan, J. Lister, *Appl. Spectrosc.*, **43**, 772–777 (1989).
30. Karl Pearson F.R.S., *Philos. Mag.*, **2**, 559–572 (1979).
31. D. Liu, J. Ma, D. W. Sun, H. Pu, W. Gao, J. Qu, X. A. Zeng, *Food Bioproc. Technol.*, **7**, 3100–3108 (2014).

32. J. Platt, *Tech. Rep.*, **98**, 14 (1998).

33. G. Mountrakis, J. Im, C. Ogole, *ISPRS J. Photogram. Remote Sens.*, **66**, 247–259 (2011).

34. K. Were, D. T. Bui, Ø. B. Dick, B. R. Singh, *Ecol. Indic.*, **52**, 394–403 (2015).

35. V. Rodriguez-Galiano, M. Sanchez-Castillo, M. Chica-Olmo, M. Chica-Rivas, *Ore Geol. Rev.*, **71**, 804–818 (2015).

36. C. Cortes, V. Vapnik, *Mach. Learn.*, **20**, 273–297 (1995).

37. Y. Lecun, L. Bottou, *Proc. IEEE*, **86**, 2278–2324 (1998).

38. X. Zhang, L. Z. Zhou, L. L. Zhang, *Computer and Modernization*, 5655–5660 (2019).