

Rapid and Non-destructive Evaluation of Soybean Seed Viability Using Transmission Hyperspectral Imaging and Ensemble Learning

Fangfang Liu

Nanjing University of Industry Technology

Pengshuai Niu

Nanjing Agricultural University

Wei Lu

n.jaurobot@163.com

Nanjing Agricultural University

Article

Keywords: soybean vigor, hyperspectral imaging, stacking ensemble learning, transmittance, nondestructive detection

Posted Date: March 30th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9119890/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Rapid and Non-destructive Evaluation of Soybean Seed Viability Using Transmission Hyperspectral Imaging and Ensemble Learning

Fangfang Liu ¹, Pengshuai Niu ² and Wei Lu ^{2,*}

¹ School of computer&software engineering,Nanjing University of Industry Technology, Nanjing 210023, China;
liuff@niit.edu.cn

² College of Smart Agriculture(College of Artificial Intelligence), Nanjing Agricultural University, Nanjing 210031, China;
niupengsh@163.com, njajrobot@163.com

* Correspondence: njajrobot@163.com; Tel.:086-15195816620.

Abstract: Traditional soybean seed viability assessment methods are destructive, time-intensive, and incapable of rapid, non-destructive single-seed grading. To overcome these limitations, this study proposes a novel approach integrating Transmission Hyperspectral Imaging (THSI) with ensemble learning for rapid, non-destructive evaluation. Naturally aged soybean seeds were analyzed using full-spectrum (400–2500 nm) transmittance data to capture deep physiological information. Multiple datasets were constructed by comparing preprocessing techniques—including Smoothing, First Derivative (FD), Hilbert Transform (HT), Savitzky–Golay, Multiplicative Scatter Correction (MSC), and Standard Normal Variate (SNV)—with dimensionality reduction algorithms such as Principal Component Analysis (PCA), Successive Projections Algorithm (SPA), and Competitive Adaptive Reweighted Sampling (CARS). The proposed Stacking ensemble model integrates predictions from Random Forest (RF), Support Vector Machine (SVM), Naïve Bayes (NB), k-Nearest Neighbors (k-NN), and Logistic Regression (LR), achieving 98.33% accuracy and an F1-score of 98.12% on the CARS-HT dataset—significantly outperforming individual classifiers in both accuracy and robustness. Furthermore, the model identified 17 critical wavelengths that reveal physiological mechanisms ranging from chlorophyll degradation and antioxidant balance in the visible spectrum to water migration and lipid peroxidation in the infrared region. The method's high precision and reliability were validated, providing robust technical support for intelligent and precise soybean seed quality management.

Keywords: soybean vigor; hyperspectral imaging; stacking ensemble learning; transmittance; nondestructive detection

0. Introduction

Soybean is a globally important oilseed and protein crop, and its seed quality directly impacts agricultural productivity, food security, and farmers' economic returns. Seed vigor serves as a critical indicator for evaluating seed quality, demonstrating a significant positive correlation with field emergence rates ¹. Internally deteriorated seeds can reduce field emergence rates by 30–50% ² and may serve as vectors for pathogen transmission, posing significant challenges to germplasm resource management and sustainable agricultural development.

Conventional techniques for assessing soybean seed vigor, such as standard germination tests, tetrazolium staining, and Kjeldahl nitrogen analysis, are constrained by their destructive nature, low throughput, and inability to detect early physiological degradation ³. Although artificial accelerated aging (40°C, 95% RH) is widely used to expedite vigor assessment, it alters seed biochemical profiles—such as through aberrant lipid peroxidation and enzyme inactivation—deviating from natural aging processes under typical storage conditions ⁴. Natural aging, influenced by multifactorial environmental interactions, provides a more realistic basis for vigor modeling but requires highly sensitive and robust detection methods.

In recent years, transmission hyperspectral imaging (THSI) has achieved breakthrough progress in agricultural inspection by enabling non-destructive detection of internal biochemical components through multidimensional spectral-spatial fusion^{5,6,7}. In a recent comprehensive review, Liang et al.⁶ highlighted that while reflectance spectroscopy is highly effective at extracting seed coat texture and color features, it exhibits significant limitations in probing biochemical changes within the endosperm that are intrinsically linked to seed vigor. Conversely, the transmission mode or the more penetrative short-wave near-infrared (SWNIR) region can more effectively capture the degradation dynamics of internal storage materials. This makes THSI a critical technological pathway for the precise and non-destructive evaluation of internal seed quality, such as aging severity and hidden pest damage. Furthermore, Zhao et al.⁷ acquired near-infrared spectra of sweet corn seeds under both transmission and reflectance modes. After removing outliers and applying variable selection and preprocessing techniques, they developed a partial least squares (PLS) prediction model, which confirmed that the transmission mode is superior for determining the seed vigor index.

To enhance both detection efficiency and accuracy, this study initially applied preprocessing and dimensionality reduction to the hyperspectral dataset, followed by the construction of base classification models using machine learning algorithms. Although deep learning algorithms possess powerful feature extraction capabilities, they inherently rely on massive annotated datasets⁸ and substantial computational resources⁹. The scarcity of naturally aged seed samples severely constrains the construction of large-scale THSI datasets. In contrast, the Stacking ensemble strategy effectively integrates the respective strengths of multiple weak classifiers^{10,11}. This approach significantly mitigates the risk of overfitting and substantially improves classification robustness, particularly under small-sample conditions.

This study employs THSI technology combined with stacking ensemble learning to achieve rapid, non-destructive, and high-precision detection of seed vigor in naturally aged soybean seeds. Specifically, this study aims to (1) establish the THSI-based seed vigor evaluation system using natural aging models, overcoming the methodological limitations of artificial aging distortion; (2) develop adaptive ensemble learning algorithms for small-sample hyperspectral data to optimize both detection accuracy and computational efficiency; and (3) construct a multi-scale correlation model of "physiological deterioration-spectral response-vigor classification" to provide a new theoretical framework and technical paradigm for intelligent seed quality detection. The research outcomes can provide crucial technical support for germplasm resource bank construction, precision breeding, and the development of intelligent agricultural equipment.

2. Materials and Methods

2.1. Soybean Seed Material

The effect of long-term storage under natural conditions on the vigor of soybean seeds was examined using seeds harvested and naturally aged over different years to ensure ecological relevance and predictive accuracy of the results. Experimental materials were obtained from the seed bank of the College of Agriculture at Nanjing Agricultural University. Three soybean varieties—Gutiandou (from Gutian, Fujian), Diandou No. 1 (from Kunming, Yunnan), and Zhonghuang No. 1 (from the Huang-Huai-Hai region)—were selected with seeds from three years (2021–2023), yielding nine sample groups. Seeds with impurities or detached seed coats were manually removed, and high-quality seeds of uniform size, shape, and plumpness, without mold, were selected for analysis. In total, 100 seeds per variety were selected, resulting in 900 seeds. The soybeans received from the seed bank (maintained at a controlled temperature of 20 °C using air conditioning) were transported in foam boxes and placed into the fresh food compartment of a laboratory refrigerator (4 °C) within 3 h, where they were stored until experimentation.

2.2. Hyperspectral Data Collection

This study employed the GaiaClass series hyperspectral imaging system (Jiangsu Shuangli Hepu Technology Co., Ltd., China) for soybean seed image acquisition. The system was equipped with two Image- λ spectral cameras, enabling simultaneous acquisition of hyperspectral data from two seeds and thereby improving data collection efficiency. It was configured with a 150 W halogen cold light source and customized auxiliary devices: a vacuum pump with a vacuum suction nozzle for precise seed positioning and transfer, and a fiber collimator for optimized illumination uniformity (Figure 1). The spectral acquisition range covered 391.6–2521.6 nm, with a spectral resolution of 2.5 nm in the 391.6–1003.6 nm range and 5.6 nm in the 1003.6–2521.6 nm range. Prior to the experiment, the system was warmed up for 20 min, and optical path focusing and parameter calibration (including resolution and exposure time) were performed to eliminate artifacts such as image distortion or blurring. The optimized camera parameters were determined as follows: focal length of 25 mm, exposure time of 6 ms, and image resolution of 1681×1040 pixels.

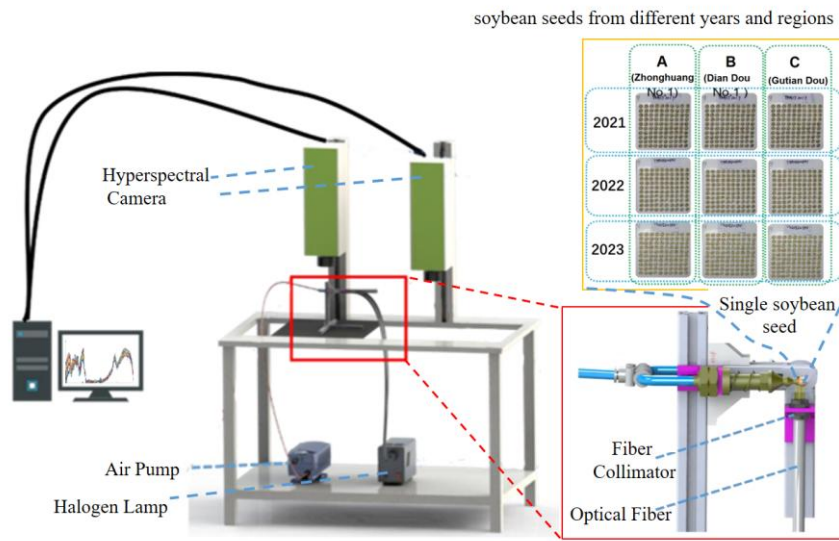


Figure 1. Schematic diagram of the hyperspectral imaging system setup in transmission mode.

Based on the dual-camera configuration, two seeds were selected from the platform and acquired simultaneously for each run. A silicone suction tip was used under negative pressure to transfer seeds of the same variety and batch one by one to preset positions, so that the two seeds were aligned with the fields of view of the two cameras, respectively. Light was projected upward onto the seeds through optical fiber and the fiber collimator; the two spectral cameras captured the transmitted light vertically from above in sync, and the computer recorded sequential one-dimensional images and the corresponding spectral information in real time to obtain the raw hyperspectral image data. To minimize the impact of dark current or lighting conditions, among other factors, black-and-white correction was performed on the hyperspectral images before exporting them from the computer of the hyperspectral system. Equation (1) represents the correction calculation:

$$R = \frac{I - D}{W - D} \quad (1)$$

where I represents the original hyperspectral image, D represents the black frame image (dark reference), W represents the white frame image (white reference), and R represents the corrected spectral image. The corrected images can be used for subsequent spectral processing and analysis.

2.3. Germination Experiment

After collecting the spectral data for the soybean seeds, a 7-day germination assay was conducted in accordance with the regulations of the International Seed Testing Association (ISTA) to evaluate the germination rate, germination index (Gi), and vigor index (Vi). To accommodate the testing capacity, the seeds were divided into 5 batches of 180 seeds each, a total of

900 seeds were tested. The spectral data and germination results were strictly maintained in a one-to-one correspondence. The specific procedures were as follows:

(1) Seed soaking: For each batch, 180 seeds representing three soybean varieties harvested across 3 years were selected for the germination test. From each sample group, 20 seeds were soaked in water at 20°C for 12 h.

(2) Seed placement: The soaked seeds were placed in germination baskets lined with moist germination paper. To maintain consistency between spectral and germination test results, the seeds were arranged in numerical order with 3 cm spacing between each seed to minimize the risk of mold contamination.

(3) Incubation: The germination baskets were placed in an artificial climate chamber (model PQX-250L). The temperature was maintained at 25 ± 1 °C and relative humidity of approximately 80%, optimal for soybean seed germination.

(4) Observation and recording: The germination status of the soybean seeds subjected to various natural aging treatments was recorded twice daily at 08:00 and 19:00. According to the ISTA standard, seeds with a radicle length of 1 cm were classified as germinated. The first 24 h after soaking and placement were recorded as the first day of germination.

(5) Calculation of results: After 7 days, the G_i and V_i of the seeds were calculated and documented. Equation (2) presents the formula for calculating the G_i :

$$G_i = \sum \frac{G_t}{Dt} \quad (2)$$

where G_t represents the number of seeds germinated each day and Dt indicates the corresponding day of germination.

Equation (3) presents the formula for calculating the V_i :

$$V_i = G_i \times S \quad (3)$$

where S represents the root shoot length of the seedlings on day 7 of germination

Based on single-seed germination and seedling morphology observed during the experiment, individual seeds were classified into three vigor categories for single-seed assessment and field screening:

(1) Non-viable: No germination within 7 days.

(2) Low-vigor: Germination within 7 days, but $S < 2$ cm, or exhibiting weak seedling growth (e.g., short/abnormal roots, yellowed or defective cotyledons).

(3) High-vigor: Germination within 7 days with $S \geq 2$ cm and robust seedling growth (well-developed roots, sturdy stem, and intact cotyledons).

2.4. Spectral Data Extraction and Preprocessing

A total of 450 raw hyperspectral image files were acquired, containing all 900 soybean seeds (two seeds per image). To extract valid spectral data, the raw images were first parsed using the Geospatial Data Abstraction Library (GDAL) library in Python and converted into the Environment for Visualizing Images (ENVI) standard format¹². To determine the optimal image segmentation strategy, multiple thresholding methods were compared. Initially, regions of interest (ROIs) were manually delineated using ENVI software to establish a ground truth. Subsequently, using the 650 nm band image (which exhibited the highest clarity) as a baseline, automatic segmentation was performed using global thresholding (fixed threshold $v=127$), adaptive mean thresholding, and Gaussian thresholding (Figure 2A). The results indicated that the adaptive mean thresholding method yielded the best segmentation performance, most closely matching the manually marked ground truth.

Consequently, a binary mask was generated using this method. The mask was then applied to the full-band hyperspectral data cube through a pixel-by-pixel AND operation to accurately extract the target region for individual seeds (Figure 2B). In transmission mode, because the light source converged and entered from the bottom of the seed, combined with the uneven

thickness and heterogeneous internal structure of the seed itself, the transmitted light intensity of the sample exhibited an uneven distribution (brighter in the center and darker at the edges). To eliminate this within-sample spectral variability, the spectral values of all pixels within the ROI of each individual seed were averaged to represent the sample's spectral data for subsequent analysis.

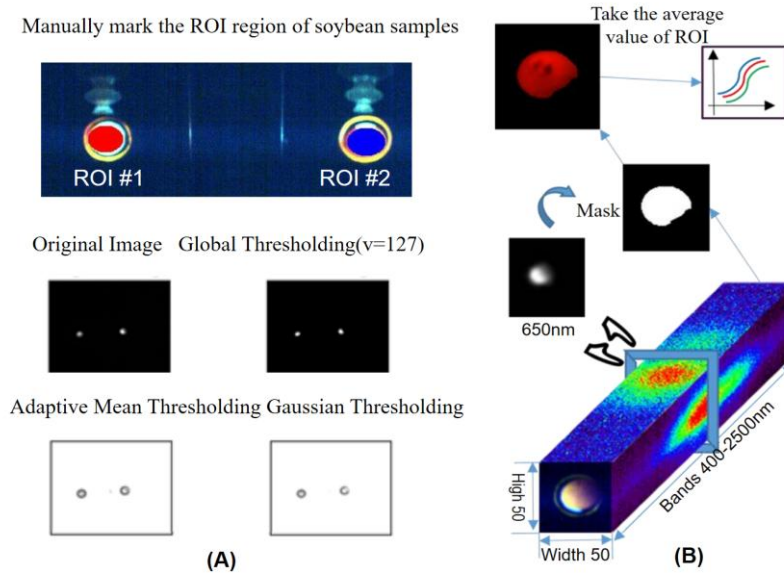


Figure 2. ROI extraction and masking for soybean samples: (A) comparison of thresholding methods (full field, two seeds); (B) masking and spectral extraction for a single seed

2.5. Spectral Data Preprocessing

The raw hyperspectral data were normalized to generate standardized spectral datasets. As depicted in the full-sample normalized spectra (Figure 3A), the soybean seeds exhibited a highly consistent trend across the entire 400–2500 nm range, regardless of their viability, variety, or aging level. To better illustrate the spectral differences among the various sample groups, the average spectrum for each group was calculated and compared (Figure 3B).

The average spectra revealed distinct transmission characteristics primarily in the 500–1500 nm region. Specifically, prominent transmission peaks (corresponding to lower light absorption) were observed near 700 nm and 1100 nm. A pronounced absorption valley (minimum transmittance) appeared around 850 nm, followed by another distinct transmission peak at approximately 1350 nm. Beyond this active region, within the wavelength range of 1450–2500 nm, the spectral curves remained relatively flat with lower transmittance. These distinctive peaks and valleys reflect the key internal biochemical absorption characteristics of the soybean seeds.

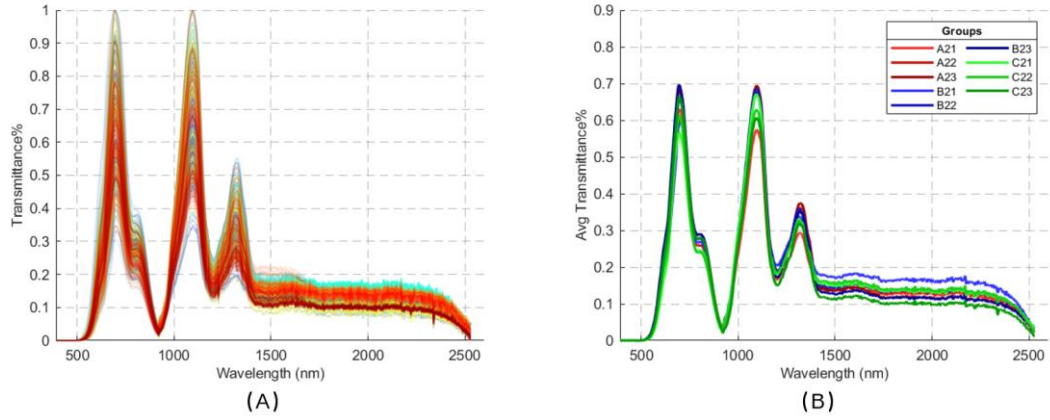


Figure 3. Spectral data:(A)Normalized Spectra, (B)Average Spectra Comparison

To minimize the effects of instrumentation and environmental factors on model performance, smooth denoising was applied before employing five preprocessing techniques: multiplicative scatter correction (MSC), first derivative (FD), Hilbert transform (HT), Savitzky–Golay smoothing (SG), and standard normal variate correction (SNV) ¹³. After optimizing the parameters, the order and window size for the SG algorithm were set to three and nine, respectively. Raw hyperspectral data were preprocessed to mitigate spectral baseline shifts caused by light scattering, resulting in smoother and more tightly clustered spectral curves (Figure 4). Smooth (Figure 4A), HT (Figure 4D), and SG (Figure 4E) all resulted in smoother spectral curves, while MSC (Figure 4B) smoothed the curves and eliminated scattering effects. FD (Figure 4C) enhanced the varying parts of the spectral curves, while SNV (Figure 4F) improved the reflectance ratio.

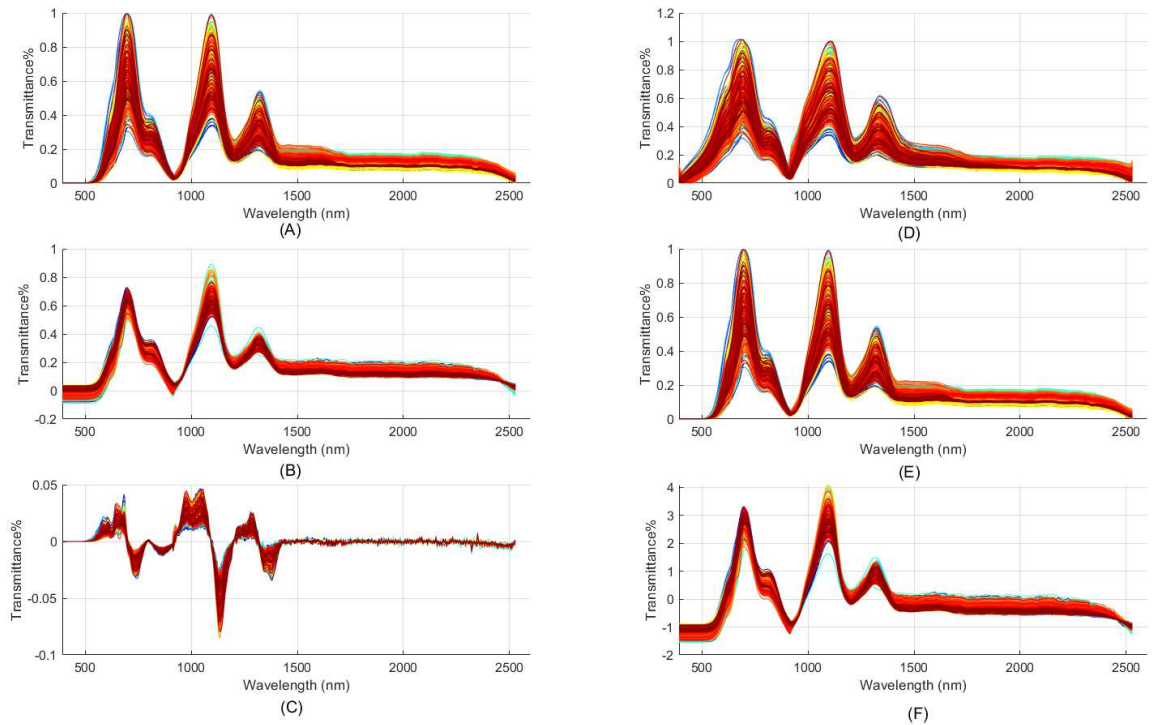


Figure 4. Preprocessed results of the soybean hyperspectral dataset: (A) smooth, (B) multiplicative scatter correction (MSC), (C) first derivative (FD), (D) Hilbert transform (HT), (E) Savitzky–Golay smoothing (SG), and (F) standard normal variate correction (SNV).

2.6. Modeling with Weak Classifiers

To further improve computational efficiency and processing speed, three algorithms were employed for data dimensionality reduction and feature wavelength extraction—principal component analysis (PCA) ¹⁴, successive projections algorithm

(SPA) ¹⁵, and competitive adaptive reweighted sampling (CARS) ^{16,17}. These algorithms were applied only to the training set of each cross-validation fold, rather than to the entire dataset, ensuring an unbiased evaluation of model performance. PCA reduces the dimensionality of hyperspectral data by transforming the original variables into a smaller set of principal components that capture the dominant spectral information. SPA selects feature wavelengths by iteratively choosing those with the largest projection values relative to the remaining wavelengths. CARS constructs partial least squares (PLS) models and uses adaptive reweighted sampling with cross-validation to eliminate less informative wavelengths and retain the subset that minimizes the root mean square error of cross-validation (RMSECV).

The selected weak classifiers were k-nearest neighbor (KNN), random forest (RF), support vector machine (SVM), logistic regression (LR), and naïve Bayes (NB) ^{18–22}. These algorithms are theoretically well established, easy to implement and interpret, and have been widely applied in hyperspectral image classification. In this study, KNN was implemented with $K=8$; SVM used a radial basis function (RBF) kernel ($\gamma = \text{“scale”}$, $C = 1.0$); RF consisted of 500 decision trees with bootstrap sampling and random feature subsets; LR adopted a multinomial formulation for multi-class seed vigor prediction; and NB followed a Gaussian naïve Bayes model. The hyperparameters of these classifiers were determined based on preliminary experiments (including grid search) and empirical recommendations from previous studies, ensuring stable classification performance while maintaining reasonable computational efficiency.

2.7. Stacking Ensemble Learning Model

Stacking, also known as stacked generalization or stacking ensemble, is a machine-learning ensemble algorithm that has gained widespread adoption across various industries ^{23–25}. Its core idea is to construct a two-layer cascaded model framework, consisting of base learners in the first layer and a meta-learner in the second layer. In this study, a 5-fold cross-validation scheme was used to build the Stacking model for soybean seed vigor classification: the first layer employs base learners such as KNN, RF, SVM, NB, and LR, whose predictions are combined into a meta-feature vector that serves as the input to the second layer; the second layer trains a meta-learner on these meta-features and selects the optimal combination of base learners and meta-learner configuration based on their performance on the validation set. A schematic of the Stacking-based soybean seed classification model is shown in Figure 5. The procedure is as follows.

(1) Data splitting and feature extraction : The hyperspectral data were split into a training set D and a test set S in a 4:1 ratio. The training set D contained 80 seeds per vigor category (720 samples in total), and the test set S contained 20 seeds per category (180 samples in total). Dimensionality reduction and feature wavelength selection (PCA, SPA, or CARS) were performed only on the training set D . Under the cross-validation setup, the training set D was divided into K equal folds, each containing 144 seeds (16 per category), denoted $D_1, D_2, \dots, D_k$. Defining $D'_j = D - D_j$, D_j and D'_j represent the j -th fold training set and test set, respectively, in the K -fold cross-validation. The test set S was used only for final evaluation and did not participate in any feature selection or training step.

(2) Base learner training and meta-feature generation: Suppose T base learners are used on the training set D . The t -th base learner $h_t^{(j)}$ is obtained by training the t -th learning algorithm on the j -th fold training set D_j . For each sample x_i in the j -

th fold test set D'_j , let $Pre_{.t,i} = h_t^{(j)}(x_i)$. The set of predictions for x_i is $Pre_i = (Pre_{.1,i}; Pre_{.2,i}; \dots; Pre_{.t,i})$, and the collection $D_{new} = \{(Pre_i)\}$ forms the training set for the second-layer meta-learner.

(3) Meta-learner training: The second-layer meta-learner (e.g., LR or SVM) is trained on this meta-feature matrix and the corresponding labels over D .

(4) Final prediction: After applying K -fold cross-validation with each base learner to the test set S , a set of predictions $Pres = \{p_1, p_2, \dots, p_k\}$ is obtained. The predictions from the T base learners are merged and averaged to form the new test set $S_{new} = \{Pres_{.1}, Pres_{.2}, \dots, Pres_{.t}\}$, which is used as the input to the second-layer meta-learner for final evaluation.

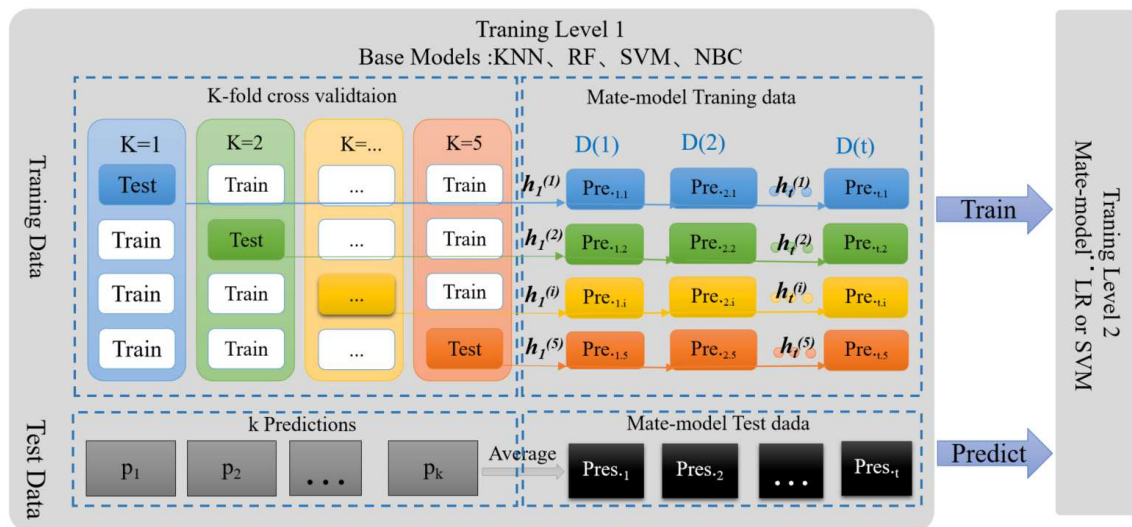


Figure 5. Schematic diagram of the soybean classification model using stacking ensemble learning.

3. Results and Discussion

3.1 Germination Results of Soybean Seeds

The vigor characteristics of naturally aged soybean seeds were determined through ISTA-compliant germination assays (Table 1). The germination index (GI) and vigor index (VI) exhibited significant downward trends with prolonged aging duration, consistent with lower vigor in older seeds; genotypic differences predominantly governed the specific patterns of vigor decline.

Table 1. Germination index (GI) and vigor index (VI) of three soybean cultivars under natural aging

	Soybean cultivar	2021	2022	2023
GI	Gutian Dou	0.67	6.5	10
	Dian Dou No.1	4.42	10.83	17.17
	Zhonghuang No.1	5.58	13.7	13.37
VI	Gutian Dou	1.37	396.18	655.5
	Dian Dou No.1	119.34	954.12	1705.84
	Zhonghuang No.1	104.15	1356.3	1627.8

The traditional variety 'Gutiandou' exhibited significant sensitivity to aging; its GI and VI for the 2021 aged seeds were substantially lower than those of the modern varieties from the same year. Although physiological and biochemical parameters were not directly measured in this study, previous research indicates that its seed coat structure or the genetic stability

of endogenous antioxidants in 'Gutiandou' endogenous antioxidants may be inferior, leading to exacerbated membrane lipid peroxidation²⁶. In contrast, 'Diandou No. 1' and 'Zhonghuang No. 1' demonstrated excellent storage tolerance, maintaining relatively high vigor in their aged seeds. Notably, the *Gi* of 'Zhonghuang No. 1' fluctuated only slightly between the 2022 and 2023 harvest years, which likely reflects the complexity of the synergistic regulation between its mitochondrial respiratory efficiency and antioxidant system²⁷. These findings suggest that the aging sensitivity of different varieties is fundamentally governed by their initial material foundations, which are dictated by their genetic backgrounds.

3.2. Spectral Data Dimensionality Reduction

3.2.1. PCA Dimensionality Reduction

After preprocessing, principal component analysis (PCA) was applied to each dataset to retain 99% of the original feature information. The number of retained features and the variance explained by the first three principal components (PC1, PC2, PC3) for each preprocessing method are summarized in Table 2. The retained feature count ranged from 7 (Smooth, SG) to 22 (FD) depending on the method. Three-dimensional scatter plots of PC1, PC2, and PC3 were generated to visualize the distribution of samples by harvest year (Figure 6).

Table 2. Number of retained features and variance explained by PC1-PC3 (99% information preserved)

Preprocessing Method	Smooth	MSC	FD	SG	SNV
Number of Wavelengths	7	9	22	7	10
PC1(%)	72.95	54.85	71.27	72.95	54.02
PC2(%)	10.27	22.69	5.17	10.27	22.89
PC3(%)	3.75	7.25	3.63	3.75	7.38

MSC, multiplicative scatter correction; FD, first derivative; SG, Savitzky–Golay smoothing; Smooth, Smoothing; SNV, standard normal variate correction

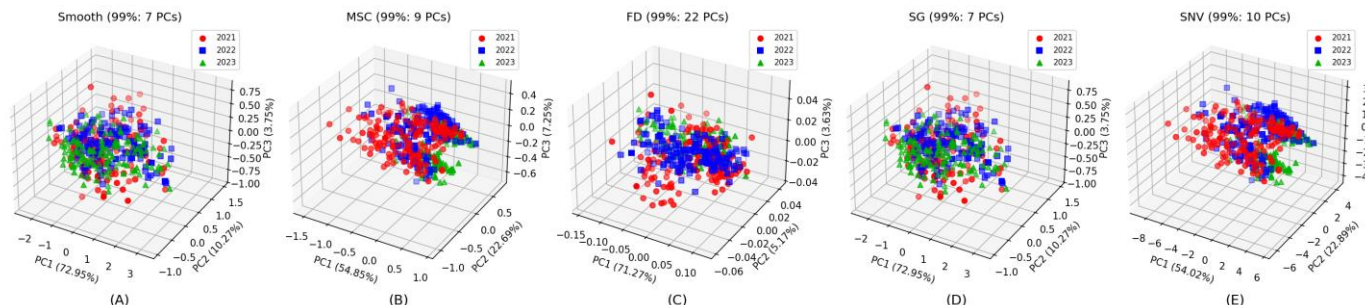


Figure 6. Three-dimensional PCA scatter plots of the hyperspectral data by harvest year (2021, 2022, 2023) under five preprocessing methods:

(A) Smooth, (B) MSC, (C) FD, (D) SG, (E) SNV.

In the three-dimensional principal component analysis (PCA) spaces across all preprocessing methods, the data points from the three harvest years (2021–2023) exhibited highly interlaced and severely overlapping distributions. In comparison, the data processed using the first derivative (FD) (Figure 6C) displayed a more dispersed spatial distribution and required a greater number of principal components to retain 99% of the variance, highlighting the effectiveness of this method in extracting fine-grained, localized spectral features. Concurrently, the extensive overlap of harvest years in the 3D space suggested that the initial principal components were likely driven more by inherent genotypic differences among the varieties than by aging-induced spectral changes. Therefore, simple linear dimensionality reduction methods are insufficient for effectively distinguishing seed vigor levels. This limitation underscores the necessity of introducing advanced machine learning models for robust classification.

3.2.2. SPA Feature Extraction

SPA was applied for dimensionality reduction and feature wavelength selection. The optimization step was fixed at 1, and the number of optimal feature wavelengths was determined by minimizing the root mean square error (RMSE) of the model. The corresponding feature spectral data were then extracted from the full spectra based on the selected wavelengths. As shown in Figure 7, which displays the feature selection results from the FD preprocessed data, the RMSE was lowest (0.36613) when 18 feature wavelengths were selected. The optimal wavelengths—462.5, 497.1, 516.9, 594.4, 612.0, 675.3, 713.7, 1105.7, 1212.1, 1245.7, 1492.2, 1609.9, 1643.5, 1721.9, 1749.9, 1867.6, 1946.0 and 2270.9nm—are indicated in the figure.

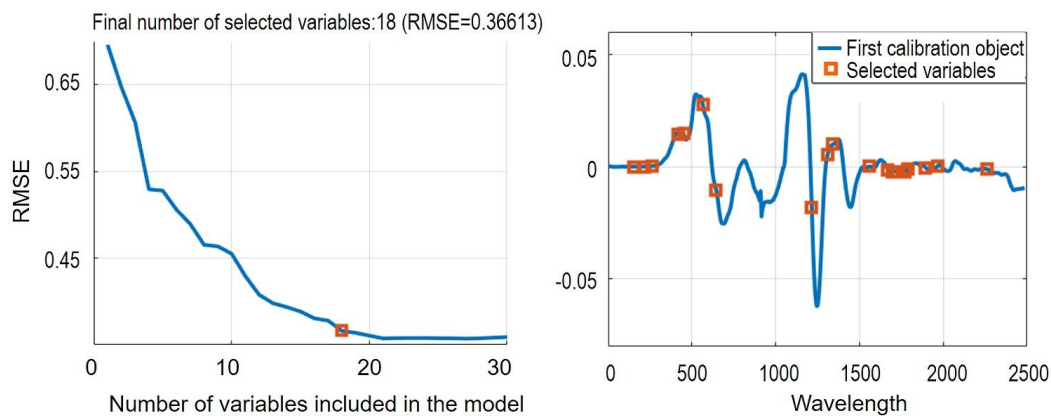


Figure 7. Dimensionality reduction results of the SPA for the soybean sample transmittance hyperspectral dataset

3.2.3. CARS Feature Extraction

The CARS method was applied to reduce the dimensionality of the original spectral data. A partial least squares regression (PLSR) model was constructed using 50 Monte Carlo sampling iterations and a 5-fold cross-validation strategy. The optimization outcomes were depicted using the data preprocessed via the HT method as a case study for CARS dimensionality reduction (Figure 8). The number of feature wavelengths was determined by selecting the wavelengths from the sampling iterations that minimized RMSECV in the PLSR model. During the iterative process, the number of retained wavelengths decreased steadily. Initially, this reduction was rapid, followed by a slower, more gradual decline. From iteration 0 to 32, the eliminated wavelengths had minimal effect on RMSECV, causing a slight decrease in its value. However, beyond 32 iterations, the RMSECV started to rise, leading to a significant increase. Therefore, the characteristic wavelengths at 32 sampling times can be extracted.

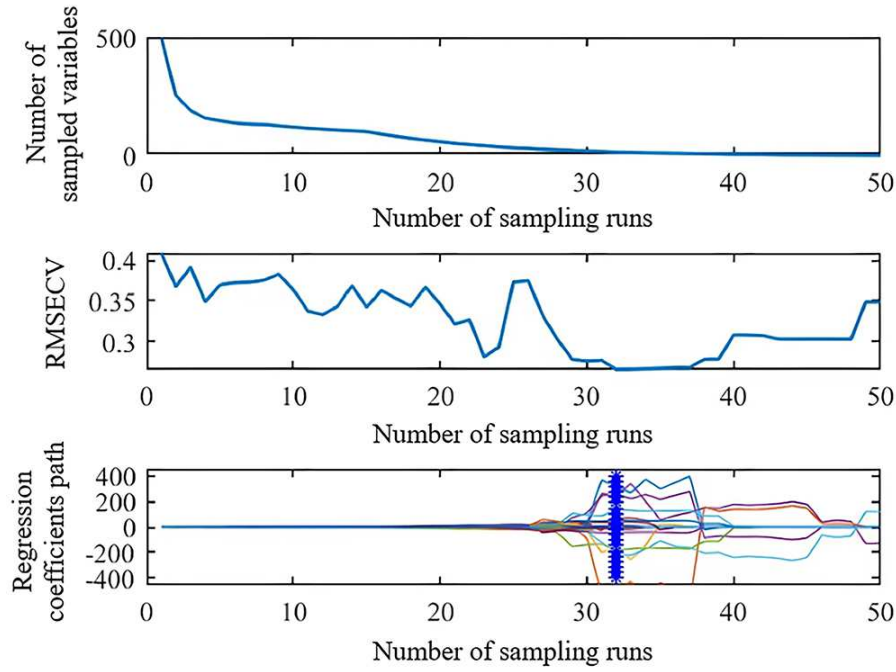


Figure 8. Dimensionality reduction results of competitive adaptive reweighted sampling (CARS) for the transmittance hyperspectral dataset.

3.3. Classification of Soybean Seed Vigor Based on Weak Classifiers

The training results based on the first-layer base learners are presented in Table 3. In this study, five models—KNN, RF, SVM, LR, and NB—were employed to predict soybean seed vigor, using the validation set (144 seeds) accuracy as the primary evaluation metric, where a higher accuracy indicates better model performance. The results demonstrate that the soybean seed vigor classification models established by PCA-FD-LR, CARS-HT-SVM, and CARS-HT-KNN achieved the optimal performance, with testing set accuracies of 95.83%, 95.13%, and 94.44%, respectively. Regarding preprocessing methods, HT and FD yielded superior results; specifically, the number of models achieving a testing accuracy of no less than 90% was five for HT and two for FD, indicating that HT performed the best overall. In terms of dimensionality reduction methods, CARS and PCA generally exhibited better efficacy, yielding five and two models with accuracies of $\geq 90\%$, respectively. From the perspective of algorithm types, RF and LR delivered the best performance, each having two dataset combinations that achieved an accuracy of $\geq 90\%$. Conversely, models trained on datasets processed with other combinations of dimensionality reduction or preprocessing methods exhibited relatively lower performance.

Table 3. Training and test set accuracy of base classifiers by preprocessing–dimensionality combination.

Preprocessed dimensionality-reduced dataset	KNN		RF		SVM		LR		NB	
	Train set	Test set	Train set	Test set	Train set	Test Set	Train set	Test Set	Train set	Test set
PCA–Smooth	86.27	81.25	100	85.42	89.8	69.44	78.84	70.14	81.09	75.69
PCA–MSC	87.85	85.42	95.66	81.25	92.43	72.22	80.81	70.83	80.48	79.17
PCA–FD	93.47	86.81	100	93.06	100	86.81	100	95.83	94.12	85.42
PCA–SG	86.37	79.17	100	85.42	86.29	64.58	80.28	75.69	81.07	75.69

PCA-SNV	88.54	85.42	100	84.03	87.41	64.58	79.51	68.75	77.96	73.61
SPA-FD	85.86	77.78	100	87.5	92.51	73.61	83.67	79.17	80.26	81.25
SPA-MSc	79.72	75	100	84.03	95.56	75.69	74.37	67.36	67.22	70.14
SPA-SG	79.59	78.47	93.45	79.17	91.36	72.22	70.09	55.56	69.99	72.22
SPA-SNV	83.25	78.47	95.24	82.64	90.03	72.22	77.3	61.81	65.23	70.14
CARS-FD	85.53	82.64	100	84.03	94.18	72.92	82.95	71.53	72.06	68.06
CARS-MSc	71.49	60.42	88.16	62.5	89.08	62.5	78.37	76.39	53.82	60.42
CARS-SG	75.35	68.75	90.31	72.22	91.03	63.19	72.88	69.44	58.05	61.81
CARS-SNV	73.04	65.28	91.85	72.92	92.92	72.92	63.93	53.47	55.18	62.5
CARS-HT	100	94.44	100	90.97	100	95.14	97.15	91.67	95.19	93.06

KNN, k-nearest neighbor; RF, random forest; SVM, support vector machine NB, naïve Bayes; LR- logistic regression; PCA, principal component analysis; MSC, multiplicative scatter correction; FD, first derivative; SNV, standard normal variate correction; SG, Savitzky–Golay smoothing; SPA, successive projection algorithm; HT, Hilbert transform; CARS, competitive adaptive reweighted sampling

The training results based on the first-layer base learners are presented in Table 3. In this study, five models—KNN, RF, SVM, LR, and NB—were employed to predict soybean seed vigor, using the validation set (144 seeds) accuracy as the primary evaluation metric, where a higher accuracy indicates better model performance. The results demonstrate that the soybean seed vigor classification models established by PCA-FD-LR, CARS-HT-SVM, and CARS-HT-KNN achieved the optimal performance, with testing set accuracies of 95.83%, 95.14%, and 94.44%, respectively. Regarding preprocessing methods, HT and FD yielded superior results; specifically, the number of models achieving a testing accuracy of no less than 90% was five for HT and two for FD, indicating that HT performed the best overall. In terms of dimensionality reduction methods, CARS and PCA generally exhibited better efficacy, yielding five and two models with accuracies of $\geq 90\%$, respectively. From the perspective of algorithm types, RF and LR delivered the best performance, each having two dataset combinations that achieved an accuracy of $\geq 90\%$. Conversely, models trained on datasets processed with other combinations of dimensionality reduction or preprocessing methods exhibited relatively lower performance.

To further compare the performance, the classification results of the two best-performing methods from Table 3—LR based on PCA-FD and SVM based on CARS-HT—were visualized using confusion matrices on both the modeling and prediction sets (Figure 9). As shown in the figures, the modeling set comprised 720 samples. Under a 5-fold cross-validation scenario, the training subset contained 576 seeds (99 non-viable, 161 low-vigor, and 316 high-vigor), while the prediction subset consisted of 144 seeds (28 non-viable, 40 low-vigor, and 76 high-vigor). For the modeling set, the confusion matrices for both models were strictly diagonal, yielding an accuracy of 100%, which indicates sufficient fitting to the training data. On the prediction set, the PCA-FD-LR model correctly classified 138 seeds (accuracy: 95.83%), and the CARS-HT-SVM model correctly classified 137 seeds (accuracy: 95.14%). The misclassifications in both models were primarily concentrated at the boundaries between non-viable and low-vigor, as well as between low-vigor and high-vigor categories. The high-vigor category exhibited the highest recognition rate, and importantly, there was no direct misclassification between the non-viable and high-vigor extremes. These results are consistent with the biological characteristics of seed vigor, which represents a

continuous degradation process where boundary samples are inherently difficult to distinguish. This demonstrates that the selected methods possess strong robustness and practical value, particularly in identifying high-vigor seeds.

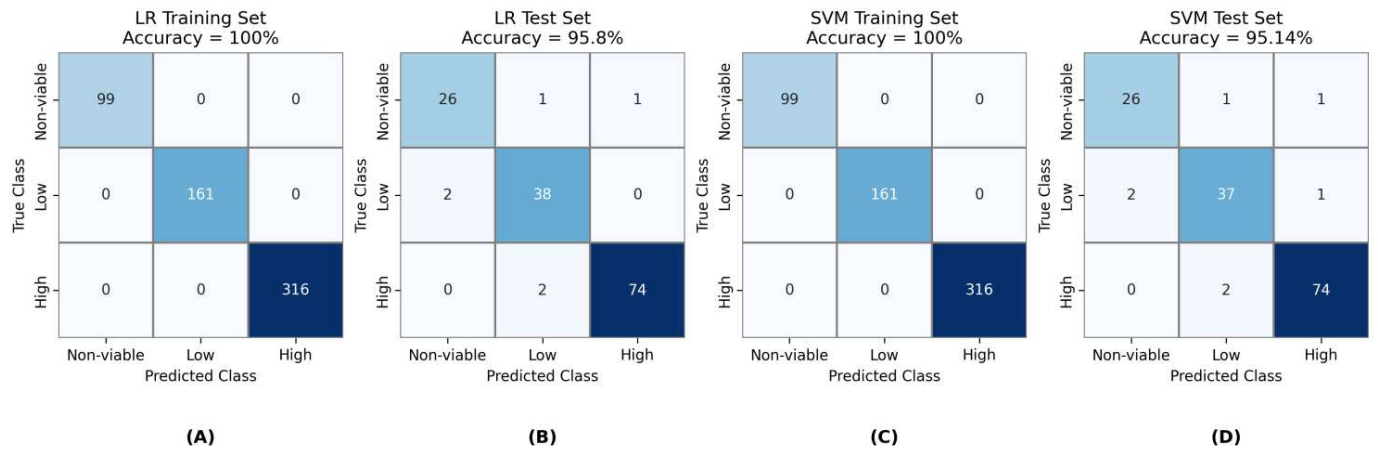


Figure 9. Confusion matrices for PCA-FD-LR and CARS-HT-SVM: (A) LR training set, (B) LR validation set, (C) SVM training set, (D) SVM validation set.

3.4 Classification Results of Soybean Seeds Based on Stacking Ensemble Learning

To construct the optimal ensemble, RF and LR were comparatively evaluated as meta-learners. Experimental comparisons revealed that the Stacking model utilizing KNN, RF, SVM, and NB as base learners, with LR as the meta-learner, achieved the best overall performance. An independent test set comprising 180 seeds was used for validation, and the results are presented in Table 4.

Table 4. Stacking ensemble model performance (training accuracy, test accuracy, and F1-score) by preprocessing–dimensionality combination.

	PCA					SPA					CARS			
	Smooth	MSC	FD	SG	SNV	FD	MSC	SG	SNV	FD	MSC	HT	SG	SNV
Train Set	99.31	98.87	100	99.78	96.06	100	100	100	100	100	83.89	100	94.69	92.37
Test Set	86.11	83.89	97.22	84.44	85	91.11	84.44	82.78	83.33	86.11	65	98.33	80	76.11
F1	85.32	82.95	96.88	83.61	84.12	90.45	83.58	81.92	82.41	85.28	63.52	98.12	79.18	75.26

FD, first derivative; SNV, standard normal variate correction; SG, Savitzky–Golay smoothing; SPA, successive projection algorithm; HT, Hilbert transform; CARS, competitive adaptive reweighted sampling; MSC, multiplicative scatter correction; Smooth, smoothing

Table 4 details the training set accuracy, testing set accuracy, and F1-score of the Stacking models across various combinations of preprocessing and dimensionality reduction methods. On the training set, the accuracy of most combinations approached or reached 100%, indicating sufficient model fitting. The testing set accuracy and F1-score, which are more reflective of the model's generalization ability, further demonstrated the model's robustness. When evaluated jointly by testing accuracy and F1-score, the CARS-HT combination performed the best, achieving a testing accuracy of 98.33% and an F1-score of 98.12%. Furthermore, the PCA-FD and SPA-FD combinations achieved testing accuracies of 97.22% and 91.11%,

with corresponding F1-scores of 96.88% and 90.45%, respectively. These results are significantly superior to the performance of any single first-layer classifier on the corresponding datasets.

The stacked model based on the CARS-HT dataset established a multidimensional spectral fingerprint characterizing the critical biochemical mechanisms of soybean seed vigor by screening 17 key feature wavelengths (526.9, 559.3, 591.9, 607.0, 652.5, 667.7, 677.9, 822.1, 848.2, 920.8, 982.4, 1139.3, 1458.6, 1604.3, 1755.5, 2310.1, and 2444.6 nm). During natural aging, the contents of malondialdehyde (MDA), proteins, moisture, chlorophyll, and catalase—key biochemical indicators affecting seed vigor—change significantly with prolonged storage duration^{28, 29}.

In the visible region (526.9–677.9 nm), the band at 526.9 nm reflects antioxidant capacity through transmittance variations associated with chlorophyll degradation products; its intensity is negatively correlated with lipid peroxidation and seed deterioration³⁰. Meanwhile, 677.9 nm indicates oxidative status via the electronic transition characteristics of aromatic amino acids and residual pigments, further confirming the superior antioxidant properties of high-vigor seeds³¹. Within the short-wave near-infrared region (822.1–1604.3 nm), 822.1 nm and 920.8 nm elucidate water activity and oligosaccharide dynamics through the overtones of O-H bonds and C-H/C-O stretching vibrations, respectively^{32,33}. Notably, moisture content exceeding 13% markedly accelerates enzymatic deterioration, while oligosaccharides enhance seed desiccation tolerance by maintaining cell membrane integrity³². The N-H vibrational signal at 1458.6 nm specifically monitors the 7S/11S globulin ratio and conformational transitions³⁴, directly influencing the efficiency of amino acid metabolic supply during germination. Concurrently, the stretching vibrations near 1604.3 nm reveal the accumulation of lipid peroxidation products, such as MDA, and protein structural changes³⁵. In the long-wave near-infrared region (1755.5–2444.6 nm), the 1755.5 nm band characterizes fatty acid unsaturation and lipid oxidation levels under aging stress. Finally, the 2310.1 nm and 2444.6 nm wavelengths assess cellular structural stability through the degree of protein cross-linking and the cellulose degradation state, respectively. By integrating these multidimensional physiological mechanisms—including antioxidant defense, dynamic utilization of storage proteins, and cellular structural stability—this spectral fingerprint provides a highly interpretable physicochemical basis for the CARS-HT stacked model to achieve high-precision seed vigor classification³⁶.

4. Conclusions

In this study, a rapid, fully non-destructive detection method for evaluating the vigor of naturally aged soybean seeds was developed using transmission hyperspectral imaging (400–2500 nm) combined with Stacking ensemble learning. By comparing various preprocessing and dimensionality reduction techniques, 17 key characteristic wavelengths associated with critical biochemical mechanisms (e.g., moisture dynamics, protein conformation, and lipid oxidation) were successfully identified.

The empirical results confirmed that the Stacking ensemble learning algorithm—utilizing multiple weak classifiers as base learners and logistic regression as the meta-learner—significantly outperformed individual machine learning models. The optimal CARS-HT-Stacking model achieved an exceptional classification accuracy of 98.33% alongside an F1-score of 98.12% on an independent test set.

These findings demonstrate that the synergistic application of transmission hyperspectral technology and advanced ensemble learning can effectively overcome the challenges of non-linear spectral interference and limited sample sizes. This research not only provides a highly accurate and interpretable technical paradigm for single-seed vigor evaluation but also lays a crucial theoretical and methodological foundation for the intelligent screening of germplasm resources and precision agriculture.

References

1. Hamman, B.; Egli, D. B; and Gwen, K. Seed vigor, soilborne pathogens, preemergent growth, and soybean seedling emergence. *Crop Sci.* 2002, 42(2), 451-457. [doi:10.2135/2002.4510](https://doi.org/10.2135/2002.4510).

2. Taylor A G. Seed storage, germination, quality, and enhancements. The physiology of vegetable crops. Wallingford UK: CABI, 2020: 1-30. doi:[10.1079/9781786393777.00](https://doi.org/10.1079/9781786393777.00).
3. Rahman, A.; Cho, B.K. Assessment of seed quality using non-destructive measurement techniques: a review. *Seed Sci. Res.* 2016, 26(4), 285-305. doi:[10.1017/S0960258516000234](https://doi.org/10.1017/S0960258516000234).
4. Goel, A., Goel, A. K., & Sheoran, I. S. 2003. Changes in oxidative stress enzymes during artificial ageing in cotton (*Gossypium hirsutum* L.) seeds. *Journal of Plant Physiology*. 2003, 160(9): 1093-1100. doi:[10.1078/0176-1617-00881](https://doi.org/10.1078/0176-1617-00881).
5. Yin, H, et al. Detection of moisture content and size of pumpkin seeds based on hyperspectral reflection and transmission imaging techniques. *Journal of Food Composition and Analysis*. 2023, 124: 105651. doi:[10.1016/2023.105651](https://doi.org/10.1016/2023.105651).
6. Liang Y, Li Z, Shi J, et al. Advances in hyperspectral imaging technology for grain quality and safety detection: a review. *Foods*, 2025, 14(17): 2977. doi:[10.3390/14172977](https://doi.org/10.3390/14172977).
7. Zhao, B., Wang, A. W., & Zhao, H. Determination of sweet corn seed vigor by near infrared spectra and chemometrics. *Journal of Southern Agriculture*, 2022, 53(7): 1875. doi:[10.5555/20220536492](https://doi.org/10.5555/20220536492).
8. Li, He, et al. Rapid detection of the viability of naturally aged maize seeds using multimodal data fusion and explainable deep learning techniques. *Food Chem*, 2025, 478: 143692. doi:[10.1016/2025.143692](https://doi.org/10.1016/2025.143692).
9. Shuai L, Li Z, Chen Z, et al. A research review on deep learning combined with hyperspectral Imaging in multiscale agricultural sensing. *Computers and Electronics in Agriculture*, 2024, 217: 108577. doi:[10.1016/2023.108577](https://doi.org/10.1016/2023.108577).
10. Chu Y R, Jo M S, Kim G E, et al. Non-Destructive Seed Viability Assessment via Multispectral Imaging and Stacking Ensemble Learning. *Agriculture*, 2024, 14(10): 1679. doi:[10.3390/14101679](https://doi.org/10.3390/14101679).
11. Zhao Z, Alzubaidi L, Zhang J, et al. Robust and explainable framework to address data scarcity in diagnostic imaging. *Computers in Biology and Medicine*, 2025, 197: 111052. doi:[10.1016/2025.111052](https://doi.org/10.1016/2025.111052).
12. Gao T, Chandran A K N, Paul P, et al. HyperSeed: An End-to-End method to process hyperspectral images of seeds. *Sensors*, 2021, 21(24): 8184. doi:[10.3390/s21248184](https://doi.org/10.3390/s21248184).
13. Choi J H, Park S H, Jung D H, et al. Hyperspectral imaging-based multiple predicting models for functional component contents in Brassica juncea. *Agriculture*, 2022, 12(10): 1515. doi:[10.3390/12101515](https://doi.org/10.3390/12101515).
14. Uddin, M.P.; Mamun, M.A.; Hossain, M.A. PCA-based feature reduction for hyperspectral remote sensing image classification. *IETE Tech. Rev.* 2021, 38, 377–396. doi:[10.1080/02564602.2020.1740615](https://doi.org/10.1080/02564602.2020.1740615).
15. Chen, J.; Bai, T.; Zhang, N.; Zhu, L.; Zhang, X. Hyperspectral detection of sugar content for sugar-sweetened apples based on sample grouping and SPA feature selecting methods. *Infrared Phys. Technol.* 2022, 125: 104240. doi:[10.1016/2022.104240](https://doi.org/10.1016/2022.104240).
16. Zhang H, Zhan B, Pan F, et al. Determination of soluble solids content in oranges using visible and near infrared full transmittance hyperspectral imaging with comparative analysis of models. *Postharvest Biology and Technology*, 2020, 163: 111148. doi:[10.1016/2020.111148](https://doi.org/10.1016/2020.111148).
17. Li, X.; Fu, X.; Li, H. A CARS-SPA-GA feature wavelength selection method based on hyperspectral imaging with potato leaf disease classification. *Sensors (Basel)*. 2024, 24, 6566. doi:[10.3390/s24206566](https://doi.org/10.3390/s24206566).
18. Ge Y, Song S, Yu S, et al. Rice seed classification by hyperspectral imaging system: A real-world dataset and a credible algorithm. *Computers and Electronics in Agriculture*, 2024, 219: 108776. doi:[10.1016/2024.108776](https://doi.org/10.1016/2024.108776).
19. Chen M, Chang Z, Jin C, et al. Classification and recognition of soybean quality based on hyperspectral imaging and random forest methods. *Sensors*, 2025, 25(5): 1539. doi:[10.3390/s25051539](https://doi.org/10.3390/s25051539).
20. Wu Q, Xu L, Zou Z, et al. Rapid nondestructive detection of peanut varieties and peanut mildew based on hyperspectral imaging and stacked machine learning models. *Frontiers in plant science*, 2022, 13: 1047479. doi:[10.3389/2022.1047479](https://doi.org/10.3389/2022.1047479).

21. Liu W, Zeng S, Wu G, et al. Rice seed purity identification technology using hyperspectral image with LASSO logistic regression model. *Sensors*, 2021, 21(13): 4384.doi:[1424-8220/21/13/4384](https://doi.org/10.3390/s21134384).
22. Singh, A. K., Sunkara, R., Kadambi, G. R., & Palade, V.. Spectral–Spatial Classification With Naive Bayes and Adaptive FFT for Improved Classification Accuracy of Hyperspectral Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2023,17, 1100-1113.doi:[10.1109/2023.3327346](https://doi.org/10.1109/2023.3327346).
23. Jiang, M.; Liu, J.; Zhang, L.; Liu, C. An improved Stacking framework for stock index prediction by leveraging tree-based ensemble models and deep learning algorithms. *Physica. A. Stat.* 2020, 541, 122272. doi:[10.1016/2019.122272](https://doi.org/10.1016/2019.122272).
24. Ribeiro, M.H.D.M.; dos Santos Coelho, L. Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series. *Appl. Soft Comput.* 2020, 86, 105837. doi:[10.1016/2019.105837](https://doi.org/10.1016/2019.105837).
25. Cui, S.; Yin, Y.; Wang, D.; Li, Z.; Wang, Y. A stacking-based ensemble learning method for earthquake casualty prediction. *Appl. Soft Comput.* 2021, 101, 107038. doi:[10.1016/2020.107038](https://doi.org/10.1016/2020.107038).
26. Rao P J M, Pallavi M, Bharathi Y, et al. Insights into mechanisms of seed longevity in soybean: a review. *Frontiers in Plant Science*, 2023, 14: 1206318.doi: [10.3389/2023.1206318](https://doi.org/10.3389/2023.1206318).
27. Lin Y, Xu H, Yin G, et al. Dynamic changes in membrane lipid metabolism and antioxidant defense during soybean (*Glycine max* L. Merr.) seed aging. *Frontiers in Plant Science*, 2022, 13: 908949.doi:[10.3389/fpls.2022.908949](https://doi.org/10.3389/fpls.2022.908949).
28. Ebone, L.A.; Caverzan, A.; Silveira, D.C.; Siqueira, L.O.; Lângaro, N.C.; Chiomento, J.L.T.; Chavarria, G. Biochemical profile of the soybean seed embryonic axis and its changes during accelerated aging. *Biology*. 2020, 9, 186. doi:[10.3390/9080186](https://doi.org/10.3390/9080186).
29. Sheidaei, S.; Hamidi, A.; Sadeghi, H.; Oskouei, B. Evaluation of initial seed quality and storage conditions on biochemical and physiological changes of soybean seeds. *Iranian Journal of Seed Science and Technology*. 2020, 9, 101–118.doi:[10.22034/2019.123739.1232](https://doi.org/10.22034/2019.123739.1232).
30. Tang, Z., Wang, X., Xiang, Y., et al. Application of hyperspectral technology for leaf function monitoring and nitrogen nutrient diagnosis in soybean (*Glycine max* L.) production systems on the Loess Plateau of China. *European Journal of Agronomy*,2024, 154:127098.doi:[10.1016/2024.127098](https://doi.org/10.1016/2024.127098).
31. Wang, Wei-Qing, et al. Proteomics of seed development, desiccation tolerance, germination and vigor.*Plant Physiol. Biochem.* 2015, 86, 1-15.doi:[10.1016/2014.11.003](https://doi.org/10.1016/2014.11.003).
32. Hollung, K, et al. Evaluation of nonstarch polysaccharides and oligosaccharide content of different soybean varieties (*Glycine max*) by near-infrared spectroscopy and proteomics. *Journal of Agricultural and Food Chemistry*. 2005, 53(23), 9112-9121.doi:[10.1021/jf051438r](https://doi.org/10.1021/jf051438r).
33. Guo, Z, et al. Application of visible-near-infrared hyperspectral imaging technology coupled with wavelength selection algorithm for rapid determination of moisture content of soybean seeds.*Journal of Food Composition and Analysis*. 2023, 116:105048.doi:[10.1016/j.jfca.2022.105048](https://doi.org/10.1016/j.jfca.2022.105048).
34. Duan, Z, et al. Genetic regulatory networks of soybean seed size, oil and protein contents. *Front. Plant Sci.* 2023, 14, 1160418.
35. Jin, B., Qi, H., Jia, L., et al. Determination of viability and vigor of naturally-aged rice seeds using hyperspectral imaging with machine learning. *Infrared Physics & Technology*,2022, 122, 104097.doi:[10.1016/j.infrared.2022.104097](https://doi.org/10.1016/j.infrared.2022.104097).
36. Silva, Martha Freire da, et al. Near infrared spectroscopy for the classification of vigor level of soybean seed. *Revista Ciência Agronômica*,2023, 55, e20238703.doi:[10.5935/1806-6690.20240005](https://doi.org/10.5935/1806-6690.20240005).

Acknowledgments

This work was supported by the National Natural Science Foundation of China. The authors also express sincere gratitude to the technical staff at the Artificial Intelligence Laboratory of Nanjing Agricultural University for their valuable assistance.

Author Contributions

F.L. and W.L. conceived and designed the experiments. F.L. and P.N. conducted the experiments and collected the data. F.L. performed the spectral preprocessing, feature extraction, and modelling. F.L. and P.N. analysed the results and prepared the figures. F.L. and P.N. wrote the original draft. W.L. supervised the project, acquired funding, and reviewed and edited the manuscript. All authors reviewed the manuscript.

Funding

This research was funded by the General Program of the National Natural Science Foundation of China, grant number 32071896 and Fundamental Research Funds for the Central Universities of China, grant number QTPY2025009.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to W.L.

Reprints and permissions information is available at www.nature.com

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [partialdatasets.csv](#)