European Food Research and Technology (2024) 250:191–201
https://doi.org/10.1007/s00217-023-04377-9

ORIGINAL PAPER

Identification of varieties of sorghum based on a competitive adaptive reweighted sampling-random forest process

Kai Wu[1] · Tingyu Zhu[1] · Zhiqiang Wang[1] · Xuerong Zhao[1] · Ming Yuan[1] · Du Liang[2] · Zhiwei Li[1,3]

Received: 29 June 2023 / Revised: 12 September 2023 / Accepted: 16 September 2023 / Published online: 26 October 2023

Abstract

To address issues relating to the advantages of varieties of sorghum in construction, breeding, and brewing of the seed repository, a visible–near-infrared hyperspectral imaging (VNIR-HSI) non-destructive technique was proposed to detect different varieties of sorghum. The VNIR-HSI system was used to collect spectrum images for 27 types of varieties of sorghum, and the spectral data were pre-processed using Savitzky–Golay (S–G) smoothing filters, the standard normal variate (SNV), and multiplicative scatter correction (MSC). Competitive adaptive reweighted sampling (CARS) was used for dimensionality reduction. Based on full-spectrum and characteristic spectral data, classification models were developed using a random forest (RF) algorithm. The tested results indicated that precisions of calibration and prediction sets of the RF model established based on the full-spectrum reach 94.58% and 64.44%, respectively. The CARS algorithm was adopted to extract 20 characteristic wavelengths from sorghum spectra. The precisions of the calibration and prediction sets for the CARS-RF model reach 95.00% and 84.07%, respectively. Using the confusion matrix to calculate Cohen's kappa values, the calibration and prediction Cohen's kappa values for the full sample were 0.9212 and 0.9231 respectively, indicating that the evaluation results are almost identical to the correctness results. The applied model can achieve favorable effects when detecting each cultivar of sorghum. The results show that the modeling method integrating VNIR-HSI technique with CARS-RF can provide a rapid non-destructive testing method for detection of varieties of sorghum, and offer an idea for detecting cultivars of coarse cereal crops.

Keywords Sorghum · Hyperspectral imaging · Characteristic extraction · Random forest · Identification of varieties

Introduction

Sorghum is an ancient cereal crop, characterized by strong disease resistance, wide adaptability, tolerance to barrenness, aridity, saline, and alkali stresses, and tolerance to water-soluble fertilizer. It can also be planted in arid hills and barren mountainous areas where staple crops requiring high level of water-soluble fertilizer are inappropriate to be planted [1]. Sorghum planting can increase the efficiency of the grain-production process, provide diverse types of grain, and promote the development of livestock farming [2, 3]. Moreover, Sorghum can be used as an industrial raw material applied in products, such as starch, alcohol, vitamins, and biomass. In particular, it is broadly applied in the brewing industry and has significance to broader socio-economic development [3–8]. However, the differences between varieties of sorghum are significant so accurate identification of the varieties is conducive to determination of the characteristics of different varieties of sorghum. This helps sorghum planters to choose the varieties suitable for their planting conditions and demands to set more effective planting management strategies [4]. Meanwhile, it can also ensure diversity and integrity of varieties in the seed repository and guarantee the effective utilization and development of germplasm resources. During breeding, identification of the varieties can help breeding operators to make a rational selection of the parents and avoid hybridization and mixing of different varieties. This can improve breeding efficiency and the rate of success, also facilitating the determination of

✉ Zhiwei Li
lizhiweitong@163.com

[1] College of Agricultural Engineering, Shanxi Agricultural University, Jinzhong 030801, China

[2] Sorghum Research Institute, Shanxi Agricultural University, Jinzhong 030600, China

[3] College of Information Science and Engineering, Shanxi Agricultural University, Jinzhong 030801, China

the genetic purity of hybrid offspring for hybrid and related varieties [4–6]. In doing so, breeding hybrid failure can be avoided. Identification of the varieties is of significance to improvement of varieties, new cultivar breeding, and their popularization. By means of detection of the varieties, the contents of starch and tannin in sorghum grain can be obtained. The research provides supports for brewing enterprises, also helping to justify the present research into the identification of varieties of sorghum.

Different sorghum varieties have different growth characteristics and agronomic traits. Traditional methods of detection of varieties, such as morphological and physiological methods, fail to identify precisely the varieties due to small differences among the varieties of various families and genera. Chemical analysis methods such as phenol staining, and DNA-based molecular technique require grinding of the grains, rendering the process inefficient, while preventing rapid non-destructive assay [9]. As the hyperspectral imaging technique is a rapid, non-destructive identification technique, it integrates merits, such as machine vision and spectral analysis. It has been applied in detection and identification of varieties for crops, such as wheat, corn, and millet [10]. Different sorghum varieties have unique reflectance characteristics in the hyperspectral bands, and by analyzing this spectral data, differences between varieties can be identified. Compared to traditional observation and measurement methods, hyperspectral imaging technology has higher spectral resolution and accuracy, providing more accurate variety identification results. At the same time, labor and time costs are reduced. Due to a huge amount of spectral data and redundancy generated when using the hyperspectral technique, feature extraction algorithms were used to conduct dimensionality reduction on the spectra data and chemical significance of selecting variables becomes easier to elucidate [11, 12]. Song et al. performed principal component analysis to extract 25 characteristic wavelengths of wheat spectral and used a support vector machine (SVM) to construct classification models, with an accuracy of 97.54% [13]. Zhu et al. utilized a deep convolutional neural network to identify wheat grains [14]. Yang et al. used a successive projection algorithm to extract 19 characteristic wavelengths from the hyperspectral data of waxy corn seed varieties and used an SVM to establish cultivar classification models with an accuracy of 98.2% [15]. Kabir et al. used visible–near-infrared spectroscopy on 16 varieties of millet and integrated principal component analysis (PCA) dimensionality reduction with the *k*-nearest-neighbor (KNN) algorithm, linear discriminant analysis, logistic regression, random forest (RF), and SVM algorithms, respectively to build cultivar classification methods [16]. The results indicated that RF and SVM were the most effective. At present, research into rapid identification of varieties of sorghum using VNIR-HSI technique is sparse, thus, the research is undertaken with the aim of obtaining an efficient non-destructive testing method for distinguishing between varieties of sorghum.

In these experiments, taking 27 varieties of sorghum as research objects, Innovation includes the following three points:

(1) Visible and near-infrared spectroscopy were used to acquire the sample information.
(2) Competitive adaptive reweighted sampling (CARS) was then integrated with RF to obtain a model for the identification of varieties of sorghum.
(3) Indices, such as identification precision and Cohen's kappa value, were adopted to test the performance of the proposed method which was applied across the entire sample and each single cultivar therein.

In doing so, a rapid, non-destructive method of identification of varieties of sorghum was developed (Fig. 1).

Materials and methods

Experimental samples

A total of 3240 samples of 27 varieties of sorghum (120 samples for each cultivar) cultivated from the Sorghum Research Institute of Shanxi Agricultural University were selected. The sorghum seeds selected that were plump and highly regular in shape were classified, then sealed for storage in a jar. The type of the varieties of sorghum is represented by V1–V27, as shown in Fig. 2, among which, the varieties (V1, V3, V9, V13, V15, V16, and V24) are glutinous varieties of sorghum, having a white color. The other varieties of sorghum appear red in color. Other physical properties of the sorghum seeds are similar. In this experiment, non-glutinous and glutinous sorghum seeds were randomly arranged to increase the adaptability of the model.

Hyperspectral imaging system

The VNIR-HSI (Headwall Photonics, USA) scanning platform was used to finish the image acquisition of sample varieties of sorghum. This system consists of a canning platform, a micro near-infrared hyperspectral imager with an aperture of 1.4 and focal length of 25 mm, light source, a controller, and computer. Its spectral range is 380–1000 nm at a spectral resolution of 0.727 nm, with a total of 856 wave bands. The image acquisition parameters are as follows: the object distance is 370 mm, the pushing distance is 100 mm, and the platform was driven at a speed of 2.938 mm s^{-1}, in doing so, clear, undistorted images can be captured.

To reduce interference caused by systemic light sources and the dark current, a black-and-white correction

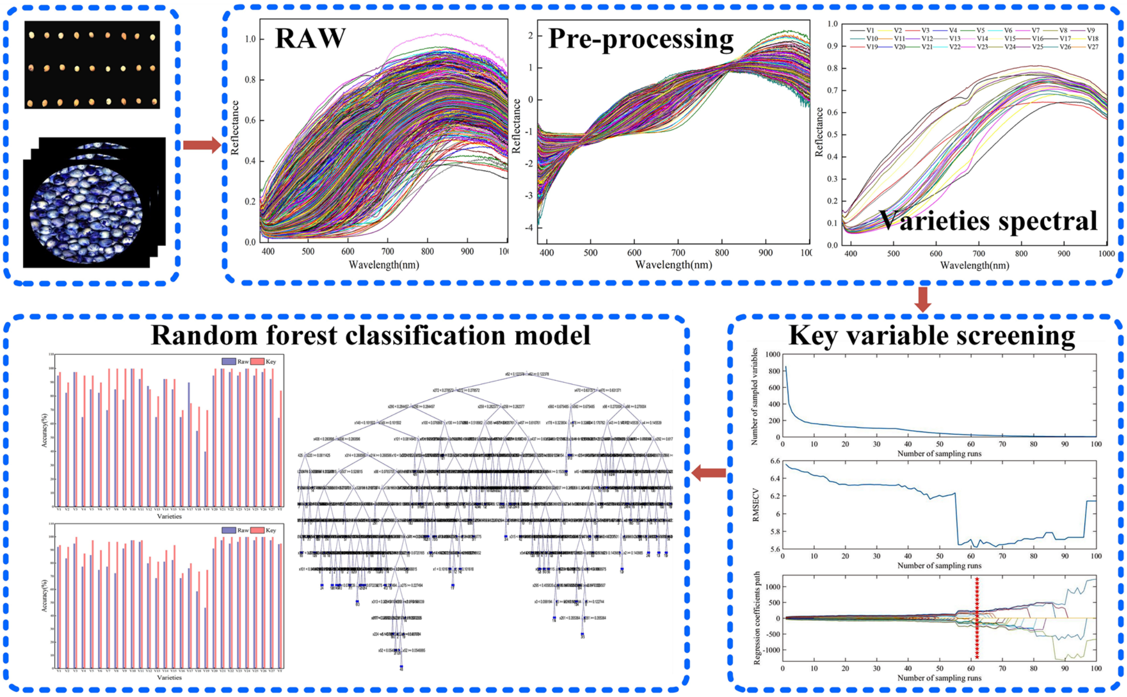
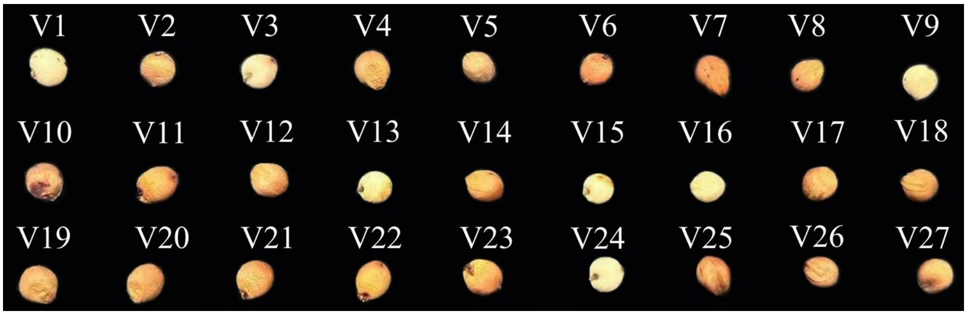


Fig. 1 Analysis process flow chart

Fig. 2 Sample diagram



was applied to the hyperspectral images based on the equation $R = \frac{R_0 - R_b}{R_w - R_b}$, among which R refers to the corrected hyperspectral images; R_0 denotes the original hyperspectral images; R_w represents the white background image of a standard white correction board with a reflectivity approaching 99.9%; R_b represents the dark black background image with a reflectivity of 0% taken after closing the lens cover.

The samples of varieties of sorghum were placed into a sample dish with a diameter of 40 mm and depth of 15 mm, leveled smooth, and compacted. Afterward, they were placed on the movable scanning platform to allow collection of hyperspectral images. Image acquisition on the samples of each cultivar was conducted 120 times, producing a total of 3240 hyperspectral images.

Data extraction and pre-processing

Data extraction

Hyperspectral images contain both the spectral and image information from sorghum samples. Each pixel point on the image corresponds to a curve on the diffuse reflectance spectrum. ENVI 5.0 [17] was adopted to extract the spectral data of the region of interest (ROI) for each hyperspectral image: the reflectivity of each pixel point was calculated, and its arithmetic mean was taken as basic data for subsequent processing.

Data pre-processing

The data pertaining to the influences of noise and light scattering of the instrument in the data acquisition process were obtained. This work successively used Savitzky–Golay (S–G) [17, 18] smoothing filters, the standard normal variate (SNV) [19, 20], and multiplicative scatter correction (MSC) [21, 22] to pre-process the spectral data. This can further eliminate random errors, remove the influence of light scattering to improve the signal-to-noise ratio (SNR), enhance the correlation between the light spectrum and data, effectively eliminating instrument noise from the spectral data. On this basis, the performance of the established model can be improved. Considering the stability and general adaptability of the constructed model, the total number of samples in the modeling part is 3240, and the calibration set and the prediction set are randomly divided according to a 2:1 ratio using Matlab to generate random numbers, where the calibration set contains 2160 samples and the prediction set contains 1080 samples.

Extraction of characteristic wavelengths

The hyperspectral can effectively provide quantification data. However, a host of variables leads to significant redundancy, reducing the model load and prediction capability. To improve the precision of the predictive model, this experiment adopted CARS for dimensionality reduction, further improving interpretability of the variables.

The CARS algorithm is used to select the optimal combination of the effective variables in the spectra by mimicking Darwin's "survival of the fittest" principle [23–29]. For the spectral variables of dimensions $m \times p$, CARS adopted the effective variables based on the following steps:

1. Based on Monte Carlo sampling (MCS), 80% of the samples were randomly selected from the correction set to construct the PLS model. The regression coefficient $|K_i|$ ($i = 1, 2, \ldots, p$) of the ith wavelength can be obtained;
2. Exponentially decreasing function (EDF) was adopted to eliminate smaller wavelength points of $|K_i|$. The retention rate of the variable $r_j = ae^{-bj}$ ($j = 1, 2, \ldots, N$): where j denotes the jth MCS; N represents the total number of MCS operations. The parameters (a and b) are constants. According to $r_1 = 1$ and $r_N = 2/p$, they can be calculated thus

$$a = (p/2)^{1/(N-1)} \quad (1)$$

$$b = \ln(p/2)/(N-1) \quad (2)$$

3. Based on an adaptive reweighted sampling (ARS) technique, the variable was further screened. According to Darwin's principle of the survival of the fittest, $w_i = |K_i| / \sum_{i=1}^{p} |K_i|$ ($i = 1, 2, \ldots, p$) was adopted for variable selection;
4. The aforementioned steps are repeated until the number of MCS reach its pre-set value N;
5. Using tenfold cross-validation, the root mean square error of cross-validation (RMSE_{CV}) can be used as the evaluation standard. Then, the value of the variable subset can be obtained by comparing difference of each MSC. The variable subset corresponding to the minimal RMSE_{CV} value is seen as the optimal variable.

The classification model and evaluation criteria

The classification model

The RF algorithm, which is widely applied in issues relating to data regression and classification, is an intelligent combinational classification algorithm [30–33]. A self-sampling method (bootstrap) is used to repeatedly and randomly select n samples from the original calibration set N to generate new calibration sample decision trees by placement. Repeating the aforementioned steps can help generate m decision trees to form an RF. The classified results of new data are determined according to the scores of the polling of classification trees. The classification capability of single tree may be small. However, after many decision trees are generated, after generating statistics pertaining to the classified results for each tree in each tested sample, the most possible type of classification can be ascertained.

The steps can be described as follows:

1. N is used to express the number of training examples (samples) and M denotes the number of features;
2. The number of features m is input for determining the decision results of the last node at the decision tree, among which, m should be less than M;
3. Using a method of random sampling with replacement in the N^{th} training examples (samples), N sampling operations can be conducted to form a training set, namely, (bootstrap) and then unselected examples (samples) are used for prediction to estimate the error of the method;
4. For each node, m features are randomly selected and each node at the decision tree is determined based on these features. According to m features, the optimal method of division can be calculated. The purpose of randomly selecting the training set includes: if random sampling is not made, the training sets for each of different trees are same. Hence, the classified results of finally trained trees are identical. The objectives of sampling

with replacement involve the following: if sampling with replacement is not used, the training sample for each different tree differs, showing no intersection. That is to say, the trained results from each tree are largely different;

5. Finally, RF is voted using multiple trees to determine the type of the samples.

Evaluation criteria

The RF method was used to build a prediction and classification model of varieties of sorghum. Then the performance of the model was estimated using the classification accuracy. The accuracy can be calculated thus

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}} \times 100 \tag{3}$$

where ACC refers to the accuracy; TP refers to true positive; TN refers to true negative; FP refers to false positive; FN refers to false negative.

Meanwhile, in order to visualize the effect of the algorithm, the confusion matrix (CM) of the sorghum variety classification results was constructed and Cohen's kappa value was calculated using the formula:

$$K = \frac{P_\mathrm{o} - P_\mathrm{e}}{1 - P_\mathrm{e}} \tag{4}$$

where K: Cohen's kappa; P_o: proportion of observation-consistent units; P_e: proportion of chance consistent units. Cohen's kappa is calculated as − 1 to 1, but usually the kappa falls between 0 and 1, which can be divided into five groups to indicate different levels of agreement: 0.0–0.20 very low agreement (low), 0.21–0.40 fair agreement, 0.41–0.60 moderate agreement (moderate), 0.61–0.80 high agreement (substantial), and 0.81–1 almost perfect.

Results and discussion

Spectral characteristics

Figure 3 describes the original spectral curve of 3240 samples selected in this experiment and pre-processed spectral curve. Comparison of Fig. 3a and d shows that the influences of light scattering and noise on the spectra after pre-processing with S–G, SNV, and MSC methods are eliminated as soon as possible to improve SNR, which is conducive to improving the subsequent classification accuracy of the model.

Figure 4 depicts the average spectra of 27 varieties of sorghum. Within the wavelength range of 380–1000 nm, the

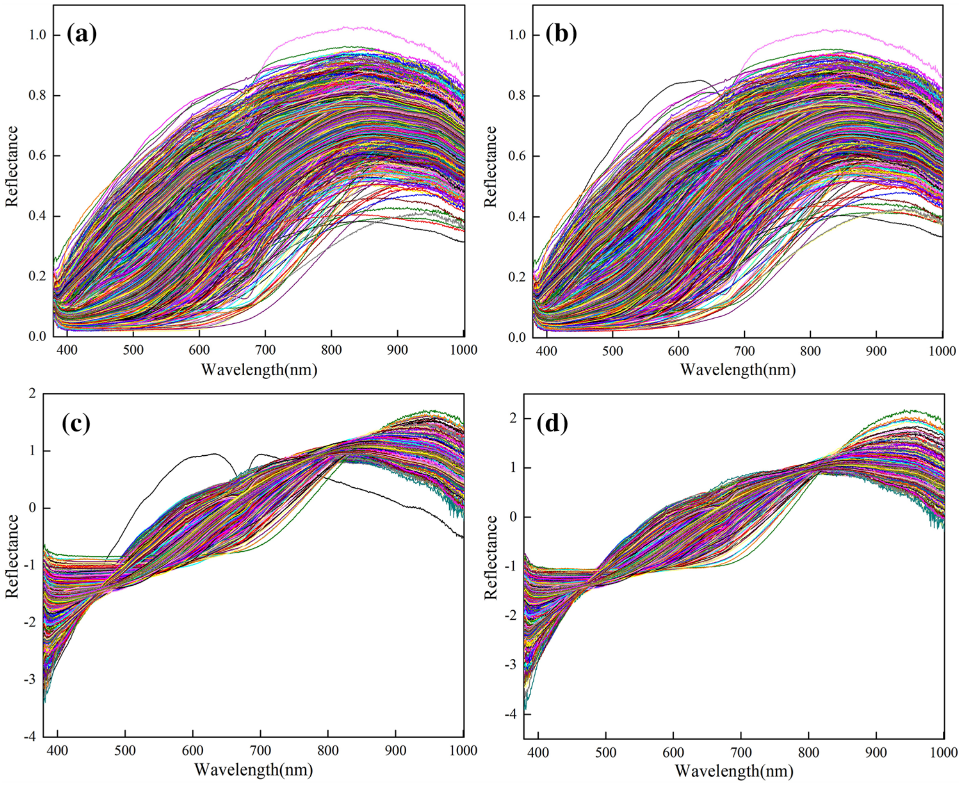


Fig. 3 Raw spectra and pretreatment of the samples. **a** Raw spectra. **b** Spectral pre-processing by S–G. **c** Spectral pre-processing by S–G and SNV. **d** Spectral pre-processing by S–G SNV and MSC

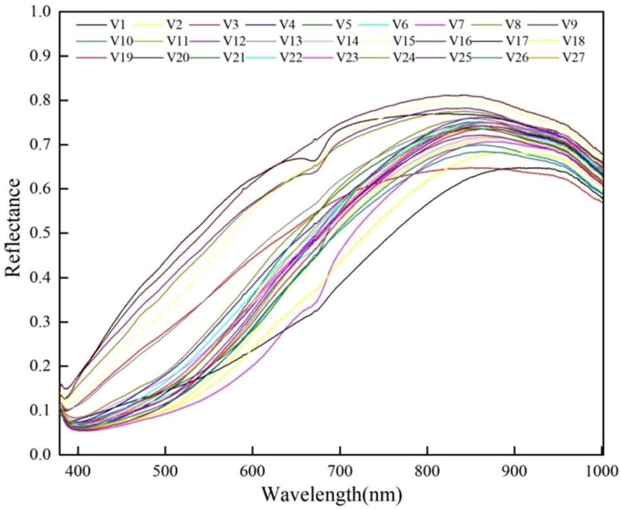


Fig. 4 Average spectra of 27 samples

general trends of the sorghum spectral curves are similar. Wave peak and trough show relatively little change. Meanwhile, some curves intersect and overlap. Figure 4 displays the pre-processing of the average spectra for 27 varieties of sorghum. After pre-processing, the curve intersection and overlapping can be reduced, the noise is reduced, allowing clearer distinction between different samples. Among these, the spectral curves of the varieties (V1, V3, V9, V13, V15, V16, and V24) at wavelengths ranging from 400 to 850 nm are shown above the rest of the varieties. This information can be used to differentiate red sorghum from white sorghum: a trough appears at around 670 nm in the spectral curve, possibly caused by a bathochrome effect, while the spectral curve at wavelengths greater than 850 nm presents a declining trend. This is possibly related to the stretching vibration of molecular bonds for O_2, O–H bonds, and hydroxyl functional groups [28]. These differences provide an effective discrimination basis for varieties of sorghum when using spectral identification.

Characteristic extraction

The aim of CARS is to eliminate irrelevant variables and reduce collinearity among variables. Figure 5 describes with the increase in the number of MCS, changes in the number of sample variables in the subsets of sorghum samples, $RMSE_{CV}$, and regression coefficient are shown in the subsets of sorghum samples. As the number of MCS is increased, the variables selected from the effect of EDF present an

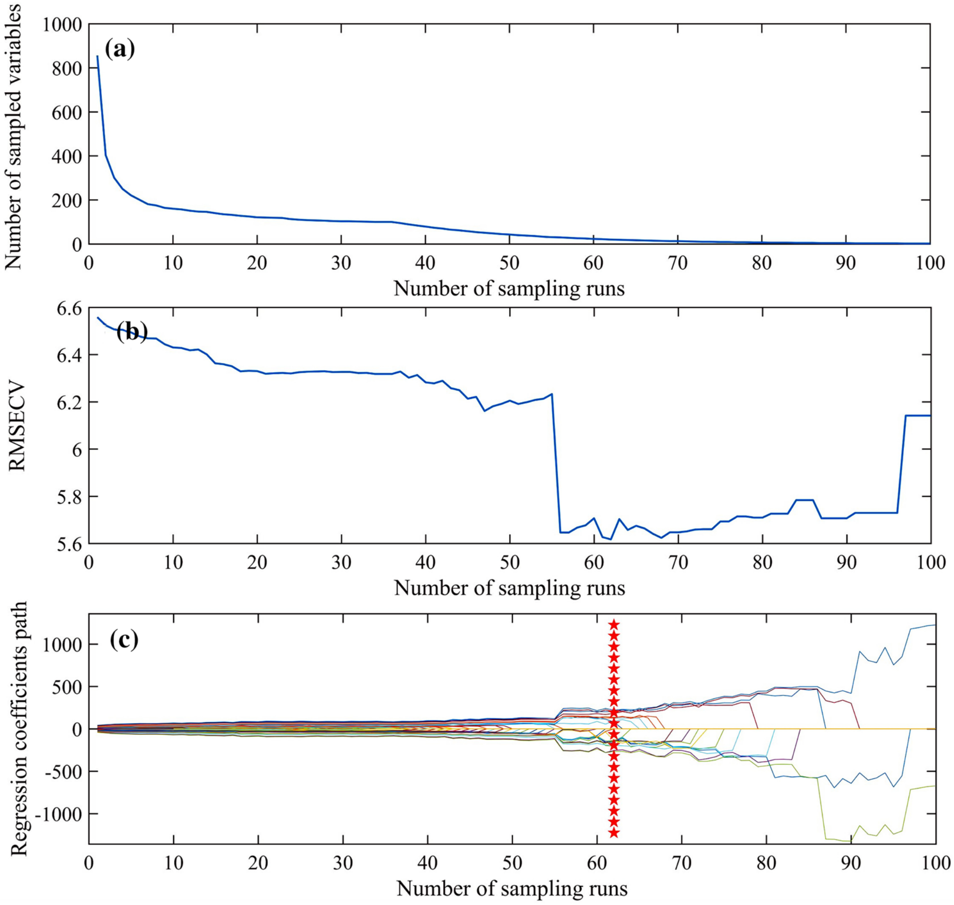


Fig. 5 Key extraction results of the CARS algorithm. **a** Changes in the number of waveband variables. **b** Variation of RMSEcv. **c** Path of variable regression coefficients. Five-pointed star denotes the optimal point where root mean square error of cross-validation ($RMSE_{CV}$) values achieve the lowest

exponential decrease and then gradually tend to stabilize; the value of $\mathrm{RMSE}_{\mathrm{CV}}$ first decreases, then increases, indicating an absence of correlation between the variables initially eliminated during the variable screening process and components to be tested. Thereafter, the variables irrelevant to the components to be tested are added to the variable subsets; effective wavelengths are retained at the labeled position, at this time, the value of $\mathrm{RMSE}_{\mathrm{CV}}$ is minimal. The screened variables are the optimal variable combination, as shown in the figure, when the number of sampling operations is 62, $\mathrm{RMSE}_{\mathrm{CV}}$ is 5.6174 (its minimum value). At this time, there are no redundant wavelengths to be screened. The subset contains 20 variables, the corresponding wavelengths are 425, 426, 429, 430, 588, 589, 591, 592, 661, 662, 668, 669, 672, 673, 686, 880, 881, 885, 911, and 915 nm.

Result of classification

Identification precision

Full-spectral data for 27 types of varieties of sorghum and 20 characteristic spectral data were respectively used to establish RF-based classification models. When using the full-spectrum data, the number of trees (ntree) is 1000, and the number of features (mtry) randomly sampled is 29. The experimental results demonstrate that the accuracy of the calibration set is 94.58% while the accuracy of the prediction set is 64.44%; when using characteristic spectral data, the number of trees generated (ntree) is 1000 and the number of randomly sampled features (mtry) is 4. Among them, ntree and mtry are two important parameters of random forests. ntree is the number of base classifiers included, with a default of 500; mtry is the number of variables included in each decision tree, with a default of logN. By optimizing the parameters, the error is minimized when mtry is taken to be 29 for full-spectrum data modeling, mtry is minimized when mtry is taken to be 4 for eigenspectral data modeling, and the error is stable when ntree is taken to be 1000.

Experimental results indicate that the accuracy of the correction set is 95.00% while the accuracy of the prediction set is 84.07%. The overall accuracy of the testing sets for the classification models built using CARS -RF can be significantly increased, as shown in Fig. 6.

Cohen's kappa value

In this paper, the confusion matrix is established for the classification results of the classification model established based on the feature spectral data, and its Cohen's kappa value is calculated, and the calibration and prediction Cohen's kappa values of all samples are 0.9212 and 0.9231, respectively, as shown in Fig. 7.

Discussion

Using CARS algorithm, 20 characteristic wavelengths were screened from 27 types of varieties of sorghum. The characteristic wavelength is compressed to 2.4% of the total number of full-wave bands. The spectral data are saved in the matrix with array dimensions of 3240 × 20, thereby reducing the computational time. The characteristic wavelengths are between 420–430 nm, 580–600 nm, 660–690 nm, and 880–920 nm. The characteristic wavelengths occur at 420–430 nm, corresponding to the blue–violet region, which corresponds to the adsorption peak of compounds such as flavonoids contained in sorghum seeds [34, 35], while the wavelength range of 580–600 nm corresponds to the green light region, corresponding to the adsorption peak of pigments including chlorophyll contained in sorghum seeds [36–38]. The wavelength range of 660–690 nm corresponds to the red light region, which mainly involve excitation and fluorescence of chlorophyll contained in sorghum seeds; while at wavelengths of around 880–920 nm, due to the effects of stretching vibrations for C=C, C=O, and C=N, the changes in optical properties of substances including protein, starch, cellulose, and moisture content contained in sorghum at this wavelength range contribute to the phenomenon [39, 40].

According to the classification results, it can be found that when modeling using full-spectrum data, the accuracies of the calibration and prediction sets are 94.58% and 64.44%, respectively. When using the characteristic spectral data, the accuracy of the calibration set is 95.00% and the accuracy of the prediction set is 84.07%. The total accuracy of the prediction set for the classification model established based on CARS-RF is increased by 19.63%. This is because irrelevant variables are eliminated during the extraction of characteristic variables while retaining related characteristic wavelengths pertaining to organic matter contained in sorghum as soon as possible.

When the CARS-RF model was used for classification prediction of each of 21 types of varieties of sorghum (V1–V11, V15, and V20–V27), the accuracies of the correction set are above 92.50%, and the accuracies of the prediction set are higher than 90.00%. The accuracies of some varieties reach 100%, while the accuracies of the prediction set for two varieties of sorghum V12–V13 ranged between 81.25 and 90% and the accuracies of the prediction set are beyond the range of 80.00–85%; while the accuracies of the correction set for four types of varieties of sorghum V16–V19 are in the range of 73.75–80.00%; the accuracies of the prediction set ranged between 70.00 and 75.00%. Owing to physical properties of four types of varieties of sorghum V16–V19 are similar, some deviations occur in the polling process of RF.

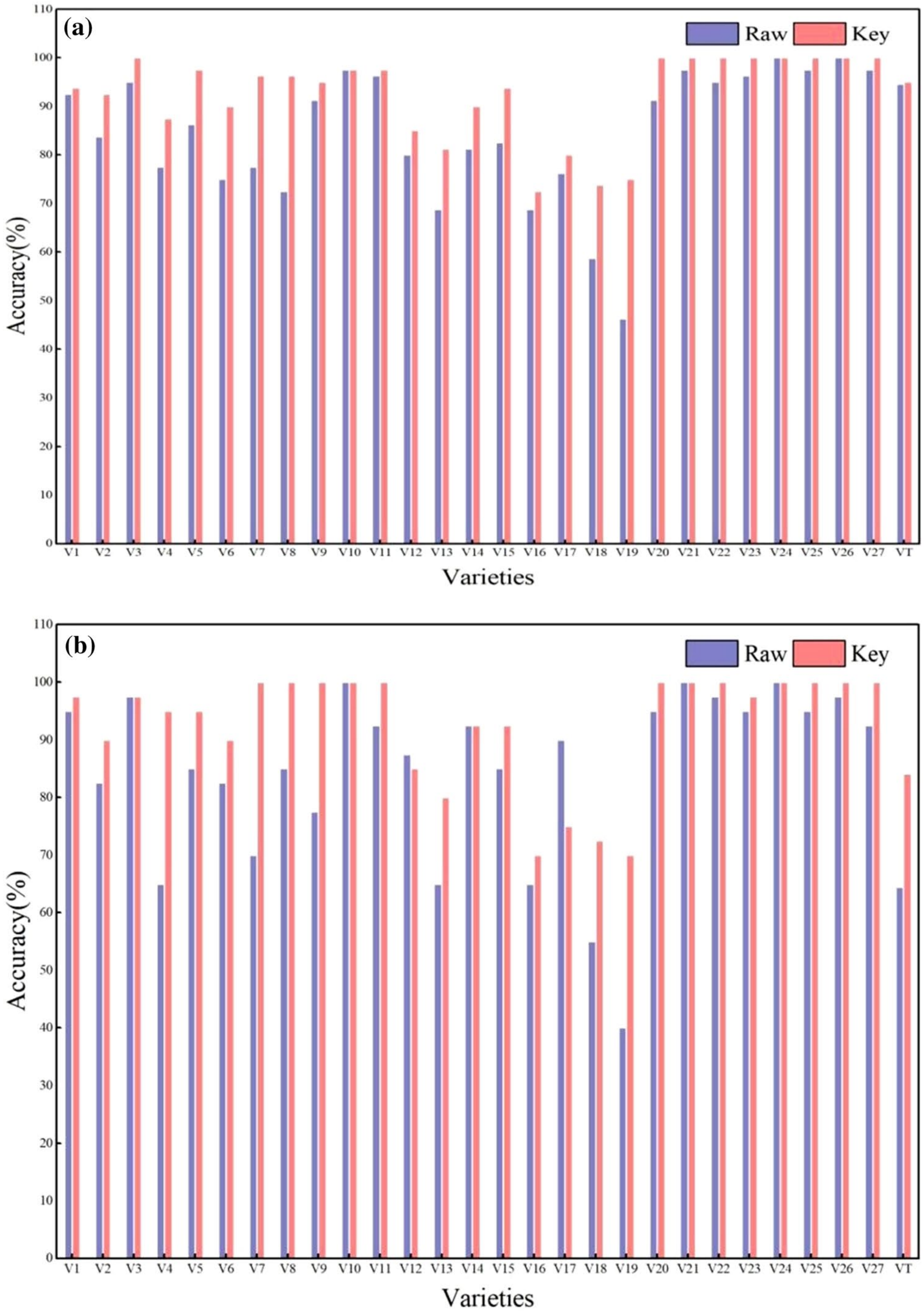


Fig. 6 Results of classification for the calibration and prediction in single and overall sample. **a** Calibration results of RF model based on full and key wavelengths. **b** Prediction results of RF model based on full and key wavelengths

As can be seen from Table 1, using Cohen's kappa value to evaluate the accuracy of the single-species model, the Cohen's kappa values of the calibration and prediction sets of V1–V15, V19–V27 were all in the range of 0.81–1.0, it shows that the evaluation results are almost identical to the accuracy results. The Cohen's kappa values for V16–V18 calibration and prediction sets ranged from 0.61 to 0.8, indicating a high level of consistency. The calibration and prediction Cohen's kappa values for the full sample were 0.9212 and 0.9231 respectively, ranging from 0.81 to 1.0, also falling into the category of almost perfect agreement.

In summary, the CARS-RF was used for modeling of varieties of sorghum. Its identification accuracy meets the classification requirements of the varieties, presenting a certain application value.

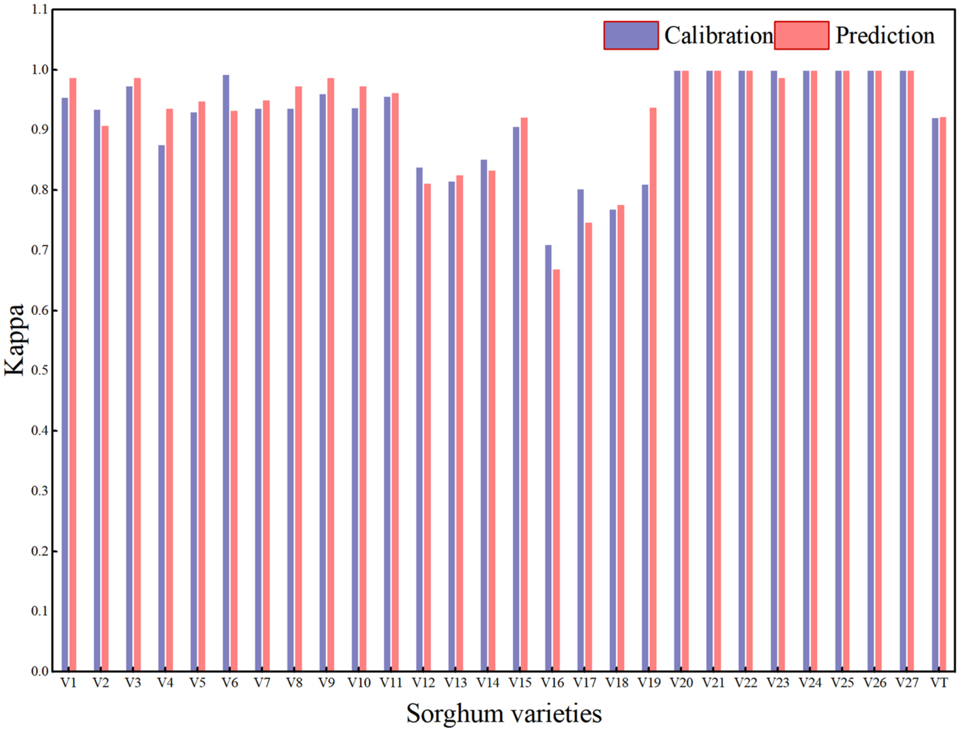


Fig. 7 Results of Cohen's kappa value for the calibration and prediction in single and overall sample

Table 1 Recognition accuracy and Cohen's kappa value for the prediction set

Variety	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
Identification precision	0.9750	0.9000	0.9750	0.9500	0.9500	0.9000	1.0000	1.0000	1.0000	1.0000
Cohen's kappa value	0.9872	0.9078	0.9872	0.9363	0.9481	0.9331	0.9505	0.9740	0.9875	0.9740
Variety	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20
Identification precision	1.0000	0.8500	0.8000	0.9250	0.9250	0.7000	0.7500	0.7250	0.7000	1.0000
Cohen's kappa value	0.9621	0.8119	0.8256	0.8338	0.9215	0.6699	0.7464	0.7664	0.9381	1.0000
Variety	V21	V22	V23	V24	V25	V26	V27	VT		
Identification precision	1.0000	1.0000	0.9750	1.0000	1.0000	1.0000	1.0000	0.8407		
Cohen's kappa value	1.0000	1.0000	0.9872	1.0000	1.0000	1.0000	1.0000	0.9231		

Conclusion

This experiment took 27 types of varieties of sorghum as research objects to identify sorghum seeds using the method integrating the hyperspectral non-destructive detection technique with machine learning. The hyperspectral imaging system was used for acquisition of varieties of sorghum across the average spectral range of 380–1000 nm. Using CARS algorithm, irrelevant variables can be eliminated but effectively retain the characteristic wavelength related to organic matter contained in sorghum the further to realize spectral dimensionality reduction. Meanwhile, RF was used to establish a model capable of identifying varieties of sorghum. The accuracy of a part of the correction set in the model for the varieties V1–V27 is 95.00%. The accuracy of the prediction set is 84.07%. Using the Confusion matrix to calculate Cohen's kappa values, the calibration and prediction Cohen's kappa values for the full sample were 0.9212 and 0.9231 respectively, indicating that the evaluation results are almost identical to the correctness results. The results indicated that non-glutinous and glutinous varieties of sorghum with similar physical properties can be distinguished using the CARS-RF model.

Sorghum variety identification with CARS-RF, not only can accurate, rapid and non-destructive identification of sorghum varieties be achieved, but it can also provide a way of thinking for varietal identification of other crops, as well as helping with seed management, variety protection and germplasm resource management.

Author contributions Conceptualization, KW; methodology, KW, TZ and XZ; investigation, MY, ZW, and DL; writing—original draft, KW; writing—review and editing, KW and ZL. All authors have read and agreed to the published version of the manuscript.

Funding This work was supported by Major Special Projects of National Key R&D, grant number 2021YFD1600301-4; Major Special Projects of Shanxi Province Key R&D, grant number 202102140601013; Major Special Projects of Shanxi Province Key R&D, grant number 201903D211005.

Data availability Available upon request from the corresponding author.

Declarations

Conflict of interest The authors declare no conflict of interests. The funders have no role in the experimental design, data collection and analysis, decision to publish, or preparation of the manuscript.

Compliance with ethics requirements This article does not contain any studies with human or animal subjects.

References

1. Lopes MS, Araus JL, Van Heerden PDR et al (2011) Enhancing drought tolerance in C4 crops. J Exp Bot 62(9):3135–3153. https://doi.org/10.1093/jxb/err105
2. Wang Y, Chan KX, Long SP et al (2021) Towards a dynamic photosynthesis model to guide yield improvement in C4 crops. Plant J 107(2):343–359. https://doi.org/10.1111/tpj.15408
3. Khoddami A, Messina V, Vadabalija Venkata K et al (2021) Sorghum in foods: Functionality and potential in innovative products. Crit Rev Food Sci Nutr 2021:1–17. https://doi.org/10.1080/10408398.2021.1960793
4. Hao H, Li Z, Leng C et al (2021) Sorghum breeding in the genomic era: opportunities and challenges. Theor Appl Genet 134:1899–1924. https://doi.org/10.1007/s00122-021-03789-z
5. Mayor L, Demarco P, Lira S et al (2023) Retrospective study in US commercial sorghum breeding: I. Genetic gain in relation to relative maturity. Crop Sci 63(2):501–510. https://doi.org/10.1002/csc2.20897
6. Dabija A, Ciocan ME, Chetrariu A et al (2021) Maize and sorghum as raw materials for brewing, a review. Appl Sci 11(7):3139. https://doi.org/10.3390/app11073139
7. Shehzad T, Okuizumi H, Kawase M et al (2009) Development of SSR-based sorghum (*Sorghum bicolor* (L.) Moench) diversity research set of germplasm and its evaluation by morphological traits. Genet Resour Crop Evol 56:809–827. https://doi.org/10.1007/s10722-008-9403-1
8. Kaur B, Sandhu KS, Kamal R et al (2021) Omics for the improvement of abiotic, biotic, and agronomic traits in major cereal crops: applications, challenges, and prospects. Plants 10(10):1989. https://doi.org/10.3390/plants10101989
9. Endalamaw C, Adugna A, Mohammed H (2017) Correlation and path coefficient analysis of agronomic and quality traits in a bioenergy crop, sweet sorghum [*Sorghum bicolor* (L.) Moench]. Afr J Biotechnol 16(47):2189–2200. https://doi.org/10.5897/AJB2017.16241
10. Cong S, Liu C, Zhu Z et al (2021) Study on identification of multiple pesticide residues in lettuce leaves based on hyperspectral technology. In: Advances in artificial intelligence and security: 7th international conference, ICAIS 2021, Dublin, Ireland, July 19–23, 2021, proceedings, Part III 7. Springer International Publishing, pp 537–550. https://doi.org/10.1007/978-3-030-78621-2_45
11. Jun S, Xin Z, Hanping M et al (2016) Identification of pesticide residue level in lettuce based on hyperspectra and chlorophyll fluorescence spectra. Int J Agric Biol Eng 9(6):231–239. https://doi.org/10.3965/j.ijabe.20160906.2519
12. Belmerhnia L, Djermoune EH, Carteret C et al (2021) Simultaneous variable selection for the classification of near infrared spectra. Chemom Intell Lab Syst 211:104268. https://doi.org/10.1016/j.chemolab.2021.104268
13. Song XZ, Tang G, Zhang LD et al (2017) Research advance of variable selection algorithms in near infrared spectroscopy analysis. Spectrosc Spectr Anal 37(4):1048–1052. https://doi.org/10.3964/j.issn.1000-0593(2017)04-1048-05
14. Zhu J, Li H, Rao Z et al (2023) Identification of slightly sprouted wheat kernels using hyperspectral imaging technology and different deep convolutional neural networks. Food Control 143:109291. https://doi.org/10.1016/j.foodcont.2022.109291
15. Yang H, Cheng Y, Li G (2021) A denoising method for ship radiated noise based on Spearman variational mode decomposition, spatial-dependence recurrence sample entropy, improved wavelet threshold denoising, and Savitzky–Golay filter. Alex Eng J 60(3):3379–3400. https://doi.org/10.1016/j.aej.2021.01.055
16. Kabir MH, Guindo ML, Chen R et al (2021) Geographic origin discrimination of millet using Vis–NIR spectroscopy combined with machine learning techniques. Foods 10(11):2767. https://doi.org/10.3390/foods10112767
17. Song X, Du G, Li Q et al (2020) Rapid spectral analysis of agro-products using an optimal strategy: dynamic backward interval PLS-competitive adaptive reweighted sampling. Anal Bioanal Chem 412:2795–2804. https://doi.org/10.1007/s00216-020-02506-x
18. Chen Y, Cao R, Chen J et al (2021) A practical approach to reconstruct high-quality Landsat NDVI time-series data by gap filling and the Savitzky–Golay filter. ISPRS J Photogramm Remote Sens 180:174–190. https://doi.org/10.1016/j.isprsjprs.2021.08.015
19. Mishra P, Marini F, Biancolillo A et al (2021) Improved prediction of fuel properties with near-infrared spectroscopy using a complementary sequential fusion of scatter correction techniques. Talanta 223:121693. https://doi.org/10.1016/j.talanta.2020.121693
20. Wrobel TP, Liberda D, Koziol P et al (2020) Comparison of the new Mie extinction extended multiplicative scattering correction and resonant mie extended multiplicative scattering correction in transmission infrared tissue image scattering correction. Infrared Phys Technol 107:103291. https://doi.org/10.1016/j.infrared.2020.103291
21. Mishra P, Rutledge DN, Roger JM et al (2021) Chemometric pre-processing can negatively affect the performance of near-infrared spectroscopy models for fruit quality prediction. Talanta 229:122303. https://doi.org/10.1016/j.talanta.2021.122303
22. Zhang Z, Ding J, Zhu C et al (2020) Combination of efficient signal pre-processing and optimal band combination algorithm to predict soil organic matter through visible and near-infrared spectra. Spectrochim Acta Part A Mol Biomol Spectrosc 240:118553. https://doi.org/10.1016/j.saa.2020.118553
23. Tang H, Meng X, Su X et al (2021) Hyperspectral prediction on soil organic matter of different types using CARS algorithm. Trans CSAE 37:105–113. https://doi.org/10.11975/j.issn.1002-6819.2021.2.013
24. Lennon JT, den Hollander F, Wilke-Berenguer M et al (2021) Principles of seed banks and the emergence of complexity from

dormancy. Nat Commun 12(1):4807. https://doi.org/10.1038/s41467-021-24733-1

25. Liu J, Dong Z, Xia J et al (2021) Estimation of soil organic matter content based on CARS algorithm coupled with random forest. Spectrochim Acta Part A Mol Biomol Spectrosc 258:119823. https://doi.org/10.1016/j.saa.2021.119823
26. Liu J, Jin S, Bao C et al (2021) Rapid determination of lignocellulose in corn stover based on near-infrared reflectance spectroscopy and chemometrics methods. Biores Technol 321:124449. https://doi.org/10.1016/j.biortech.2020.124449
27. Bai Z, Hu X, Tian J et al (2020) Rapid and nondestructive detection of sorghum adulteration using optimization algorithms and hyperspectral imaging. Food Chem 331:127290. https://doi.org/10.1016/j.foodchem.2020.127290
28. Huang H, Hu X, Tian J et al (2021) Rapid and nondestructive prediction of amylose and amylopectin contents in sorghum based on hyperspectral imaging. Food Chem 359:129954. https://doi.org/10.1016/j.foodchem.2021.129954
29. Caporaso N, Whitworth MB, Fisk ID (2018) Near-Infrared spectroscopy and hyperspectral imaging for non-destructive quality assessment of cereal grains. Appl Spectrosc Rev 53(8):667–687. https://doi.org/10.1080/05704928.2018.1425214
30. Wang C, Wu XH, Li LQ et al (2018) Convolutional neural network application in prediction of soil moisture content. Spectrosc Spect Anal 38(1):36–41. https://doi.org/10.3964/j.issn.1000-0593(2018)01-0036-06
31. Georganos S, Grippa T, Niang Gadiaga A et al (2021) Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. Geocarto Int 36(2):121–136. https://doi.org/10.1080/10106049.2019.1595177
32. Dhanaraj RK, Ramakrishnan V, Poongodi M et al (2021) Random forest bagging and x-means clustered antipattern detection from sql query log for accessing secure mobile data. Wirel Commun Mob Comput 2021:1–9. https://doi.org/10.1155/2021/2730246
33. Sheykhmousa M, Mahdianpari M, Ghanbari H et al (2020) Support vector machine versus random forest for remote sensing image classification: a meta-analysis and systematic review. IEEE J Select Top Appl Earth Observ Remote Sens 13:6308–6325. https://doi.org/10.1109/JSTARS.2020.3026724
34. Gano B, Dembele JSB, Ndour A et al (2021) Using uav borne, multi-spectral imaging for the field phenotyping of shoot biomass, leaf area index and height of West African sorghum varieties under two contrasted water conditions. Agronomy 11(5):850. https://doi.org/10.3390/agronomy11050850
35. Wang K, Guo P, Luo AL (2017) A new automated spectral feature extraction method and its application in spectral classification and defective spectra recovery. Mon Not R Astron Soc 465(4):4311–4324. https://doi.org/10.1093/mnras/stw2894
36. Indahl UG, Naes T (1998) Evaluation of alternative spectral feature extraction methods of textural images for multivariate modelling. J Chemometr J Chemometr Soc 12(4):261–278. https://doi.org/10.1002/(SICI)1099-128X(199807/08)12:4%3c261::AID-CEM513%3e3.0.CO;2-Z
37. Sun L, Zhao G, Zheng Y et al (2022) Spectral–spatial feature tokenization transformer for hyperspectral image classification. IEEE Trans Geosci Remote Sens 60:1–14. https://doi.org/10.1109/TGRS.2022.3144158
38. Wang N, Yao D, Ma L et al (2021) Multi-site clustering and nested feature extraction for identifying autism spectrum disorder with resting-state fMRI. Med Image Anal 75:102279. https://doi.org/10.1016/j.media.2021.102279
39. Sun Y, Liu B, Yu X et al (2021) Perceiving spectral variation: unsupervised spectrum motion feature learning for hyperspectral image classification. IEEE Trans Geosci Remote Sens 60:1–17. https://doi.org/10.1109/TGRS.2022.3221534
40. Cheng Z, Zhang LQ et al (2010) Successive projections algorithm and its application to selecting the wheat near infrared spectral variables. Spectrosc Spectr Anal 30(4):949–952. https://doi.org/10.3964/j.issn.1000-0593(2010)04-0949-04

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.