Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Vis-NIR hyperspectral imaging for sorghum variety discrimination: a systematic evaluation of preprocessing, feature selection, and classification models

Zhen Dong
Jilin Business and Technology College
Wen Xu
1034338581@qq.com
Jilin Business and Technology College
Yu Sun
Jilin Business and Technology College
Liang Zhang
Jilin Business and Technology College
Xiaoming Sun
Jilin Business and Technology College
Haijuan Tian
Jilin Business and Technology College
Fujian Xu
Jilin Business and Technology College

Research Article

Keywords: Sorghum variety discrimination, Vis-NIR hyperspectral imaging, Savitzky–Golay first derivative, Linear discriminant analysis, Competitive adaptive reweighted sampling

Posted Date: August 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10464423/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Vis-NIR hyperspectral imaging for sorghum variety discrimination: a systematic evaluation of preprocessing, feature selection, and classification models

Zhen Dong · Wen Xu* · Yu Sun · Liang Zhang · Xiaoming Sun · Haijuan Tian · Fujian Xu

(School of Grain Engineering and Nutritional Science, Jilin Business and Technology College, Changchun 130507, China)

* Corresponding author: Wen Xu (1034338581@qq.com)

Abstract

Rapid and non-destructive identification of sorghum varieties is essential for quality control and traceability in the grain supply chain. Hyperspectral imaging technology offers a viable means for sorghum variety classification. However, a systematic comparison of preprocessing, feature selection, and classification strategies under a standardized experimental framework has not yet been conducted. This study applied visible–near-infrared (Vis-NIR) hyperspectral imaging (374–1010 nm, 300 bands) to classify nine sorghum varieties from 1,800 spectra. We evaluated five preprocessing strategies plus the untreated raw spectral baseline, five feature representations, and five classifiers under strict data-leakage controls, 95% Wilson confidence intervals, and McNemar's test. The feature representations comprised the full spectrum, principal component analysis (PCA), successive projections algorithm (SPA), competitive adaptive reweighted sampling (CARS), and random forest variable importance with 95% cumulative threshold (RF-95). The classifiers included partial least squares discriminant analysis (PLS-DA), linear discriminant analysis (LDA), radial basis function support vector machine (RBF-SVM), random forest (RF), and a one-dimensional convolutional neural network (1D-CNN). Linear discriminant analysis with Savitzky–Golay first derivative preprocessing (SG+1stDer) on full spectra achieved the highest overall accuracy (OA) of 90.28% (95% confidence interval [CI]: 86.78–92.93%). CARS closely matched full-spectrum performance (89.44% for LDA) while retaining only 83% of spectral bands. The 1D-CNN performed poorly (OA ≤ 46%), indicating that deep learning provides limited benefit on moderate-scale spectral datasets without extensive data augmentation. These findings establish SG+1stDer + LDA as a robust benchmark for sorghum variety discrimination and show that well-regularized classical classifiers remain the practical choice for routine Vis-NIR grain screening.

Keywords Sorghum variety discrimination · Vis-NIR hyperspectral imaging · Savitzky–Golay first derivative · Linear discriminant analysis · Competitive adaptive reweighted sampling

1. Introduction

Sorghum (*Sorghum bicolor* (L.) Moench) is the fifth most important cereal grain globally, serving as a staple food, livestock feed, and increasingly as a feedstock for biofuel production (Bollam et al. 2021). China is among the major sorghum producers, cultivating diverse varieties for grain sorghum for food and brewing, forage sorghum for animal feed, and sweet sorghum for bioethanol (Ramalingam et al. 2021). The morphological and physicochemical differences among varieties directly influence their processing quality, nutritional value, and market price (Adebo

and Kesa 2023). Rapid and reliable variety discrimination is therefore essential for breed authentication, quality control along the supply chain, and prevention of economic fraud through mislabeling of premium varieties.

Conventional variety identification relies on morphological inspection, chemical analysis, or DNA-based markers (Elmasry et al. 2019). Morphological assessment requires trained personnel and is inherently subjective; chemical assays are destructive or time-consuming; and DNA markers remain cost-prohibitive for routine screening of large sample batches (Feng et al. 2019; Liang et al. 2025). Vis-NIR hyperspectral imaging (HSI) has emerged as a non-destructive alternative that simultaneously captures spectral and spatial information across hundreds of narrow wavelength bands (Lu et al. 2020; Patel et al. 2024). Vis-NIR HSI probes the absorption signatures of C-H, O-H, and N-H bond overtones and combination bands (Beć et al. 2022), which are sensitive to the biochemical composition of grain kernels including starch, protein, moisture, and phenolic compounds (Said et al. 2025). Vis-NIR HSI has been extensively applied to grain variety discrimination across multiple cereal species. For maize seeds, hybrid deep learning architectures combining convolutional and recurrent networks have achieved high classification accuracy across multiple varieties (Zhu et al. 2025). Convolutional neural networks applied to hyperspectral data have proven effective for rice seed variety classification (Wang et al. 2025). A dual-modality deep fusion framework integrating spectral and spatial features from Vis-NIR HSI has been demonstrated for soybean seed classification (Zhao et al. 2026). For wheat grain classification, concatenating Vis-NIR and short-wave infrared spectral ranges enhanced variety-level discriminative performance for both single kernels and bulk samples (Özdoğan and Gowen 2025). Recent critical reviews of HSI data analysis strategies have documented advances in spectral preprocessing, wavelength selection, and classifier optimization, providing a methodological framework for grain variety discrimination (Medina–García et al. 2025; Rogers et al. 2023).

Spectral preprocessing remains a critical methodological component, with Savitzky–Golay (SG) smoothing, standard normal variate (SNV), multiplicative scatter correction (MSC), and derivative methods forming the standard toolkit that underpins modern grain classification pipelines (Rinnan et al. 2009). Feature selection and wavelength optimization substantially influence classification performance in Vis-NIR spectroscopy by removing redundant or uninformative spectral variables. PCA is widely employed for unsupervised dimensionality reduction, transforming high-dimensional spectral data into a compact set of orthogonal components that capture the dominant variance structure (Pan et al. 2026). Among supervised approaches, SPA selects a minimally redundant wavelength subset (Qin et al. 2023). CARS uses Monte Carlo sampling with PLS regression coefficients to adaptively retain informative variables (Li and Yang 2023). RF ranks spectral variables by their mean decrease in impurity across an ensemble of decision trees (Wang et al. 2021). For classification, PLS-DA remains the most widely adopted chemometric baseline (Anderson. and Walsh. 2022). LDA with shrinkage regularization has been applied to near-infrared classification of peanut kernel origin and variety (Zhu et al. 2024). RBF-SVM have demonstrated strong performance for spectral classification tasks where non-linear class boundaries are present (Jiang et al. 2021). RF provides an ensemble alternative, ranking spectral variables by their mean decrease in impurity across an ensemble of decision trees (Wang et al. 2021). Among deep learning approaches, 1D-CNN have been applied to hyperspectral imaging for wheat seed variety classification (Zhao et al. 2022). Collectively, these five classifiers span the

methodological spectrum from classical chemometrics to deep learning, yet their comparative performance has not been systematically evaluated for sorghum grain variety discrimination.

Despite substantial progress, three methodological gaps persist in the sorghum variety discrimination literature. First, no published study has systematically compared multiple preprocessing methods, feature selection strategies, and classification models within a unified experimental framework for sorghum grain classification. Previous studies typically evaluate a single preprocessing method and a single classifier, making cross-study comparisons unreliable because spectral acquisition protocols, sample preparation, and evaluation metrics vary across investigations. Second, most sorghum HSI studies rely exclusively on overall accuracy for model comparison, without confidence intervals or statistical significance tests such as McNemar's test, limiting the rigor of reported findings. Third, the applicability of deep learning to moderate-scale Vis-NIR spectral datasets—typically fewer than 2,000 samples—has not been critically evaluated in the sorghum domain.

The objectives of this study were: (1) to evaluate six preprocessing strategies—Raw, SG, SNV, MSC, SG+1stDer, and SG+SNV—and identify the optimal approach for sorghum Vis-NIR spectra; (2) to compare five feature representations (full spectrum, PCA, SPA, CARS, RF-95) and five classifiers (PLS-DA, LDA, RBF-SVM, RF, and 1D-CNN) within a unified experimental framework; (3) to quantify model differences using 95% confidence intervals and McNemar's test; and (4) to assess the feasibility of 1D-CNN for spectral classification on a moderate-scale dataset (n = 1,440 training samples), establishing a benchmark for deep learning studies in grain variety discrimination.

2. Materials and Methods

2.1 Sample Preparation and Hyperspectral Data Acquisition

Nine sorghum varieties were obtained from Jilin Academy of Agricultural Sciences, China: JL6 (Jiliang 6), JN1 (Jiniang 1), JZ12 (Jinza 12), JZ124 (Jiza 124), JZ320 (Jiza 320), JZ323 (Jiza 323), JZ330 (Jiza 330), JZ341 (Jiza 341), and KS (Kangsi). For each variety, 200 intact, undamaged seeds were selected, yielding a total of 1,800 samples. All samples were equilibrated under controlled conditions (22 °C, 45% relative humidity) in darkness for 24 h prior to spectral acquisition.

Hyperspectral images were acquired using a Vis-NIR hyperspectral imaging system (Hangzhou Hyperspectral Imaging Technology Co., Ltd.) comprising a tungsten lamp light source, lens, camera, and computer. The spectral acquisition range spanned 374–1010 nm with a spectral resolution better than 2.8 nm, a sampling interval of 2.2 nm, and 300 spectral bands. For each measurement session, 200 grains per variety were evenly placed on the sample stage and each sample set was scanned in triplicate (Dong et al. 2026). The hyperspectral data acquisition system and workflow are illustrated in Fig. 1.

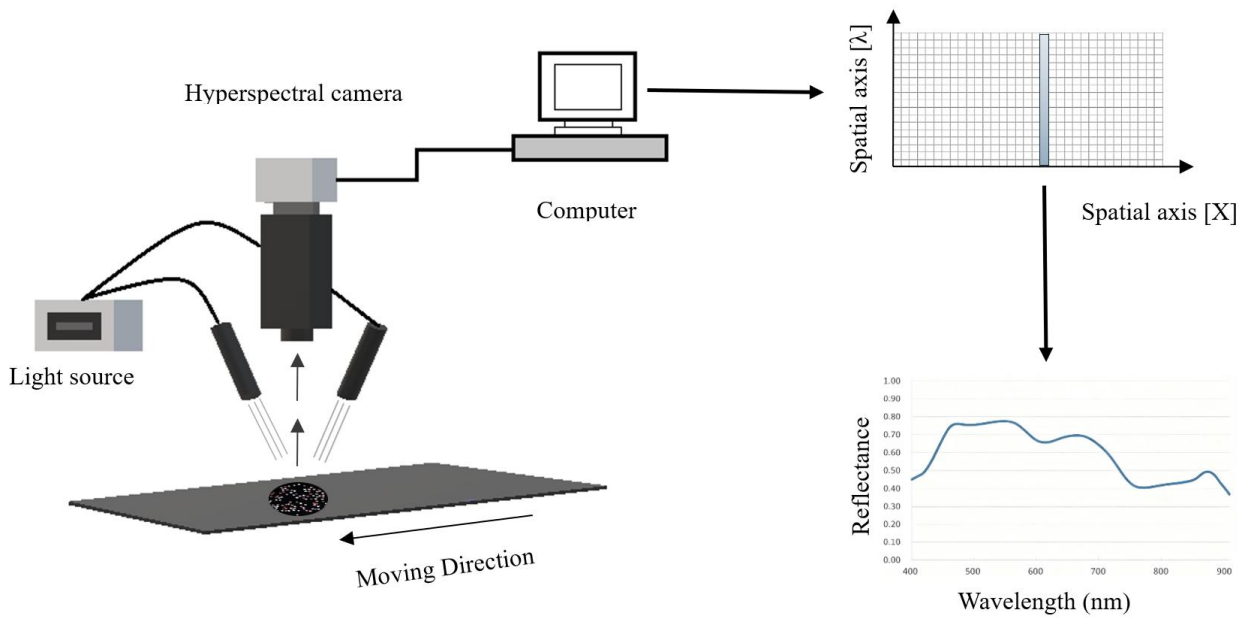

Fig. 1 Schematic drawing of the hyperspectral data acquisition system and associated workflow

2.2 Region of Interest (ROI) Extraction and Spectral Correction

ENVI® v5.3 software (Exelis Visual Information Solutions, Inc., USA) was used to extract regions of interest (ROIs) from the acquired images. The intact sorghum grains were selected as the ROI. The spectral reflectance of all pixels within each ROI was extracted, and their average was calculated as the representative spectrum for each sample. To reduce noise originating from detector sensitivity discrepancies, camera hardware differences, and external environmental interference, a black-and-white board correction was performed on the acquired raw hyperspectral images, following the established calibration procedure (Dong et al. 2026).

2.3 Anomaly Detection and Dataset Partitioning

A two-stage anomaly detection procedure was applied independently to each variety. First, physical threshold screening removed spectra with reflectance values outside the valid range, a standard deviation below 0.001, or a spectral range below 0.01—conditions indicative of detector saturation or failure. Second, PCA-based combined T^2 and Q-statistic outlier detection was performed, with principal components retained to achieve ≥95% cumulative variance (Peila et al. 2022). Hotelling T^2 control limits were calculated using the F-distribution at the 95% confidence level, and Q-residual limits were computed with the Jackson–Mudholkar weighted chi-square approximation. Samples exceeding either limit were flagged as outliers; all 1,800 spectra passed both screening stages, with no anomalies detected. The dataset was randomly split using stratified sampling (random seed = 42) into training (70%, 1,260 spectra), validation (10%, 180 spectra), and test (20%, 360 spectra) sets, with the proportion of each variety preserved across all subsets.

2.4 Unsupervised Exploratory Analysis

To inspect the natural clustering structure of the spectral data, PCA and t-distributed stochastic neighbor embedding (t-SNE) were applied to the full cleaned dataset (1,800 spectra, 300 bands) (Grazina et al. 2020). As unsupervised methods, they used the complete dataset—including the eventual test set—without risk of information leakage. PCA was first performed on the raw reflectance spectra; the first two principal components were used to visualize the clustering tendency of the varieties in spectral space. Subsequently, t-SNE was applied as a non-linear complement to PCA with a perplexity of 30 and a random seed of 42 for reproducibility, running for 1,000 iterations with a random initialization. The resulting two-dimensional embedding was color-coded by variety to reveal local neighborhood structure and overlaps that PCA, as a linear method, may not fully resolve. Both methods served exclusively for exploratory visualization; no reduced features from this step were used in subsequent classification.

2.5 Spectral Preprocessing and Optimization

Five preprocessing strategies and the untreated raw spectra were compared to address physical scattering effects, baseline drift, and random noise inherent in diffuse reflectance measurements: (1) raw spectra without preprocessing (Raw, baseline); (2) SG smoothing (window size 7, second-order polynomial); (3) SNV for individual spectrum centering and scaling; (4) MSC for additive and multiplicative scattering effects; (5) SG smoothing followed by first-order derivative (SG+1stDer), using the same SG parameters; and (6) SG smoothing followed by SNV (SG+SNV).

The optimal strategy was identified through a systematic validation procedure. For each method, a PLS-DA model was trained on the independent training set (n = 1,260) and evaluated on the held-out validation set (n = 180, stratified by variety). The number of latent variables was capped at 20, constrained by the minimum of the number of training samples and spectral variables. The method yielding the highest validation accuracy was selected as optimal. When the accuracy difference between two or more methods was ≤0.5 percentage points, a threshold defining engineering equivalence, the method with the higher signal-to-noise ratio (SNR) was preferred. The selected strategy was subsequently applied to the combined training and validation sets and to the independent test set for all downstream modeling. All global statistics, including the MSC reference spectrum, were derived exclusively from the original training set to prevent data leakage.

2.6 Feature Dimensionality Reduction and Wavelength Selection

Five feature representations were constructed and compared. The full spectrum (Full) comprised all 300 wavelength bands. For PCA-based reduction, PCA was applied to the preprocessed spectra; components retaining ≥95% cumulative variance were selected, yielding 15 principal components explaining 95.46% of the total variance. SPA selected 30 wavelength bands with minimal redundancy by iteratively identifying variables that are least collinear with those already chosen, beginning with the wavelength most correlated with the PLS score vector. CARS was performed across 50 Monte Carlo runs, each drawing 80% of the training samples at random. A PLS model was fitted in each run, and the absolute regression coefficients served as wavelength weights. The averaged weights were ranked, and the wavelengths whose cumulative weight reached 95% were retained, yielding 249 bands. For RF-95, a random forest model with 500 trees was trained under five-fold stratified cross-validation on the preprocessed

spectra. Variable importance was ranked by the mean decrease in Gini impurity averaged across folds, and the top wavelengths accounting for 95% of the cumulative importance were retained, subject to an upper limit of 120 bands.

2.7 Classification Models

Five classifiers were benchmarked: PLS-DA, LDA, RBF-SVM, RF, and 1D-CNN. All models were implemented in Python using scikit-learn (version 1.3.0) and TensorFlow 2.x with a fixed random seed of 42. RBF-SVM required z-score standardization prior to training. Although LDA is scale-invariant, it was standardized for pipeline consistency, while PLS-DA, RF, and the 1D-CNN used unstandardized preprocessed spectra. Standardization parameters were fitted on the training set and refit on the augmented training set only during the final retraining stage.

PLS-DA was implemented as a PLS2 regression model using one-hot encoded class labels. The number of latent variables was set to the minimum of 20 (to mitigate overfitting), the number of spectral variables, or one less than the number of training samples. Each test sample was assigned to the class with the highest predicted response value. LDA was applied assuming multivariate normality and equal covariance matrices. For full-spectrum input (300 features), the within-class covariance matrix becomes singular; shrinkage estimation via the Ledoit–Wolf estimator was therefore employed using scikit-learn's shrinkage="auto" with the "lsqr" solver. RBF-SVM was configured with C = 10 and gamma = "scale", corresponding to $(n_features \cdot Var(X))^{-1}$, with class_weight="balanced" and probability estimates via Platt scaling. RF consisted of 500 decision trees with $\sqrt{n_features}$ features per split, class_weight="balanced", and remaining parameters at scikit-learn defaults.

The 1D-CNN comprised a convolutional layer (32 filters, kernel size 7, ReLU, "same" padding), global average pooling, dropout (rate 0.2), a hidden dense layer (64 units, ReLU), and a nine-unit softmax output. The model was compiled with Adam (learning rate 0.001) and sparse categorical cross-entropy loss, with class weighting applied via sample_weight in model.fit() where $w_c = N_total / (n_classes \cdot N_c)$. Online augmentation included Gaussian noise ($\sigma = 0.01 \times \sigma_batch$, where σ_batch denotes the per-batch spectral standard deviation) and random baseline scaling within [0.95, 1.05]. Augmentation was applied during five-fold stratified cross-validation to determine the optimal epoch count (patience = 20). The final model was retrained on the full augmented training set for the optimal epoch count without early stopping. Batch size was 32; all training was CPU-based.

2.8 Evaluation Metrics and Statistical Analysis

Model performance was assessed using multiple complementary metrics. The OA was defined as the proportion of correctly classified samples in the test set. Wilson 95% confidence intervals were computed for OA to quantify estimation uncertainty. Weighted precision, weighted recall, and weighted F1-score were calculated with class weights proportional to class support in the test set. Macro F1-score was computed as the unweighted average of per-class F1-scores, giving equal importance to all varieties regardless of sample size. Macro area under the receiver operating characteristic curve (macro AUC) was calculated using the one-vs-rest strategy. For pairwise classifier comparisons, McNemar's test (Mcnemar 1947) was applied to the contingency table of correct and incorrect predictions between two classifiers on the test set, using a significance threshold of $\alpha = 0.05$.

3. Results and Discussion

3.1 Spectral Characterization and Unsupervised Exploratory Analysis

The mean reflectance spectra of the nine sorghum varieties exhibited broadly consistent overall profiles across the 374–1010 nm Vis-NIR range (Fig. 2). In the visible region, all varieties displayed strong pigment-driven absorption troughs, attributable to chlorophylls, carotenoids, and anthocyanins in the grain pericarp. Clear inter-variety offsets in mean reflectance were evident throughout the NIR range. These offsets are primarily governed by physical light-scattering characteristics of the grain surface and internal structure. Superimposed on these physical effects, overtones and combination bands of O–H, C–H, and N–H functional groups—associated with starch, protein, and moisture content—produce characteristic chemical absorption features. The combined physical and chemical spectral differences enable discrimination of the nine sorghum varieties.

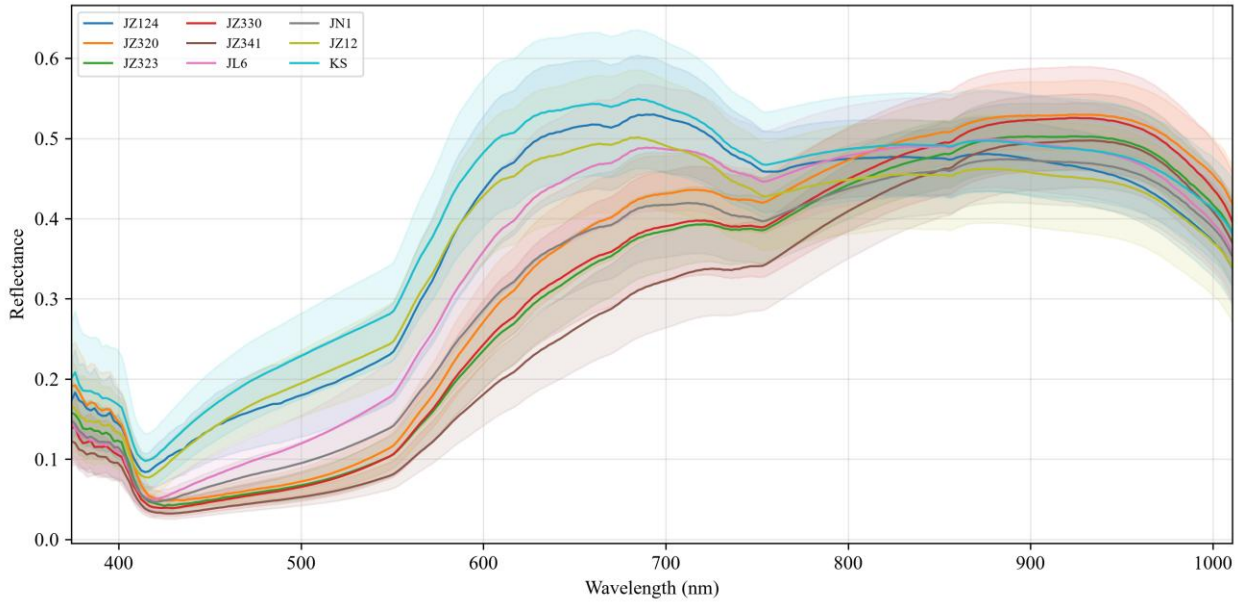

Fig. 2 Mean reflectance spectra of the nine sorghum varieties (±1 standard deviation shown as shaded bands).

Unsupervised dimensionality reduction was applied to inspect the natural cluster structure of the data. The PCA score plot of the first two principal components (PC1 vs. PC2) showed partial separation among varieties (Fig. 3a); these two components collectively accounted for over 96% of the total spectral variance. However, the negative silhouette score (−0.082) and low five-nearest-neighbor cross-validation accuracy (36.7%) confirmed that linear projection alone could not adequately resolve the nine varieties. Some pairs were well separated: Jiza 124 and Jiza 341 achieved near-perfect pairwise discriminability (1-NN accuracy of 99.8%), and Jiza 341 was clearly distinguishable from Kangsi (99.2%). In contrast, other pairs remained near chance level—notably Jiza 124 versus Kangsi (55.0%) and Jinza 12 versus Kangsi (51.0%)—reflecting substantial overlap of their spectral distributions in PCA space. The closest centroid pair, Jiza 323 and Jiza 330 (distance = 0.175), also showed substantial overlap. The t-SNE embedding (perplexity = 30, Fig. 3b) produced tighter within-variety clustering, with a higher five-nearest-neighbor accuracy (49.2%) and a larger inter- to intra-class distance ratio (1.59) compared with PCA (1.54).

Nevertheless, the silhouette score remained negative (−0.068) and even t-SNE could not fully separate all varieties, underscoring the need for supervised classification with optimized preprocessing and feature selection.

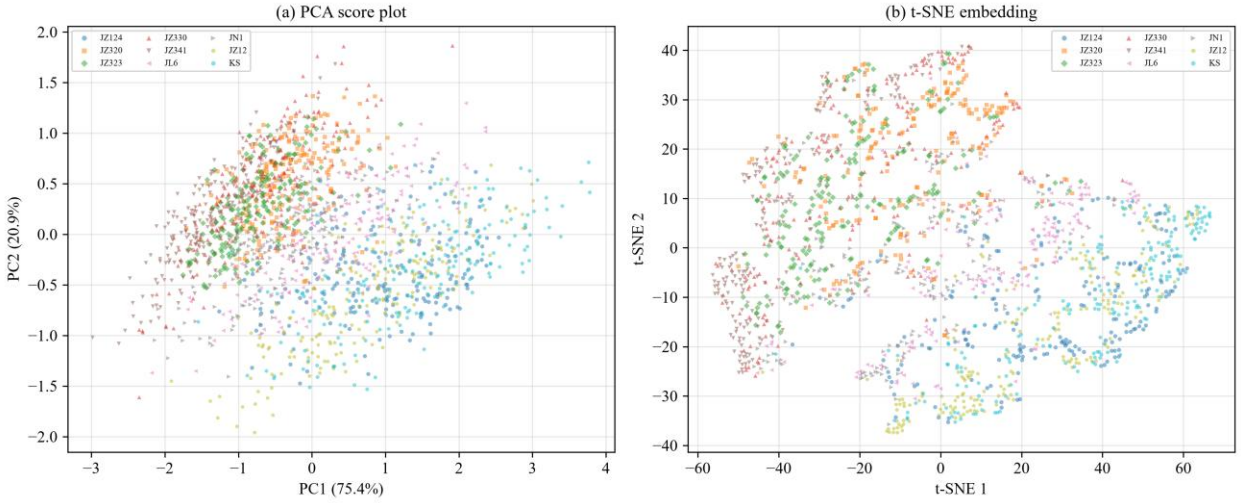

Fig. 3 PCA score plot (a) and t-SNE embedding (b) of the nine sorghum varieties based on the full spectral dataset.

3.2 Preprocessing Optimization

The validation accuracy of PLS-DA models is summarized in Fig. 4. Raw spectra yielded 76.67%, establishing the baseline without spectral correction. SG smoothing alone produced identical accuracy (76.67%), indicating that noise reduction without scatter correction did not improve discriminability. Both scatter correction methods performed substantially better: SNV achieved 82.22% and MSC reached 82.78%, confirming that correcting additive and multiplicative scattering is critical for seed spectral data. The combined method SG+SNV achieved 82.22%, equivalent to SNV alone. SG+1stDer achieved the highest validation accuracy of 85.00%, outperforming all other strategies. The dual effect of SG+1stDer—suppressing high-frequency noise while enhancing spectral resolution through derivative transformation of overlapping absorption bands—likely accounts for its superiority. Based on this validation, SG+1stDer was adopted as the optimal preprocessing method and applied to all subsequent feature selection and classification experiments. Fig. 5 compares representative raw spectra with the optimally preprocessed spectra.

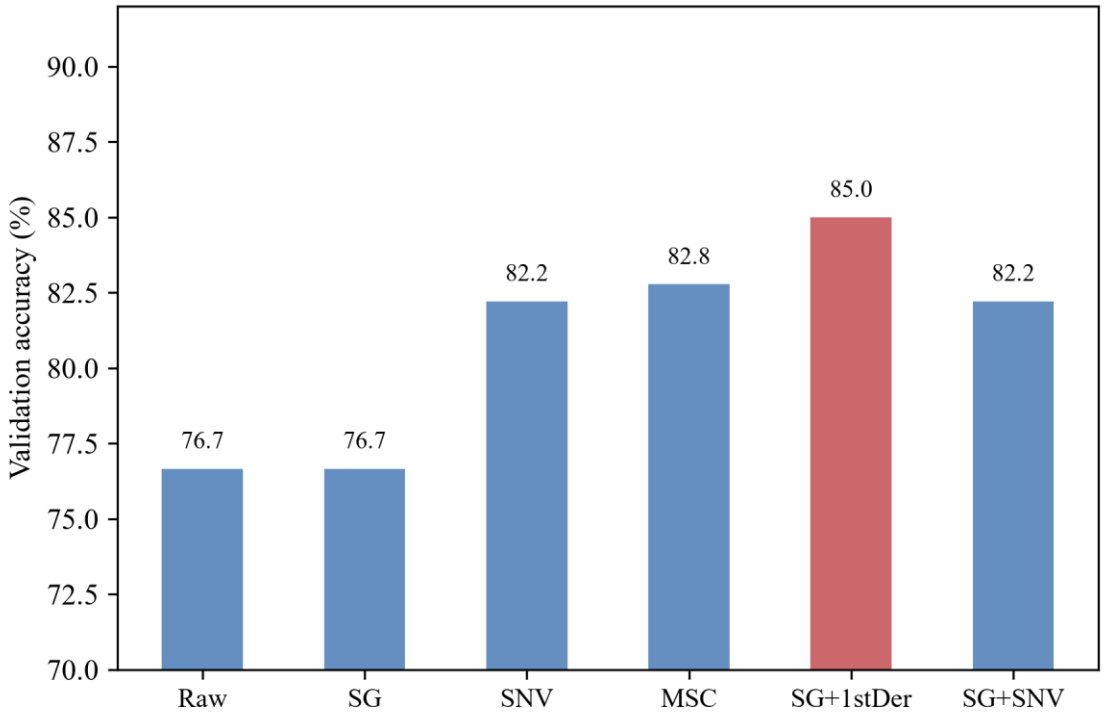

Fig. 4 Validation accuracy of PLS-DA models for five preprocessing methods and the untreated raw spectral.

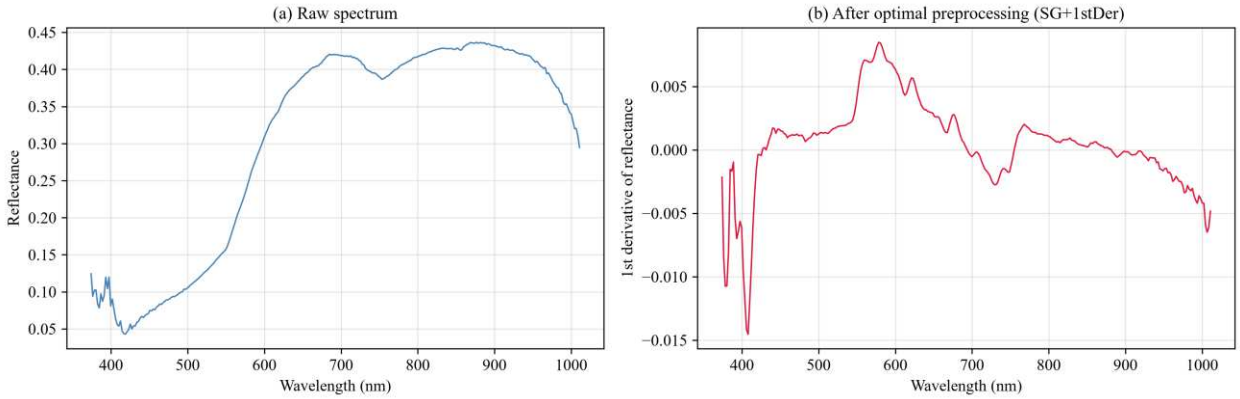

Fig. 5 Comparison of raw spectra (a) and SG+1stDer preprocessed spectra (b) for the nine sorghum varieties.

3.3 Feature Dimensionality Reduction and Wavelength Selection

PCA was applied to the SG+1stDer preprocessed spectra. The scree plot (Fig. 6a) showed that the first three components explained 73.9% of the total variance, with the cumulative variance reaching 95.5% by the 15th component. The first 15 components were retained for classifier evaluation. The PCA loading plot (Fig. 6b) revealed five interpretable spectral features across the first two principal components. PC1 exhibited a prominent positive peak near 420 nm attributable to carotenoid absorption, together with a strong negative loading at approximately 670 nm, corresponding to red-region absorption of residual chlorophyll and phenolic compounds in the grain coats. PC2 was dominated by the strongest feature in the entire loadings profile, a sharp positive peak at approximately 406 nm

also ascribed to carotenoids, paired with a secondary negative trough near 580 nm. This trough showed spectral responses opposite to the positive green peak loading of PC1 in the same wavelength region. This positive PC1 feature at 560 nm, identified as the green spectral reflectance peak, reflects the substantial seed coat colour variation among sorghum varieties and is captured as the dominant positive loading in the first principal component. In the near-infrared region, both PCs showed loading features near 730 nm, associated with O–H overtone absorption of water, and a weaker feature at approximately 980 nm, attributable to O–H overtones linked to starch and moisture content. These wavelength-specific loading patterns confirm that the spectral distinctions among the nine sorghum varieties possess a genuine biochemical basis rooted in pigment content, water status, and storage compound composition.

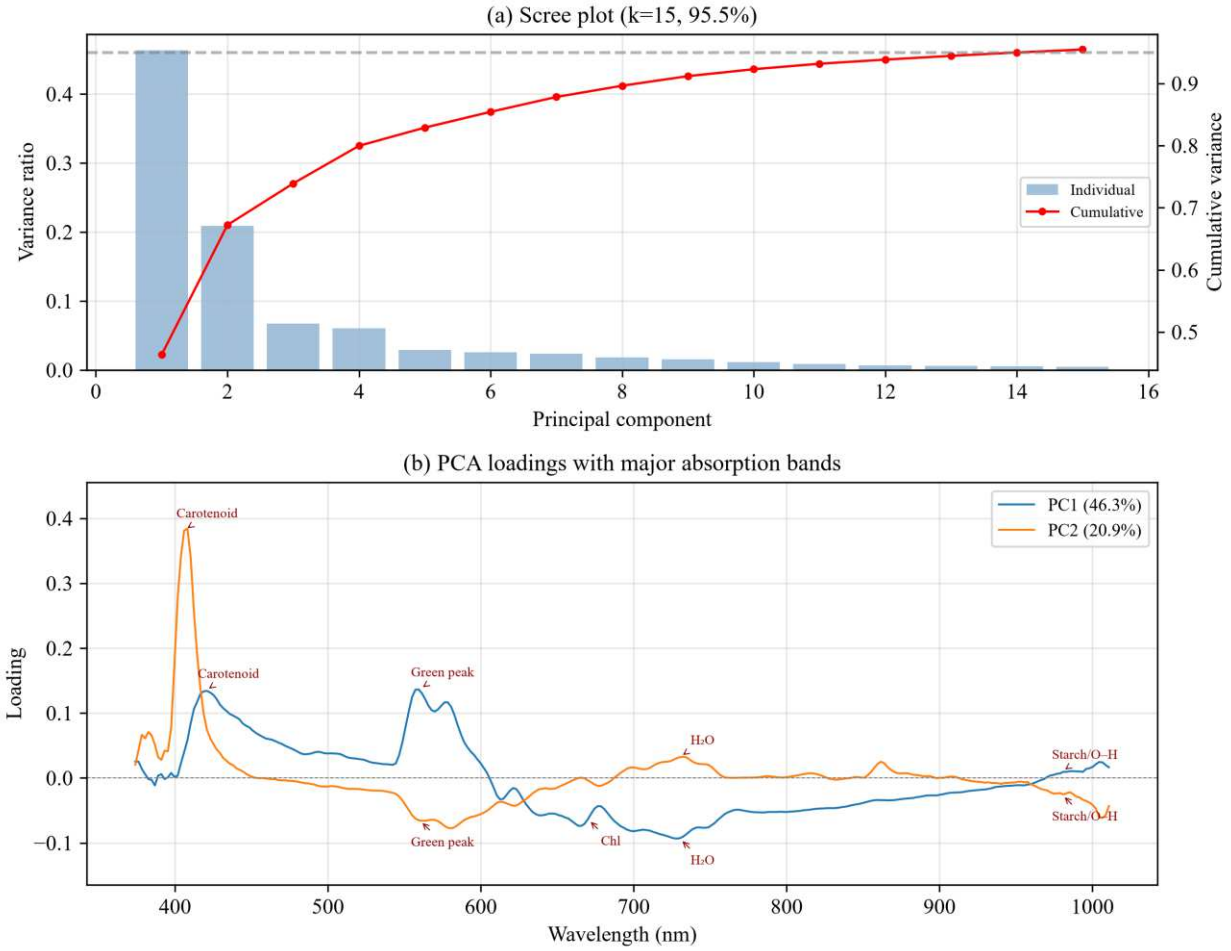

Fig. 6 PCA of the SG+1stDer preprocessed spectra: (a) scree plot together with cumulative variance, and (b) loading plots for the first two principal components.

The wavelength bands selected by SPA, CARS, and RF-95 are illustrated in Fig. 7. SPA selected the fewest bands (30), concentrated predominantly in the ultraviolet and blue spectral region (374–429 nm, 23 of 30 bands). Seven additional bands were scattered across the visible and red-edge regions (559, 602, 662, 679, and 685 nm), with one band at 751 nm and a terminal band at 1007 nm. No bands were selected by SPA in the NIR plateau (800–950 nm). CARS retained the most bands (249), preserving nearly the full spectral range (380–1011 nm). RF-95 selected 120

bands with a more balanced distribution: 56 bands in the visible region (400–680 nm, 47%), 20 bands in the red-edge (680–750 nm, 17%), and 28 bands spanning the NIR (780–932 nm), with six additional bands near 964–975 nm. Substantial overlap was observed between the CARS and RF-95 selections (95 shared bands), and both retained at least 40% of the full spectral channels.

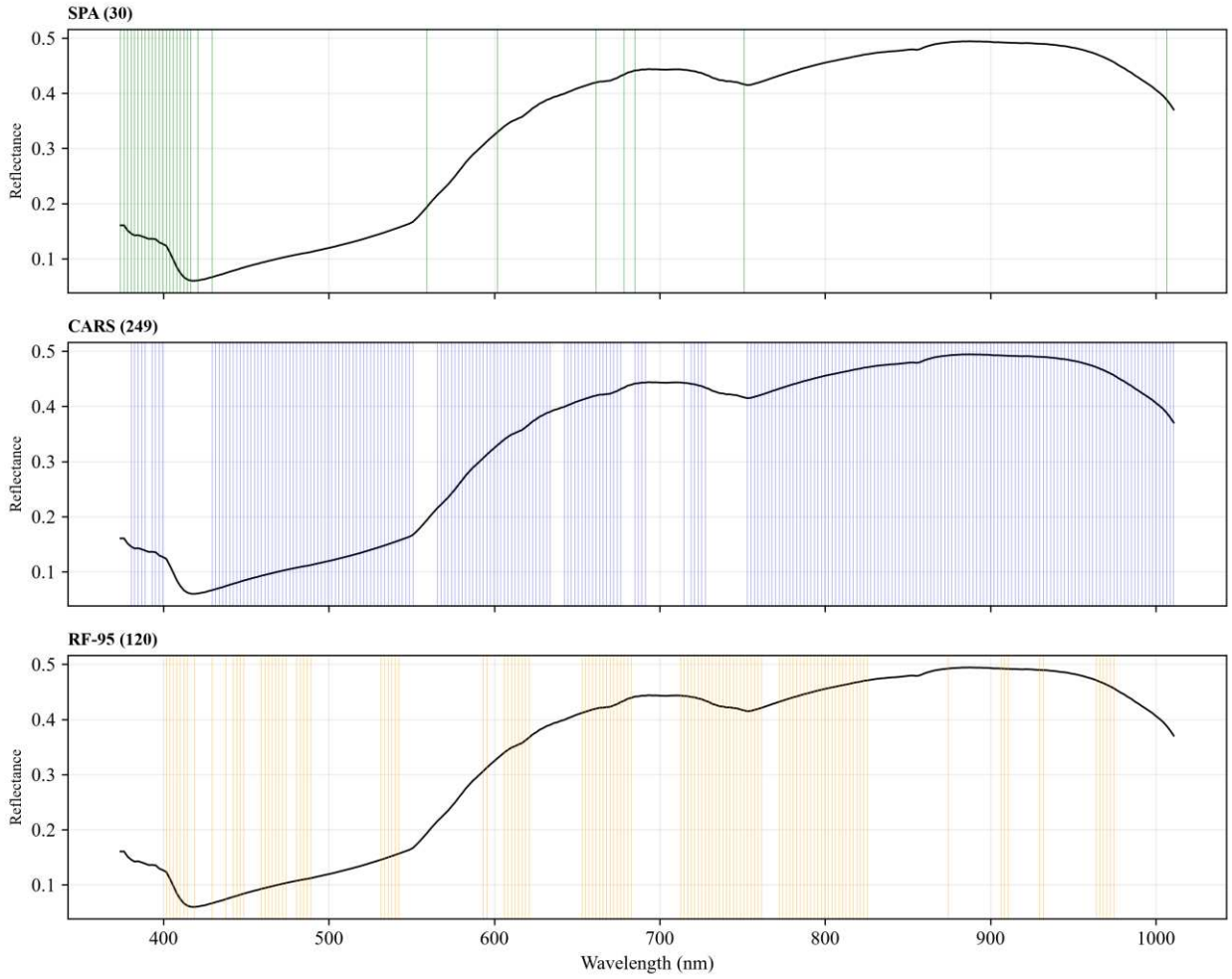

Fig. 7 Wavelength bands selected by SPA, CARS, and RF-95 (vertical markers), overlaid on the mean reflectance spectrum.

3.4 Classification Performance Comparison

Table 1 presents the performance metrics for all five classifiers across the five feature representations. The highest overall accuracy was achieved by LDA with full-spectrum SG+1stDer data (OA = 90.28%, 95% CI: 86.78–92.93%, macro AUC = 0.9898), demonstrating that LDA is highly effective for sorghum variety discrimination when combined with appropriate preprocessing. RBF-SVM with full-spectrum data was a close second (OA = 89.17%, macro AUC = 0.9868), with its confidence interval ([85.53–91.97%]) substantially overlapping that of LDA, indicating comparable performance. PLS-DA achieved 82.50% and RF achieved 81.67% with full-spectrum data, both significantly lower than LDA and RBF-SVM (McNemar's test, $p < 0.01$ for all pairwise comparisons). The 1D-

CNN trained on CPU attained only 42.78% OA with full-spectrum input, substantially below all classical models and consistent with the expectation that deep learning architectures require larger training sets or GPU-accelerated training to be competitive for small-sample spectral classification tasks.

Table 1 Performance comparison of classifiers across feature subsets.

Model	Features	OA (%)	95% CI	Prec.	Recall	F1 (w)	F1 (m)	AUC (m)
LDA	Full	90.28	[86.78–92.93]	0.9052	0.9028	0.9034	0.9034	0.9898
	CARS	89.44	[85.84–92.21]	0.8970	0.8944	0.8952	0.8952	0.9885
	PCA	67.78	[62.78–72.40]	0.6733	0.6778	0.6746	0.6746	0.9522
	RF-95	84.17	[80.04–87.57]	0.8471	0.8417	0.8422	0.8422	0.9812
	SPA	69.72	[64.79–74.24]	0.6957	0.6972	0.6943	0.6943	0.9522
PLS-DA	Full	82.50	[78.24–86.08]	0.8333	0.8250	0.8226	0.8226	0.9651
	CARS	81.39	[77.04–85.07]	0.8217	0.8139	0.8145	0.8145	0.9642
	PCA	60.83	[55.70–65.74]	0.6227	0.6083	0.5690	0.5690	0.9138
	RF-95	77.22	[72.62–81.25]	0.7780	0.7722	0.7686	0.7686	0.9543
	SPA	61.67	[56.55–66.54]	0.6228	0.6167	0.5963	0.5963	0.9055
RF	Full	81.67	[77.34–85.32]	0.8230	0.8167	0.8167	0.8167	0.9739
	CARS	80.56	[76.15–84.31]	0.8100	0.8056	0.8055	0.8055	0.9760
	PCA	78.06	[73.50–82.02]	0.7901	0.7806	0.7808	0.7808	0.9651
	RF-95	79.44	[74.97–83.30]	0.7974	0.7944	0.7937	0.7937	0.9726
	SPA	73.06	[68.25–77.38]	0.7344	0.7306	0.7287	0.7287	0.9594
RBF-SVM	Full	89.17	[85.53–91.97]	0.8922	0.8917	0.8917	0.8917	0.9868
	CARS	86.94	[83.07–90.04]	0.8715	0.8694	0.8699	0.8699	0.9865
	PCA	79.72	[75.27–83.55]	0.7955	0.7972	0.7953	0.7953	0.9696
	RF-95	85.83	[81.85–89.06]	0.8584	0.8583	0.8580	0.8580	0.9823
	SPA	75.00	[70.28–79.19]	0.7484	0.7500	0.7480	0.7480	0.9590
1D-CNN	Full	42.78	[37.77-47.94]	0.4375	0.4278	0.3990	0.3990	0.8408
	CARS	39.44	[34.53-44.58]	0.3771	0.3944	0.3686	0.3686	0.8158
	PCA	45.28	[40.21-50.44]	0.4421	0.4528	0.4410	0.4410	0.8739

RF-95	44.44	[39.40-49.61]	0.4336	0.4444	0.4335	0.4335	0.8685
SPA	36.67	[31.85-41.76]	0.3353	0.3667	0.3366	0.3366	0.8198

Metrics: CI (Wilson 95% confidence interval), Prec. (weighted precision), Recall (weighted recall), F1 (w) (weighted F1-score), F1 (m) (macro F1-score), AUC (m) (macro AUC).

The overall accuracy heatmap (Fig. 8) reveals several clear patterns in classifier and feature set interactions. First, the full spectrum yielded the highest OA for all four classical classifiers, confirming that the complete SG+1stDer preprocessed spectral profile retains the broadest discriminative information. Among dimensionality-reduced alternatives, CARS-selected features produced results closest to the full spectrum for LDA (89.44% vs. 90.28%), RBF-SVM (86.94% vs. 89.17%), and PLS-DA (81.39% vs. 82.50%), confirming CARS as an effective strategy that preserves most discriminatory wavelengths. Second, PCA and SPA caused substantial accuracy degradation for LDA (67.78% and 69.72%, respectively) and PLS-DA (60.83% and 61.67%). In contrast, RF and RBF-SVM proved more robust to PCA reduction: RF dropped by only 3.61 percentage points (to 78.06%) and RBF-SVM by 9.45 percentage points (to 79.72%). The notable underperformance of PCA for LDA, a 22.50 percentage point drop relative to the full spectrum, suggests that the linear projection directions maximizing global spectral variance are not necessarily optimal for class separation. LDA appears particularly sensitive to information loss in the reduced principal component space.

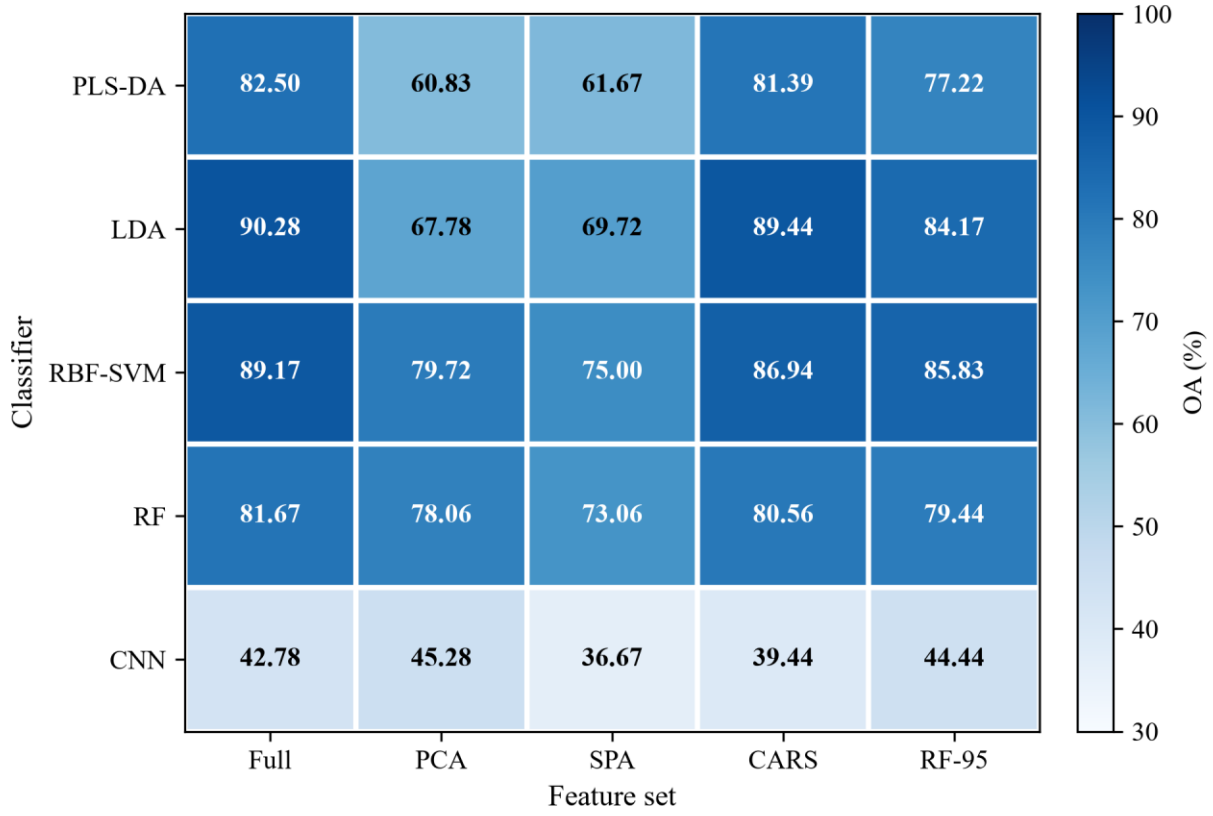

Fig. 8 Heatmap of overall accuracy (OA, %) across all classifier and feature set combinations.

Confusion matrices for LDA and RBF-SVM are shown in Fig. 9. LDA achieved the highest overall accuracy with a concentrated error profile across the nine varieties. Jiza 124 was classified perfectly (100%, 40/40), and Kangsi was identified nearly without error (98%, 39/40). The most notable confusion cluster involved Jiza 341, which had the lowest per-class accuracy (82%, 33/40), with misclassifications distributed primarily to Jiza 330 (10%, 4/40) and Jiza 323 (7.5%, 3/40). This pattern is consistent with the spectral proximity of these varieties observed in the PCA score plot (Fig. 3a) and the t-SNE embedding (Fig. 3b). RBF-SVM exhibited a similar overall confusion pattern but with a slightly broader error distribution. Jiza 330 suffered the largest accuracy decrease relative to LDA (75% vs. 85%), driven by misclassification into five other varieties, while Jinian 1 also performed notably worse under RBF-SVM (80% vs. 88% for LDA), with errors spread across Jiza 320, Jiza 330, and Jiliang 6. These results suggest that, while LDA and RBF-SVM share comparable overall accuracy, the Ledoit–Wolf shrinkage estimator employed by LDA provides slightly cleaner class boundaries, particularly for spectrally similar variety groups.

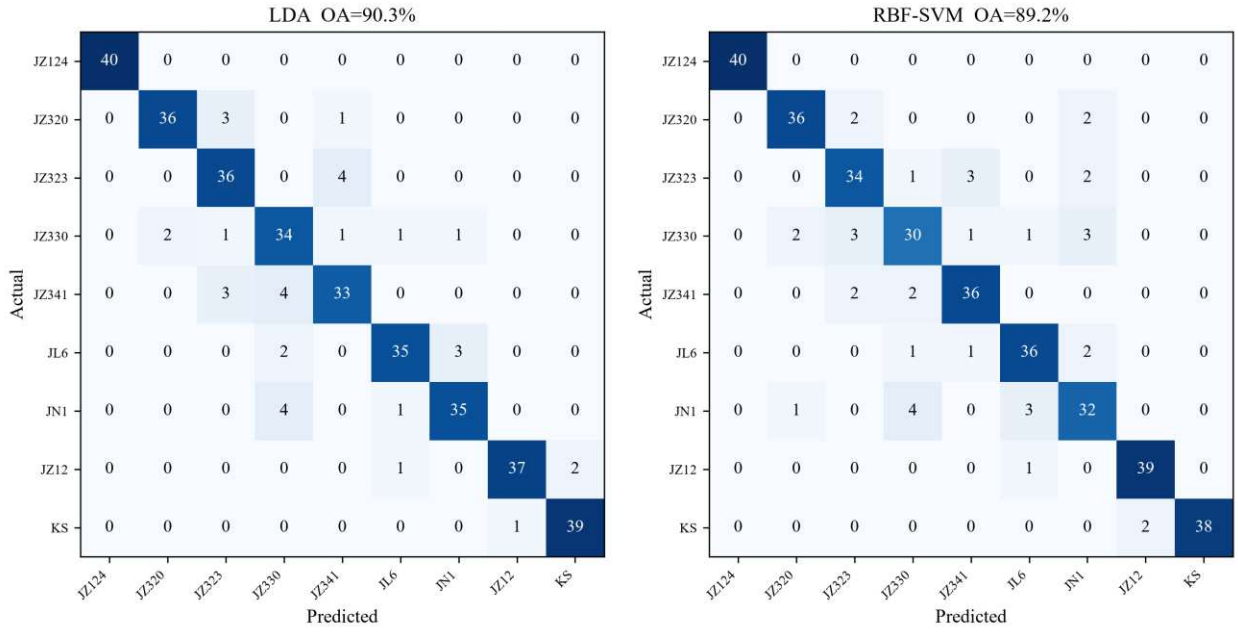

Fig. 9 Confusion matrices for LDA and RBF-SVM with full-spectrum SG+1stDer data.

Confusion matrices for PLS-DA and RF are shown in Fig. 10. Both classifiers exhibited more distributed errors across the nine varieties compared with LDA and RBF-SVM, consistent with their lower overall accuracy (82.50% and 81.67%, respectively). The error patterns were less concentrated within specific variety pairs, reflecting the reduced discriminative power of these models. McNemar's test confirmed that both PLS-DA and RF performed significantly below LDA and RBF-SVM ($p < 0.01$ for all pairwise comparisons).

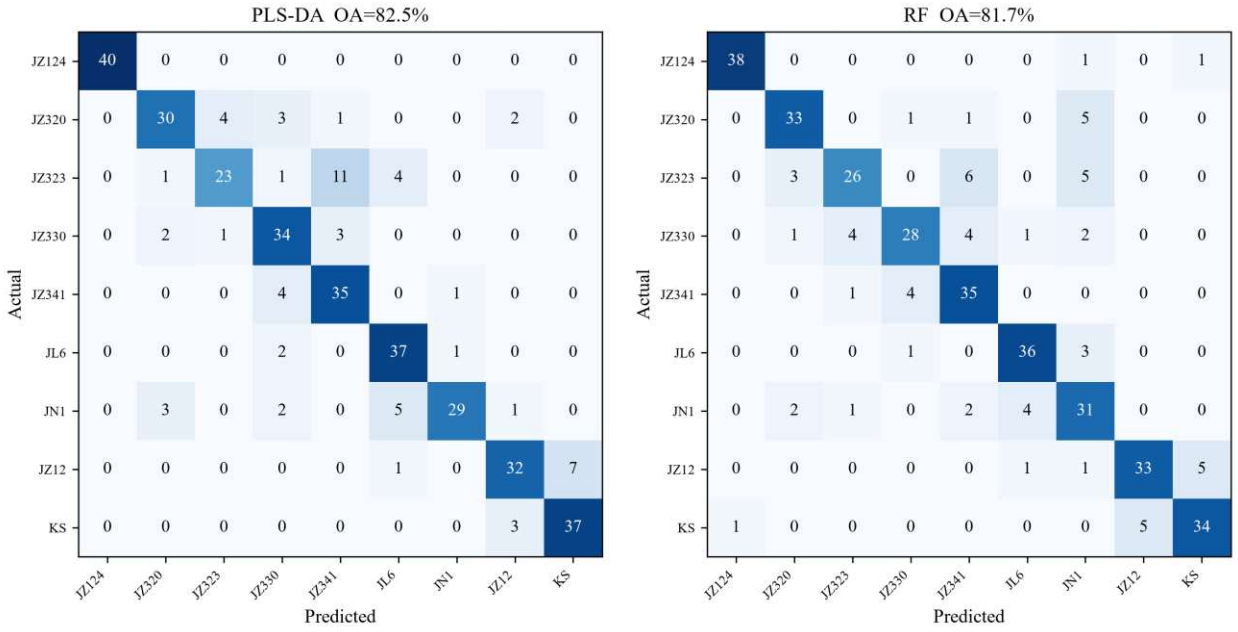

Fig. 10 Confusion matrices for PLS-DA and RF with full-spectrum SG+1stDer data.

The alignment between cross-validation and test accuracy was assessed to detect potential overfitting (Fig. 11). Across the 20 classifier and feature set combinations, the five-fold CV accuracy closely tracked the held-out test accuracy, with a mean absolute deviation of 2.60 percentage points (range: 0.07–7.92 pp). The largest deviations occurred for RF with SPA (7.92 pp) and RBF-SVM with PCA (7.08 pp). In contrast, LDA and PLS-DA produced notably smaller gaps across most feature sets, with overall mean deviations of 0.78 pp and 1.19 pp, respectively. Despite these modest discrepancies in the most aggressively reduced feature spaces, all 20 combinations remained below the 10 pp overfitting threshold, confirming that the models generalized well to unseen data across all preprocessing and feature selection pipelines.

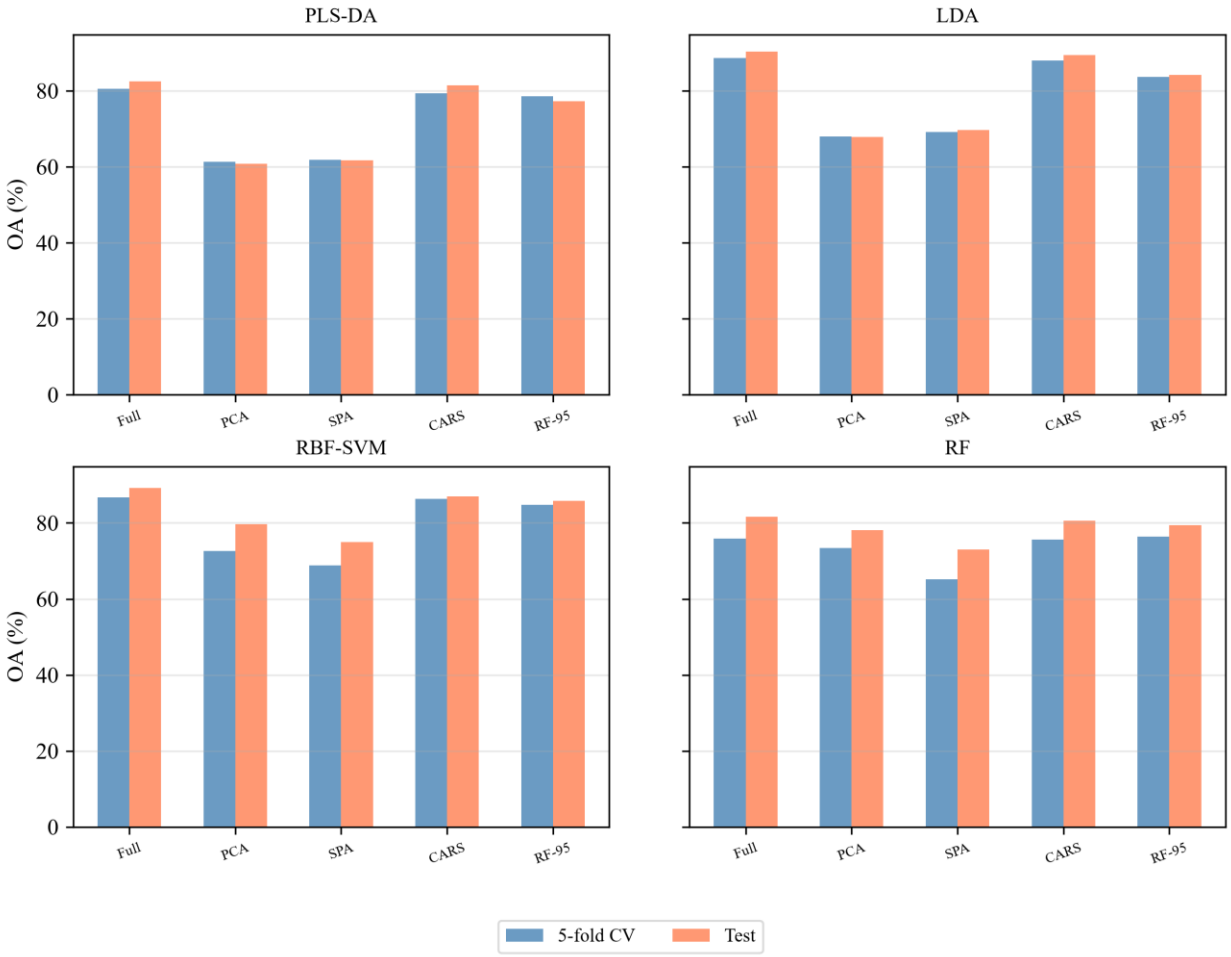

Fig. 11 Comparison of 5-fold cross-validation accuracy and held-out test accuracy for four classifier–feature set combinations.

4. Conclusion

This study presents the systematic comparison of five preprocessing methods plus the untreated raw spectral baseline, five feature representations, and five classifiers for sorghum variety discrimination using Vis-NIR hyperspectral imaging within a unified experimental framework. The central finding is that SG+1stDer preprocessing combined with LDA classification on full spectra achieves 90.28% overall accuracy (Wilson 95% CI: 86.78–92.93%), establishing a robust benchmark for this task. CARS feature selection closely approximates full-spectrum performance (89.44% for LDA) and represents a viable alternative when dimensionality reduction is prioritized. The 1D-CNN's poor performance (OA ≤ 46% across all feature sets, with a best result of 45.28% using PCA features) provides a cautionary result regarding the applicability of deep learning to moderate-scale spectral datasets without extensive training data or GPU-accelerated optimization. These findings have practical implications for rapid, non-destructive sorghum variety screening in grain quality control and breeding programs. Future research may extend the present findings along two directions. First, validating the LDA + SG+1stDer pipeline across multiple growing seasons, diverse geographic origins, and a wider range of sorghum varieties would test the

generalizability of the current benchmark. Second, the suboptimal 1D-CNN performance underscores the need to determine minimum dataset requirements, effective augmentation strategies, and tailored architectures for moderate-scale spectral classification.

Author contribution Zhen Dong: Conceptualization, Methodology, Investigation, Formal Analysis, Writing – Original Draft. Wen Xu: Supervision, Writing – Review & Editing, Funding Acquisition. Yu Sun and Liang Zhang: Investigation, Data Curation. Xiaoming Sun: Methodology, Resources. Haijuan Tian and Fujian Xu: Sample Preparation, Hyperspectral Data Acquisition.

Funding Was provided by The Education Department of Jilin Province (Grant No. JJKH20240282KJ).

Data availability Data supporting the findings of this study are available upon request from the corresponding author.

Declarations

Conflict of interest The authors declare no competing interests.

References

Adebo JA, Kesa H (2023) Evaluation of nutritional and functional properties of anatomical parts of two sorghum (Sorghum bicolor) varieties. Heliyon 9(6). https://doi.org/10.1016/j.heliyon.2023.e17296

Anderson NT, Walsh KB (2022) Review: The evolution of chemometrics coupled with near infrared spectroscopy for fruit quality evaluation. Journal of Near Infrared Spectroscopy 30(2022)(1):3-17. https://doi.org/10.1177/09670335211057235

Beć KB, Grabska J, Huck CW (2022) Miniaturized NIR Spectroscopy in Food Analysis and Quality Control: Promises, Challenges, and Perspectives. Foods 11(10):1465. https://doi.org/10.3390/foods11101465

Bollam S, Romana KK, Rayaprolu L, Vemula A, Das RR, Rathore A et al. (2021) Nitrogen Use Efficiency in Sorghum: Exploring Native Variability for Traits Under Variable N-Regimes. Front Plant Sci 12:643192. https://doi.org/10.3389/fpls.2021.643192

Dong Z, Xu F, Tian H, Hu T, Zhang J, Hu W et al. (2026) Rapid Authenticity Identification and Adulteration Level Detection of Perilla Oil Adulterated with Soybean Oil Based on Hyperspectral Imaging Technology. Food Analytical Methods 19(9):239. https://doi.org/10.1007/s12161-026-03156-y

Dong Z, Yang S, Liu W, Sun X, Xu W, Ren M (2026) The multilayer perceptron recognition model for imperfect maize kernels based on hyperspectral imaging technology. Journal of Food Measurement and Characterization 20(3):2971-2981. https://doi.org/10.1007/s11694-025-03888-1

Elmasry G, Mandour N, Wagner MH, Demilly D, Verdier J, Belin E et al. (2019) Utilization of computer vision and multispectral imaging techniques for classification of cowpea (Vigna unguiculata) seeds. Plant Methods 15(1):24. https://doi.org/10.1186/s13007-019-0411-2

Feng L, Zhu S, Liu F, He Y, Bao Y, Zhang C (2019) Hyperspectral imaging for seed quality and safety inspection: a review. Plant Methods 15(1):91. https://doi.org/10.1186/s13007-019-0476-y

Grazina L, Rodrigues PJ, Igrejas G, Nunes MA, Mafra I, Arlorio M et al. (2020) Machine Learning Approaches Applied to GC-FID Fatty Acid Profiles to Discriminate Wild from Farmed Salmon. Foods 9(11). https://doi.org/10.3390/foods9111622

Jiang H, Yang Y, Shi M (2021) Chemometrics in Tandem with Hyperspectral Imaging for Detecting Authentication of Raw and Cooked Mutton Rolls. Foods 10(9). https://doi.org/10.3390/foods10092127

Li Y, Yang X (2023) Quantitative analysis of near infrared spectroscopic data based on dual-band transformation and competitive adaptive reweighted sampling. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 285:121924. https://doi.org/10.1016/j.saa.2022.121924

Liang Y, Li Z, Shi J, Zhang N, Qin Z, Du L et al. (2025) Advances in Hyperspectral Imaging Technology for Grain Quality and Safety Detection: A Review. Foods 14(17):2977. https://doi.org/10.3390/foods14172977

Lu B, Dao PD, Liu J, He Y, Shang J (2020) Recent Advances of Hyperspectral Imaging Technology and Applications in Agriculture. Remote Sensing 12(16):2659. https://doi.org/10.3390/rs12162659

Mcnemar Q (1947) Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages. Psychometrika 12(2):153-157. https://doi.org/10.1007/BF02295996

Medina–García M, Amigo JM, Martínez-Domingo MA, Valero EM, Jiménez–Carvelo AM (2025) Strategies for analysing hyperspectral imaging data for food quality and safety issues – A critical review of the last 5 years. Microchemical Journal 214:113994. https://doi.org/ 10.1016/j.microc.2025.113994

Özdoğan G, Gowen A (2025) Wheat grain classification using hyperspectral imaging: Concatenating Vis-NIR and SWIR Data for single and bulk grains. Food Control 168:110953. https://doi.org/10.1016/j.foodcont.2024.110953

Pan T, Wang Z, Chen S (2026) A hybrid LD-PCA approach for dimensionality reduction and classification in near-infrared spectroscopy: A case study on green tea. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 358:127864. https://doi.org/10.1016/j.saa.2026.127864

Patel D, Bhise S, Kapdi SS, Bhatt T (2024) Non-destructive hyperspectral imaging technology to assess the quality and safety of food: a review. Food Production, Processing and Nutrition 6(1):69. https://doi.org/10.1186/s43014-024-00246-4

Peila C, Sottemano S, Cesare Marincola F, Stocchero M, Pusceddu NG, Dessì A et al. (2022) NMR Metabonomic Profile of Preterm Human Milk in the First Month of Lactation: From Extreme to Moderate Prematurity. Foods 11(3). https://doi.org/10.3390/foods11030345

Qin Y, Song K, Zhang N, Wang M, Zhang M, Peng B (2023) Robust NIR quantitative model using MIC-SPA variable selection and GA-ELM. Infrared Physics & Technology 128:104534. https://doi.org/10.1016/j.infrared.2022.104534

Ramalingam AP, Mohanavel W, Premnath A, Muthurajan R, Prasad PVV, Perumal R (2021) Large-Scale Non-Targeted Metabolomics Reveals Antioxidant, Nutraceutical and Therapeutic Potentials of Sorghum. Antioxidants (Basel) 10(10). https://doi.org/10.3390/antiox10101511

Rinnan Å, Berg FVD, Engelsen SB (2009) Review of the most common pre-processing techniques for near-infrared spectra. TrAC Trends in Analytical Chemistry 28(10):1201-1222. https://doi.org/10.1016/j.trac.2009.07.007

Rogers M, Blanc-Talon J, Urschler M, Delmas P (2023) Wavelength and texture feature selection for hyperspectral imaging: a systematic literature review. Journal of Food Measurement and Characterization 17(6):6039-6064. https://doi.org/10.1007/s11694-023-02044-x

Said I, Sagan V, Peterson KT, Alifu H, Maiwulanjiang A, Stylianou A et al. (2025) Seed Protein Content Estimation with Bench-Top Hyperspectral Imaging and Attentive Convolutional Neural Network Models. Sensors 25(2):303. https://doi.org/10.3390/s25020303

Wang L, Huang Z, Wang R (2021) Discrimination of cracked soybean seeds by near-infrared spectroscopy and random forest variable selection. Infrared Physics & Technology 115:103731. https://doi.org/10.1016/j.infrared.2021.103731

Wang Z, Wang Z, Zhan Y, Yu S (2025) Deep learning-based rice seed variety classification model with hyperspectral imaging. Journal of Electronic Imaging 34(5):053040. https://doi.org/10.1117/1.JEI.34.5.053040

Zhao J, Du M, Chen X, Ding H, Ma H, Li Q et al. (2026) A Novel Dual-Modality Deep Fusion Framework Based on VIS-NIR Hyperspectral Imaging for Soybean Classification. Journal of Future Foods. https://doi.org/10.1016/j.jfutfo.2025.12.014

Zhao X, Que H, Sun X, Zhu Q, Huang M (2022) Hybrid convolutional network based on hyperspectral imaging for wheat seed varieties classification. Infrared Physics & Technology 125:104270. https://doi.org/10.1016/j.infrared.2022.104270

Zhu Q, Liu Q, Ma D, Zhu Y, Zhang L, Wang A et al. (2025) Maize Seed Variety Classification Based on Hyperspectral Imaging and a CNN-LSTM Learning Framework. Agronomy 15(7):1585. https://doi.org/10.3390/agronomy15071585

Zhu R, Wu X, Wu B, Gao J (2024) High-accuracy classification and origin traceability of peanut kernels based on near-infrared (NIR) spectroscopy using Adaboost - Maximum uncertainty linear discriminant analysis. Current Research in Food Science 8:100766. https://doi.org/10.1016/j.crfs.2024.100766