


Research Article

Soybean Saponin Content Detection Based on Spectral and Image Information Combination

Hongmin Sun,¹ Xifan Meng,¹ Yingpeng Han,² Xiao Li,³ Xiaoming Li ,¹ and Yongguang Li²

¹School of Electrical and Information, Northeast Agricultural University, Harbin 150006, China

²School of Agriculture, Northeast Agricultural University, Harbin 150006, China

³China Agriculture Press, Beijing 100125, China

Correspondence should be addressed to Xiaoming Li; lixiaoming@neau.edu.cn

Received 16 October 2023; Revised 22 April 2024; Accepted 3 May 2024; Published 10 May 2024

Academic Editor: Daniel Cozzolino

Copyright © 2024 Hongmin Sun et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Soybean saponin is a natural antioxidant and is anti-inflammatory. Hyperspectral analysis technology was applied to detect soybean saponin content rapidly and nondestructively in this paper. Firstly, spectral preprocessing methods were studied, and standard normal variable (SNV) was used to remove noise information. Secondly, a two-step hybrid variable selection approach based on synergy interval partial least squares (SiPLS) and iteratively retains informative variables (IRIV) was proposed to extract characteristic variables. Then, the ensemble learning model was constructed by back propagation neural network (BPNN), deep forest (DF), partial least squares regression (PLSR), and extreme gradient boosting (EXG). Finally, image information was combined into spectral data to improve model accuracy. The prediction coefficient of determination (R^2) of the final model reached 0.9216. It can provide rapid, nondestructive, and accurate detection technology of soybean saponin content. A combination of spectral and image information will provide a new idea for application of hyperspectral.

1. Introduction

Soybean is a vitally economic food crop. Because of containing protein, fat, saponin, and amino acid, soybean is mainly used for oil production, edible proteins, building materials, and cosmetics [1, 2]. Soybean saponins are metabolic products during the growth of soybean, and their content range from 0.6% to 6.2% [3, 4]. Soybean saponin has multifaceted physiological functions such as anticancer, antiaging, antiallergic, and antiviral effects [5–8]. Prolonged consumption of soybean saponin can effectively mitigate diseases like hypertension, hyperlipidaemia, and obesity [9–11]. Soybean saponin content detection is important for quality testing for soybean breeding.

Traditionally detection methods of soybean saponin content rely on wet chemical methods, such as high-performance liquid chromatography [12], colorimetry [13], and liquid chromatography mass spectrometry [14]. However, these methods have shortage of cumbersome procedure, high cost, or subjective results [15]. Therefore, it

is crucial to develop a rapid, accurate, and low-cost method for detecting soybean saponin content.

In recent years, spectral analysis techniques have been widely used in the detection of crop nutrient content due to their advantages of fast analysis speed, easy operation, and no sample damage. Compared to near-infrared technology, hyperspectral technology offers wider wavelength range and higher information localization accuracy. Because of acquisition of spatial distribution information of spectral data, hyperspectral technology can collect pixel-level spectral data of crops, so it has significant advantages in crop analysis. Guo et al. [16] detected the moisture content of individual soybeans based on the interval variable iterative space shrinkage approach and successive projection algorithm by using a visible-near-infrared hyperspectral imaging device. Song et al. [17] conducted nondestructive detection of moisture and fatty acid content in rice using a near-infrared spectroscopic imaging device combined with PLSR. Zhang et al. [18] detected 25 different nutrients in soybeans by near-infrared reflectance spectra including soybean saponin with

R^2 of 0.35. Berhow et al. [19] developed a multiple linear regression model for detecting the content of soy isoflavones and saponins based on 3200 soybean samples. The model effectively detected isoflavone, but R^2 for the soybean saponin content detecting model was only 0.6.

Although spectral and image information combination will bring larger amount of data, it provides higher spatial resolution and extract light absorption and surface grayscale variations of measured substances. Zheng et al. [20] detected soil total nitrogen content by combining infrared spectrum and image information with R^2 value of 0.815 and root mean square error (RMSE) of 0.153. Wang et al. [21] identified damage of soybean using high-quality spatial resolution-hyperspectral imaging images by combination of hyperspectral imaging and RGB images with model accuracy of 98.36%.

Compared to single-spectrum data, combination of spectral and image information can reduce the impact of different spectra from the same substance and the same spectrum from different substances. Additionally, visual characteristics would be preserved better by combining spectral and image information because of higher spatial resolution.

Gao and Xu [22] compared single spectral information to combination of spectral and texture colour information in soluble solid content in red earth grapes and showed that combination of spectral and image information effectively improved the model's detection capability. Xu et al. [23] applied the spectral and image information combination method to detect nitrogen content in rice leaves at different growth stages. The results showed that a combination of spectral and image information reduced interference from soil and water, and R^2 increased by 0.05 to 0.09, while RMSE decreased by 0.011 to 1 for various models.

Accuracy of soybean saponin content detection was not high in previous studies. That is because the content of saponin in soybean is small and single spectral information is insufficient to express saponin. In this paper, spectral and image information in hyperspectral data was combined to improve the accuracy of the soybean saponin content detection model. Spectral preprocessing and spectral data feature band selection, ensemble learning model with skip connect, and multihead self-attention mechanisms were studied to further improve model accuracy.

2. Materials and Methods

2.1. Test Materials and Data Acquisition

2.1.1. Soybean Sample. Soybean samples are provided by the Agriculture College of Northeast Agricultural University. Ten types of soybeans are selected including Beidou 5, Sui 03-3952, Hongfeng 3, Chundou 1, Dongnong 60, Beidou 14, Huajiang 1, L-58Keburi, Zhongpin 03-5373, and Dongnong 50. 30 samples without defects are collected for each variety, and total 300 soybean samples are used in this paper. All samples are stored in a cool and well-ventilated place. The spectral-physicochemical value cooccurrence distance (SPXY) algorithm is used to divide samples into the training set and test set in ratio of 7 : 3.

2.1.2. Spectral and Image Data Acquisition. Hyperspectral data of soybean samples are collected by Hyperspec III hyperspectral imager from HEADWALL Company. Samples are placed on the tray of moving platform with moving speed of $3.5 \text{ mm}\cdot\text{s}^{-1}$, and camera exposure time is 38.84 ms. Reflection values of soybean samples are used as spectral information with 495 bands from 463 nm to 957 nm. The RGB image of soybean is output at the same time.

After collection of spectral data and image data, spectral reflectance values of each sample are extracted by ENVI 5.3 from regions of interest (ROI). Although reflectance rates may vary across different positions on soybean, overall trend remains consistent, which does not affect subsequent modelling results [24]. In this study, ROI is selected as rectangular areas with a size of 10×10 pixels in the soybean centre. The obtained spectral reflectance value is adjusted using the black and white correction method to obtain accurate spectral reflectance. The formula for black and white correction is as follows:

$$R = \frac{I - B}{w - B}, \quad (1)$$

in the equation, R represents soybean spectral reflectance, I represents soybean spectral reflection values, w represents the spectral reflection value of the whiteboard, represents the spectral reflection value of the blackboard.

2.1.3. Determination of Soybean Saponin Content. In this study, soybean saponin content is measured using liquid chromatography-mass spectrometry (LC-MS). 100 mg sample is placed in a 1.5 mL centrifuge tube and mixed with $300 \mu\text{L}$ of 75% methanol/water mixed solvent (containing 0.1% formic acid, v/v). The mixture is vortexed for 30 seconds and subjected to ultrasound treatment at 20°C for 15 minutes in a water bath. After vortexing for additional 2 minutes, the sample is centrifuged at 12,000 rpm at 4°C for 20 minutes. Then, we take the supernatant and test on the machine. Acquired mass spectrometry raw data are processed using Agilent Profinder software. Data processing steps include retention time correction, peak identification, peak extraction, peak integration, and peak alignment. Agilent Massive Parallel Processor software is used for statistical processing and combined with the KEGG database, and substance identification is conducted to determine saponin content.

2.2. Image Processing and Feature Extraction. A flowchart of image processing procedure is illustrated in Figure 1. Firstly, the acquired image is converted into a grayscale image. Subsequently, the nonlocal means denoising algorithm and the Gaussian filter are applied to eliminate noise in the grayscale image to facilitate subsequent edge detection research. Finally, soybean contour is extracted using the adaptive threshold algorithm [25].

After edge detection, feature information of soybean samples is extracted including area, perimeter, major axis, minor axis, roundness, eccentricity, aspect ratio, rectangle-area ratio, circle-area ratio, equal-area-circle diameter, and edge variation coefficient. Meanings of this feature information are described in Table 1.

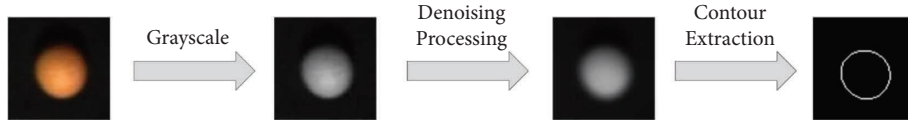


FIGURE 1: Schematic diagram of image processing.

TABLE 1: Meaning of soybean image feature information.

Feature information	Meaning
Area	Number of pixels in the soybean outline area
Perimeter	Number of pixels of the soybean contour curve
Major axis	Number of pixels of the major axis of the minimum circumscribed ellipse of soybean
Minor axis	Number of pixels of the minor axis of the minimum circumscribed ellipse of soybean
Roundness	$R = 4\pi S/L^2$. Among them, R represents roundness, and it is similar to the circle if R is closer to 1; L represents the perimeter of the soybean outline, π represents the pi ratio, and S represents the soybean area
Eccentricity	Eccentricity of the smallest circumscribed ellipse of soybean
Aspect ratio	Ratio of the major axis and minor axis of the smallest circumscribed ellipse of soybean
Rectangle-area ratio	Ratio of the area of the smallest circumscribed rectangle of soybean to the area of soybean
Circle-area ratio	Ratio of the smallest circumscribed circle area to the soybean area
Equal-area-circle diameter	Diameter of a circle equal to the soybean area
Edge variation coefficient	Ratio of standard deviation and mean value of each pixel point from the center of soybean image gravity to its edge

2.3. Spectral Data Preprocessing. Spectral data preprocessing is essential for minimizing errors during model building. Noise information would be included in spectral data such as sample background, dispersive light, signal noise, and so on. To reduce impact of abovementioned unrelated factors on the detection model and enhance spectral characteristics, removing noise information by spectral preprocessing is necessary. Common preprocessing methods include Savitzky–Golay smoothing (S-G), SNV, de-trending (DT), multiplicative scatter correction (MSC), baseline correction (BL), first derivative (FD), and second derivative (SD). These methods effectively eliminate noise from different perspectives and highlight characteristics of spectral data [26–29]. In this study, these methods are applied into soybean saponin detection and the selected suitable algorithm.

2.4. Dimensionality Reduction of Spectral Features. Full-band spectra contain a lot of redundant information, which makes the detection model complex and inaccurate. In order to reduce data dimension and obtain main characteristic bands of the spectrum, spectral data feature band selection should be done before model building. Because of selecting several spectral intervals related to the tested substance, models based on the SiPLS band selection algorithm are usually with high accuracy [30]. Extracted spectral features are stable and continuous, but there is still some redundant spectral information. IRIV iteratively reduces spectral data dimensionality by iteratively building informative variables and keeping spectral feature wavelengths with high weights in feature subsets. After removing

the irrelevant variables and interference variables, the last group of variables is processed by reverse elimination to obtain more simplified spectral characteristic wavelength [31]. However, due to multiple iterations, applying IRIV to the full band spectral set needs large calculation. In this study, method combination SiPLS and IRIV was proposed to select the spectral data feature band. Selecting valid intervals by SiPLS before selecting variables by IRIV can improve model fitting ability.

2.5. Data Combination. Due to distinct feature attributes of spectral and image information in soybeans, normalization is necessary before using spectral and image information as combination input for the detection model. Spectral and image information is scaled to map both datasets to [0, 1] interval [32]. Min-max normalization is used for data combination processing and its formula is as follows:

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}}, \quad (2)$$

in the equation, x^* represents transformed data, original x represents feature data, x_{\max} represents maximum characteristic data, and x_{\min} represents minimum characteristic data.

2.6. Model Construction and Evaluation. The two-layer stacking ensemble learning framework is built to detect soybean saponin content. The framework consists of three base learners including DF, PLSR, and EXG. Meta-learner is PBT-BPNN which is the learning rate of BPNN was

optimized by the population-based training (PBT) method. Effects of spectral information and combination information are compared in this study. The ensemble learning model construction process is illustrated in Figure 2.

Skip connect [33] and multihead self-attention mechanism [34] are introduced into BPNN in this paper, in order to enhance generalization and the expressive ability of BPNN, as well as prevent overfitting and gradient explosion in deep networks. BPNN with an optimized hidden layer can improve the detection capability of the network because of its deeper network architecture. The optimized structure of the hidden layer is shown in Figure 3. The input data sequence of this layer obtains a set of sequences by traditional BPNN nonlinear computation, and element relationships of the sequences are captured by the multihead self-attention mechanism. This ensures that no information was forgotten by the neural network after computations of multiple hidden layers, thus preventing data loss and decline in the model detection ability. The original sequence is added by skip connect to enhance the neural network's memory of the original input sequence. It can avoid weight reduction of information after multiple nonlinear transformations and solve the gradient explosion phenomenon and network performance degradation phenomenon in deep neural networks.

The structure of the multihead self-attention mechanism is shown in Figure 4. Given an input sequence, it is multiplied with three trainable parameter matrices to yield three vectors, query, key, and value. Three vectors are multiply processed by parallel computations using self-attention with distinct parameters for each computation, in order to emphasize different features of the sequence. Finally, results of multiple computations are concatenated and linearly transformed. Three vectors are first linearly transformed for each computation. Then, transformed query and key vectors are multiplied to compute attention scores. The scores are scaled and multiplied with mask for each element to prevent excessively attending to certain elements during training. The process can also enhance the generalization capability of the model. Attention weights are obtained by inputting processed attention scores to SoftMax and multiplied to the corresponding value vector and summed to be output.

Model performance is evaluated based on R^2 , RMSE, and ratio of prediction to deviation (RPD) of the model for test set. R^2 represents the fitness of the model and indicates higher fitness when its value is closer to 1, i.e., independent variables can better explain the variation of the dependent variable. RMSE is commonly used to evaluate the model error and represents higher model accuracy and less error when its value is closer to 0. RPD is an academic metric that quantifies the predictive accuracy and reliability of a model. When the RPD is less than 2.0, the model is generally considered incapable of reliable quantitative prediction; when the RPD ranges between 2.0 and 2.5, the model can make rough quantitative predictions; and when the RPD exceeds 2.5, the model is deemed to have good predictive accuracy. The result takes the average value of three tests. R^2 , RMSE, and RPD are mathematically formulated by equations (3)–(5).

$$R^2 = 1 - \frac{\sum_{i=1}^m (y_i - \hat{y})^2}{\sum_i^m (y_i - \bar{y})^2}, \quad (3)$$

$$\text{RMSE} = \sqrt{\frac{1}{m} \sum_{i=1}^m (y_i - \hat{y}_i)^2}, \quad (4)$$

$$\text{RPD} = \sqrt{\frac{\sum_{i=1}^m (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^m (\hat{y}_i - y_i)^2}}, \quad (5)$$

in the equation, m represents the sample size, y_i represents the actual value, \hat{y} represents the model calculated value, and \bar{y} represents the average value of actual values.

3. Results and Discussion

3.1. Dataset Partitioning Results. SPXY is a kind of dataset partitioning algorithm based on the Kennard Stone (KS) algorithm, which is optimized by combined spectral data (X) and chemical values (Y) in sample distance calculation. SPXY enhances model robustness and reduces regression errors because of covering multidimensional space effectively. It also mitigates the impact of imperfect dataset partitioning on final results [35]. Dataset partitioning results by SPXY are shown in Table 2 and Figure 5. From Table 2 and Figure 5, it can be observed that the soybean saponin content of the test set fell within the range of the training set. It indicated that samples were representative and SPXY method for dataset partitioning was rational.

3.2. Image Information Correlation Analysis. Correlation analysis results between soybean image feature information and soybean saponin content are presented in Table 3. Correlations between each image feature and soybean saponin content were different. Among eleven image features, the largest absolute correlation coefficient was observed for rectangle-area ratio, which was 0.289. Despite being highly significant, the Pearson correlation coefficient was relatively small. This may be due to the fact that the Pearson correlation coefficient measures linear relationships between two variables, and a larger coefficient indicates stronger linear correlation. However, if relationship between two variables is influenced by other variables, there may exist a nonlinear relationship between them. Therefore, image features such as rectangularity could exhibit a joint nonlinear relationship with soybean saponin content, leading to a lower Pearson correlation coefficient despite with highly significant correlation.

Results in Table 3 revealed that image features were highly significant related to soybean saponin content with $P < 0.01$ except for area, perimeter, minor axis, and equal-area-circle diameter. Specifically, major axis and rectangle-area ratio showed highly significant positive correlation with soybean saponin content, while roundness, eccentricity, aspect ratio, circle-area ratio, and edge variation

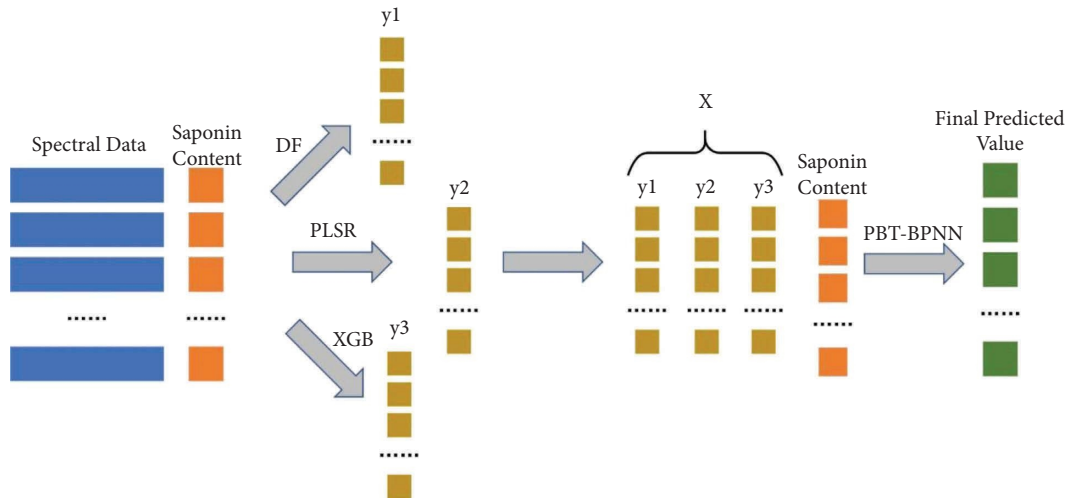


FIGURE 2: Schematic diagram of the ensemble learning principle. X represents the input, and y represents the output.

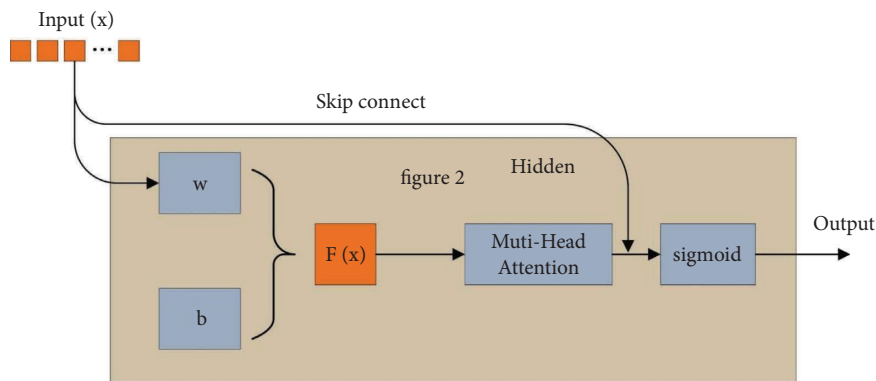


FIGURE 3: Structure of the hidden layer after optimization. W represents the weight matrix, and b represents the bias matrix. $F(x)$ represents the output.

coefficient exhibited highly significant negative correlation with soybean saponin content. High significance with $P < 0.01$ indicates that the independent variables have a notable impact on the dependent variables, and this impact is extremely unlikely to be explained by random errors statistically. Therefore, seven image features, including the major axis, roundness, eccentricity, aspect ratio, rectangle-area ratio, circle-area ratio, and edge variation coefficient, are closely correlated with the saponin content of soybeans. As a result, we have chosen these seven image features for combination with spectral information.

3.3. Spectral Reflectance Extraction Results. Figure 6 shows original spectral reflectance intensities of 300 soybean samples in 463–957 nm wavelength range. Figure 7 presents spectral reflectance values of soybean samples after black and white correction. In these two figures, each colour line represents a soybean sample. Spectral data of all samples exhibited an increasing trend in range 463–760 nm. Absorption valleys in 760–820 nm corresponded to the three types of soybean samples including Chundou 1, L-58Keburi, and Zhongpin 03-5373. Spectral reflectance of soybean samples gradually became stable in range 850–957 nm.

3.4. Result of Spectral Preprocessing. To reduce the influence of noise, stray light, and other irrelevant factors on spectral data and improve signal-to-noise ratio, various methods were applied to preprocess soybean spectral data in this study. PLSR can reliably perform regression modelling even with variables with less correlation and multicollinearity independent variables because of combining characteristics of principal component analysis, canonical correlation analysis, and linear regression analysis. PLSR has been widely used in many studies to compare the effectiveness of different spectral preprocessing methods. So, S-G, SNV, DT, MSC, BL, FD, and SD were compared for soybean saponin detection based on the PLSR model. R^2 and RMSE of the model for the test set were used to evaluate effects. Results are shown in Table 4.

According to Table 4, the PLSR model based on S-G, BL, and SD had negative effects with decreased R^2 and increased RMSE compared to original data. Important information was maybe removed with spectral preprocessing.

The detection model based on the MSC method was with higher R^2 , but also higher RMSE compared the detection model by original data. This indicated that the model had a stronger capability to explain the target variable after

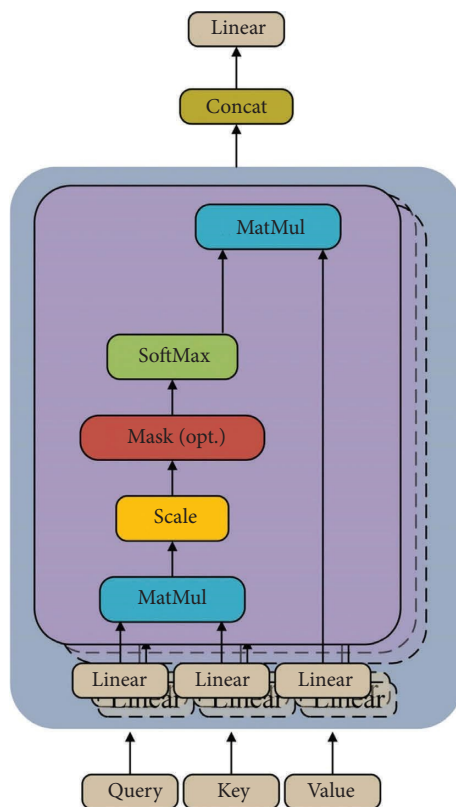


FIGURE 4: Structure of multihead self-attention mechanism.

TABLE 2: Dataset partitioning results by SPXY.

Items	Number of samples	Maximum value (g·100 g ⁻¹)	Minimum value (g·100 g ⁻¹)	Average value (g·100 g ⁻¹)	Standard deviation (g·100 g ⁻¹)	Standard error (g·100 g ⁻¹)
Dataset	300	4.46	1.41	2.78	0.55	0.032
Training dataset	210	4.46	1.41	2.84	0.56	0.039
Test dataset	90	4.42	1.49	2.76	0.54	0.057

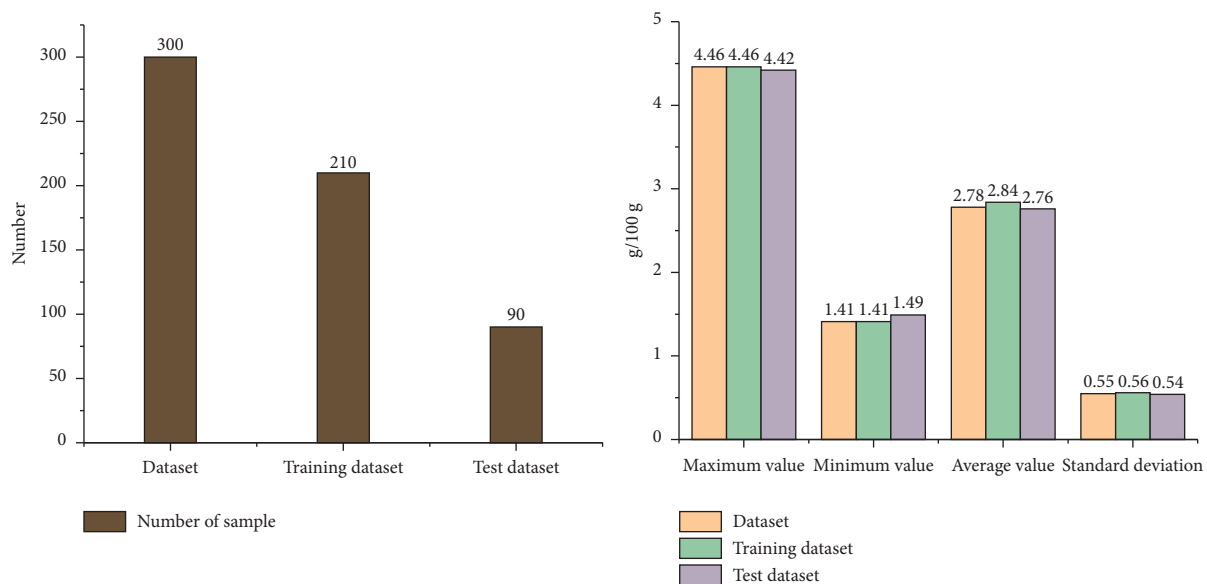


FIGURE 5: Dataset partitioning results by SPXY.

TABLE 3: Correlation analysis between different image feature information and soybean saponin content.

Feature information	Pearson correlation	Significant differences
Area	0.058	0.319
Perimeter	0.056	0.030
Major axis	0.153**	0.008
Minor axis	0.007	0.903
Roundness	-0.246**	0.000
Eccentricity	-0.209**	0.000
Aspect ratio	-0.205**	0.000
Rectangle-area ratio	0.289**	0.000
Circle-area ratio	-0.233**	0.000
Equal-area-circle diameter	0.077	0.183
Edge variation coefficient	-0.276**	0.000

Note. **significantly correlated at the $P < 0.01$ level.

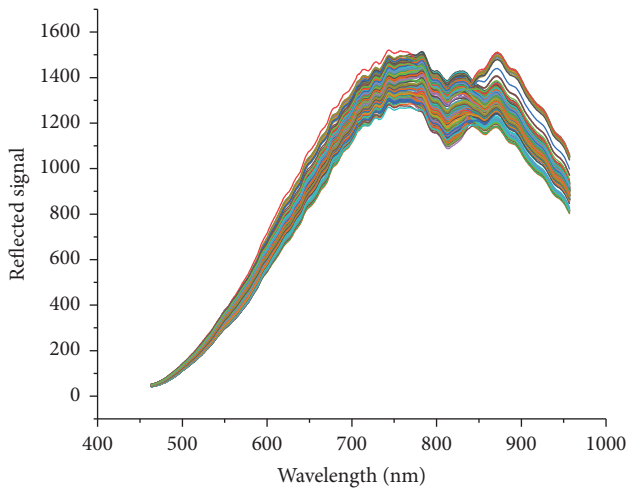


FIGURE 6: Spectral reflection intensity of samples.

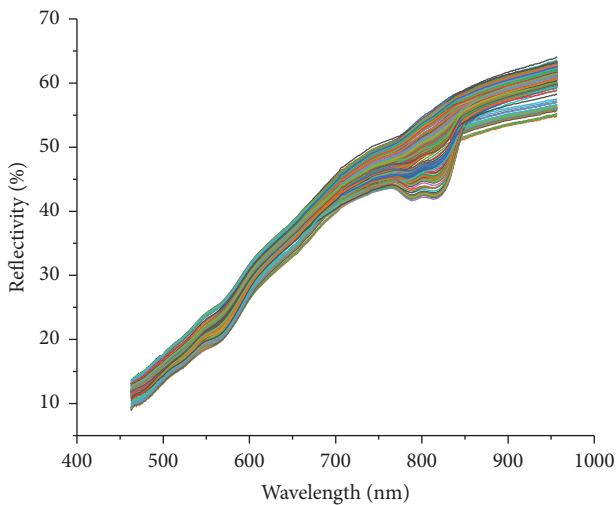


FIGURE 7: Spectral reflectance after black and white correction of samples.

TABLE 4: PLSR model results for different preprocessing methods.

Preprocessing method	R^2	RMSE
None	0.5456	3.7539×10^{-3}
S-G	0.5076	3.9076×10^{-3}
SNV	0.6392	3.4510×10^{-3}
DT	0.5838	3.5924×10^{-3}
MSC	0.6268	4.2473×10^{-3}
BL	0.5384	3.7835×10^{-3}
FD	0.5737	3.6359×10^{-3}
SD	0.5392	3.7815×10^{-3}

spectral data preprocessing. However, the error of the detection model was increased, resulting in lower accuracy. This could be attributed to loss of some important features in spectral data after preprocessing.

Models based on SNV, DT, and FD were with higher R^2 and lower RMSE. Fitting capability and accuracy of models were optimized by three valid preprocessing methods. Models based on the SNV preprocessing method had the highest R^2 and the lowest RMSE. The SNV method was chosen to preprocess spectral data in this paper. Subsequent variable selection and modelling processes utilize spectral data that has been preprocessed by SNV.

The PLSR model evaluates the efficacy of spectral preprocessing algorithms. However, the performance of these preprocessing algorithms in ensemble learning and optimized ensemble learning models may not align with their performance in the PLSR model. To validate the effectiveness of the PLSR model, we reevaluate the performance of the preprocessing algorithms using ensemble learning and optimized ensemble learning. If the resulting trends align with those observed in the PLSR model, it demonstrates the validity of the PLSR model in assessing the effectiveness of preprocessing algorithms. The ensemble learning model and optimized ensemble learning model were used to evaluate spectral preprocessing performance. Results are shown in Table 5.

Results based on the ensemble learning model and optimized ensemble learning model were the same as the PLSR model. Models based on S-G, BL, and SD exhibited poor performance. Models based on MSC only improved the R^2 without reducing RMSE. On the other hand, models based on SNV, DT, and FD methods enhanced model performance comprehensively. Models based on the SNV method achieved the highest R^2 and the lowest RMSE. This validation experiment demonstrated that model building methods would not affect results of preprocessing algorithms. That is because the purpose of spectral preprocessing is removing noisy information. So, the PLSR model can be used to evaluate a better spectral preprocessing method.

3.5. Dimensionality Reduction of Spectral Features. SiPLS can select spectral wavelength intervals containing spectral features. Figure 8 shows the RMSE of models by spectral wavelength interval combinations when SiPLS with 50 interval divisions and 3 combinations. The lowest RMSE is obtained for the 132nd wavelength interval combination, which was 3.2647×10^{-3} . Spectral wavelength intervals for

TABLE 5: Ensemble learning model and optimized ensemble learning model results for different processing methods.

Preprocessing method	Ensemble learning- R^2	Ensemble learning: RMSE	Optimized ensemble learning: R^2	Optimized ensemble learning: RMSE
None	0.6725	3.3917×10^{-3}	0.7703	2.5629×10^{-3}
S-G	0.6537	3.5021×10^{-3}	0.7548	2.7155×10^{-3}
SNV	0.7428	3.2647×10^{-3}	0.8174	2.4316×10^{-3}
DT	0.7115	3.3879×10^{-3}	0.8057	2.4512×10^{-3}
MSC	0.7404	3.7659×10^{-3}	0.7991	3.0524×10^{-3}
BL	0.6593	3.4995×10^{-3}	0.7686	2.7014×10^{-3}
FD	0.7023	3.4157×10^{-3}	0.7862	2.5261×10^{-3}
SD	0.6652	3.4522×10^{-3}	0.7689	2.7013×10^{-3}

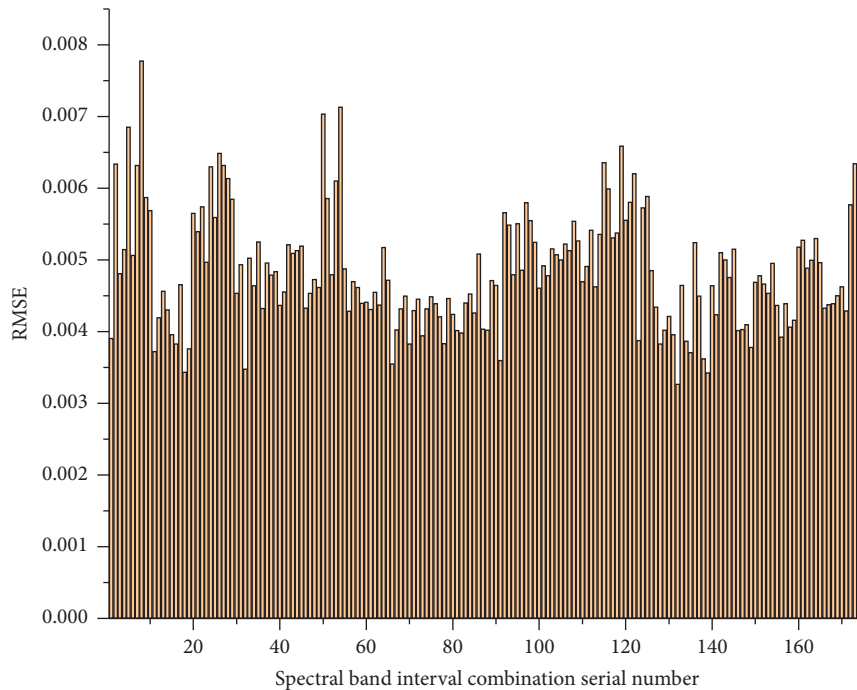


FIGURE 8: RMSE of models for different spectral band interval combinations.

the combination were 561~658 nm, 708~757 nm, and 858~907 nm, with a total of 149 spectral wavelengths, accounting for 30.1% of total wavelengths.

149 spectral wavelengths selected by SiPLS were used as input variables for the IRIV algorithm. As shown in Figure 9, irrelevant and interfering variables were eliminated from the variable combinations after 4 times of iterations, resulting in the backward elimination of 11 variables. Finally, 17 feature spectral variables related to soybean saponin content were selected including 562 nm, 575 nm, 579 nm, 597 nm, 606 nm, 710 nm, 737 nm, 739 nm, 743 nm, 836 nm, 842 nm, 847 nm, 851 nm, 852 nm, 853 nm, 854 nm, and 855 nm. Variables remained were only 3.43% of the total wavelengths. Distribution of selected spectral feature wavelengths by SiPLS-IRIV is shown in Figure 10.

3.6. Result of Ensemble Learning Modelling. Ensemble learning can reduce the risk of model overfitting and improve robustness, reliability, and accuracy of the model. Results of the detection model are shown in Table 6.

According to Table 6, six models showed a significant improvement compared to previous studies. The PLSR model with single spectral information was with the lowest R^2 of 0.7195, the highest RMSE of 3.0639×10^{-3} , and the lowest RPD of 1.8881. RPD of this model was less than 2.0. This indicated that the model possessed quantitative predictive ability, albeit not particularly outstanding. The residual attention ensemble learning model by using combined spectral image information was with the highest R^2 of 0.9216, the lowest RMSE of 1.7071×10^{-3} , and the highest RPD of 3.5714. RPD of this model was higher than 2.5. This indicated that the model possessed exceptional predictive ability, exhibiting stable and accurate performance on the test dataset. These results indicated that single spectral information had limitations in describing soybean saponin content and cannot effectively detect saponin content in soybeans. By supplementing image feature information, input dimensionality not captured by spectral data was enriched. More comprehensive descriptions of saponin content were enabled in the model. The ensemble learning model and optimized ensemble learning model

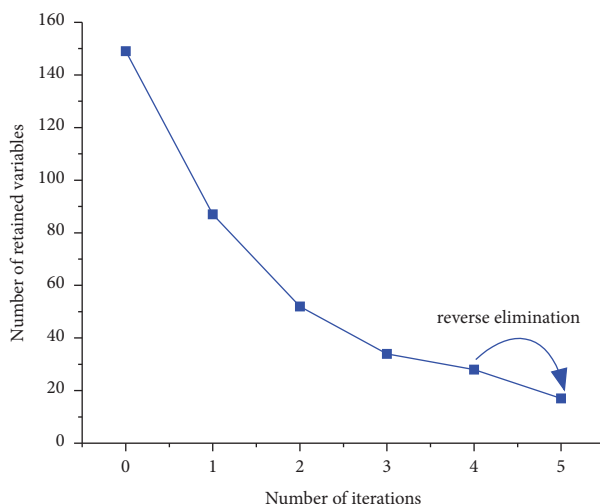


FIGURE 9: IRIV iteration process.

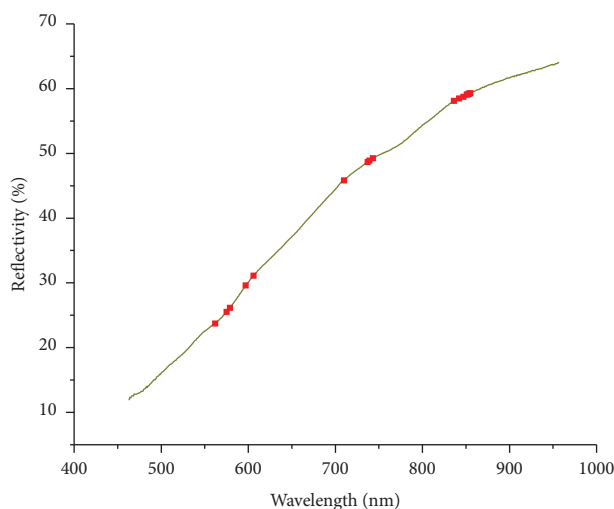


FIGURE 10: Spectral feature wavelength distribution by SiPLS-IRIV.

TABLE 6: Prediction results based on different modelling methods.

Model	Input	R^2	RMSE	RPD
PLSR	Single spectral information	0.7195	3.0639×10^{-3}	1.8881
Ensemble learning	Single spectral information	0.7942	2.5261×10^{-3}	2.2043
Optimized ensemble learning	Single spectral information	0.8406	2.3118×10^{-3}	2.5047
PLSR	Spectral image combination information	0.7804	2.9747×10^{-3}	2.1339
Ensemble learning	Spectral image combination information	0.8483	2.1685×10^{-3}	2.5675
Optimized ensemble learning	Spectral image combination information	0.9216	1.7071×10^{-3}	3.5714

have more complex structures compared to PLSR, which allow performing more accurate nonlinear transformations on the input information and resulting in improved saponin detection performance. Skip connect and multihead self-attention modules were added to optimize the ensemble learning model. The multihead self-attention module enabled a better capture of relationships between input sequences in each hidden layer and given more weight to elements that were highly correlated with saponin content. Skip connect ensured that elements with smaller forget

weights were not neglected during nonlinear transformations in hidden layers. It preserved the integrity of element information as the number of hidden layers increasing, enhanced the generalization ability of the model, and improved the accuracy of the soybean saponin content detection model.

A scatter plot of predicted values and measured values for test set is shown in Figure 11, with predicted values on the vertical axis and measured values on the horizontal axis. The linear function expressed in the figure is $y=x$, with

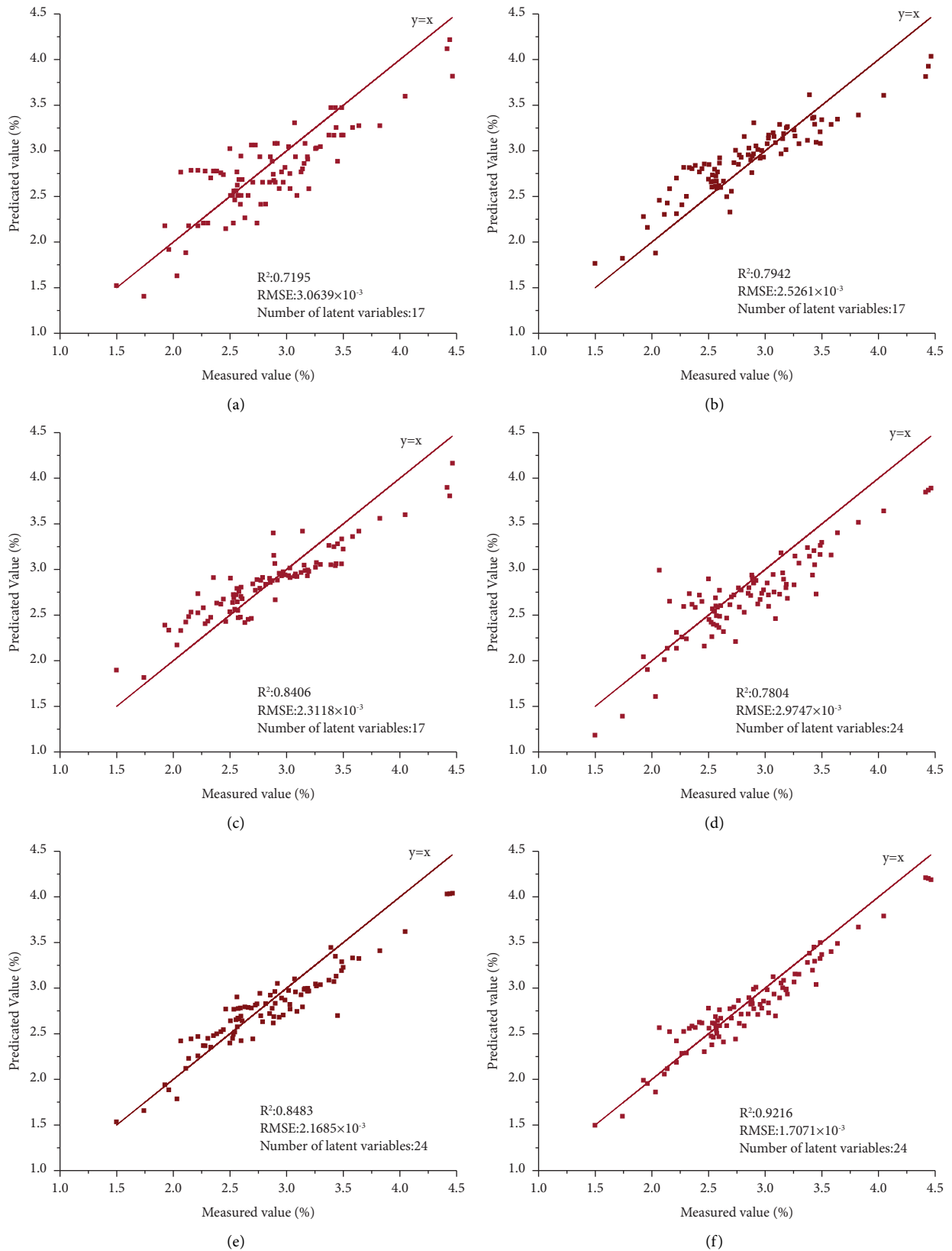


FIGURE 11: Scatter plot of predicted and measured values for the test set. A 1 : 1 line was added to the map. (a) PLSR with a single spectral information input. (b) Ensemble learning with a single spectral information input. (c) Optimized ensemble learning with a single spectral information input. (d) PLSR with a spectral image combination information input. (e) Ensemble learning with a spectral image combination information input. (f) Optimized ensemble learning with a spectral image combination information input.

a slope of 1 and a bias of 0. The closer the scattered point is to the line $y=x$, the smaller the error between the predicted value and the measured value. Conversely, the farther the scattered points are from the line $y=x$, the greater the error between the predicted value and the measured value. It intuitively demonstrates that the optimized ensemble learning model established by combined spectral image information was better. The first three figures depicted three models that were established by single spectral information. The PLSR model (Figure 11(a)) exhibited poor fit for soybean samples. Notably, for a sample with a measured saponin content of 2.1%, the model predicted a value of 2.6%, indicating significant deviation. Overall, there was considerable dispersion among the scattered points. The ensemble learning model (Figure 11(b)) demonstrated relatively more concentrated scattered points compared to the PLSR model. However, it tended to overestimate the saponin content, with most predicted values exceeding the measured values. The ensemble model optimized with the residual attention module (Figure 11(c)) performed well within the saponin content range of 2.5% to 3.0%, where predicted values closely aligned with measured values. Nevertheless, its performance deteriorated when dealing with higher saponin contents, particularly exceeding 4%. The models presented in the last three figures were established by combination information. Compared to models with single spectral information, these models demonstrated a significant improvement in the fitting ability. A scatter plot for the PLSR model (Figure 11(d)) tended to have a relatively concentrated distribution around the line $y=x$. However, the fitting performance was still unsatisfactory. The ensemble learning model's scatter plot (Figure 11(e)) showed only a few points with poor predictions, while the majority of the points were scattered around the $y=x$ line. The scatter plot of the ensemble model optimized with the residual attention module (Figure 11(f)) resembled a diagonal line, indicating minimal errors. Especially in the range of soybean saponin content from 3.5% to 4.5%, the other five models did not fit well and showed significant deviations from the measured values. However, the optimized ensemble learning model demonstrated minimal errors within this range. Overall, the residual attention ensemble learning model with combined spectral image information can accurately estimate soybean saponin content.

Through a comparative analysis of the results of this experiment and those of other research methods, it can be seen that the method employed in this research demonstrates remarkable superiority on multiple levels. Unlike previous methods based on near-infrared spectroscopy, our study leverages the unique spectral-imaging capabilities of hyperspectral technology. This allows us to simultaneously capture spectral and image data for multiple soybeans, enabling concurrent detection of multiple samples. Furthermore, the integration of hyperspectral precision with the optimized stacking ensemble learning model yields more accurate detection of soybean saponin content than ever before. When compared to traditional wet chemical methods, our approach not only ensures high-precision detection results but also excels in cost control, detection

speed, and sample preservation. Specifically, this method is cost-effective. It significantly reduces detection time. Importantly, it eliminates the need to destroy soybean samples, preserving their integrity. Additionally, it removes the influence of human proficiency on test results. Thus, our research offers a novel approach for accurate and efficient detection of soybean saponin content.

4. Conclusions

The soybean saponin content detection model based on spectroscopy and image information combination was developed in this paper. SNV was selected as the spectral preprocessing method. SiPLS-IRIV was used to perform dimensionality reduction. The ensemble learning model with skip connect and multihead self-attention modules was built to detect soybean saponin content. R^2 and RMSE values of the model were 0.9216 and 1.7071×10^{-3} . The detection method based on hyperspectral technology reduced sample processing time, improved detection efficiency, avoided sample damage, and minimized experimental errors caused by human operators. This study provides a new method for researchers in soybean breeding quality testing, making the process more efficient and convenient.

Data Availability

The raw/processed data required to reproduce these findings cannot be shared at this time as the data also form part of an ongoing study.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors would like to thank the Key Laboratory of Northeast Smart Agricultural Technology, Ministry of Agriculture and Rural Affairs.

References

- [1] Y. Shang, "The future development direction of green building energy-saving new materials," *Sichuan Cement*, vol. 301, no. 9, pp. 63-64, 2021.
- [2] H. N. Cheng, W. Wyckoff, M. K. Dowd, and Z. He, "Evaluation of adhesion properties of blends of cottonseed protein and anionic water-soluble polymers," *Journal of Adhesion Science and Technology*, vol. 33, no. 1, pp. 66-78, 2019.
- [3] E. Rekiel, W. Smulek, A. Zdziennicka, E. Kaczorek, and B. Jańczuk, "Wetting properties of saponaria officinalis saponins," *Colloids and Surfaces A: Physicochemical and Engineering Aspects*, vol. 584, 2020.
- [4] M. Yin, "Research progresses of soyasaponin," *Journal of Qilu University of Technology*, vol. 32, no. 6, pp. 34-38, 2018.
- [5] W. Zhang and D. G. Popovich, "Effect of soyasapogenol A and soyasapogenol B concentrated extracts on Hep-G2 cell proliferation and apoptosis," *Journal of Agricultural and Food Chemistry*, vol. 56, no. 8, pp. 2603-2608, 2008.

- [6] J. Wu, M. Jing, and M. Jin, "Effects of soybean saponins and soybean isoflavones on serum antioxidant capacity in aging mice," *China health care & nutrition*, vol. 26, no. 20, pp. 285–286, 2016.
- [7] T. Nagano, M. Katase, and K. Tsumura, "Dietary soyasaponin attenuates 2, 4-dinitrofluorobenzene-induced contact hypersensitivity via gut microbiota in mice," *Clinical and Experimental Immunology*, vol. 195, no. 1, pp. 86–95, 2018.
- [8] H. Nakashima, K. Okubo, Y. Honda, T. Tamura, S. Matsuda, and N. Yamamoto, "Inhibitory effect of glycosides like saponin from soybean on the infectivity of HIV in vitro," *AIDS*, vol. 3, no. 10, pp. 655–658, 1989.
- [9] Z. Tavassoli, M. Taghdir, and B. Ranjbar, "Renin inhibition by soyasaponin I: a potent native anti-hypertensive compound," *Journal of Biomolecular Structure and Dynamics*, vol. 36, no. 1, pp. 166–176, 2018.
- [10] Y. G. Lin, G. W. Meijer, M. A. Vermeer, and E. A. Trautwein, "Soy protein enhances the cholesterol-lowering effect of plant sterol esters in cholesterol-fed hamsters," *The Journal of Nutrition*, vol. 134, no. 1, pp. 143–148, 2004.
- [11] S. Kamo, Y. Takada, T. Yamashita et al., "Group B soyasaponin aglycone suppresses body weight gain and fat levels in high fat-fed mice," *Journal of Nutritional Science and Vitaminology*, vol. 64, no. 3, pp. 222–228, 2018.
- [12] Y. Huang, M. Yan, and D. Zhao, "Large-scale isolation and preparation of soybean saponin by high-speed countercurrent chromatography combined with preparative HPLC," *Food Science*, vol. 34, no. 6, pp. 27–32, 2013.
- [13] P. Xue, L. Zhao, X. Zheng, F. Zhang, and G. Ren, "Chemical structure and analysis methods of soybean saponins: a review," *Modern Food Science and Technology*, vol. 34, no. 9, pp. 291–297, 2018.
- [14] G. Sagratini, G. Caprioli, F. Maggi et al., "Determination of soyasaponins I and β g in raw and cooked legumes by solid phase extraction (SPE) coupled to liquid chromatography (LC)–Mass spectrometry (MS) and assessment of their bioaccessibility by an in vitro digestion model," *Journal of Agricultural and Food Chemistry*, vol. 61, no. 8, pp. 1702–1709, 2013.
- [15] X. Ma and Y. Wang, "Study on influencing factors of accuracy of chemical analysis results," *China Petroleum and Chemical Standard and Quality*, vol. 43, no. 9, pp. 49–51, 2023.
- [16] Z. Guo, J. Zhang, C. Ma 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*, vol. 116, 2023.
- [17] Y. Song, S. Cao, X. Chu et al., "Non-destructive detection of moisture and fatty acid content in rice using hyperspectral imaging and chemometrics," *Journal of Food Composition and Analysis*, vol. 121, 2023.
- [18] G. Zhang, P. Li, W. Zhang, and J. Zhao, "Analysis of multiple soybean phytonutrients by near-infrared reflectance spectroscopy," *Analytical and Bioanalytical Chemistry*, vol. 409, no. 14, pp. 3515–3525, 2017.
- [19] M. A. Berhow, M. Singh, M. J. Bowman, N. P. J. Price, S. F. Vaughn, and S. X. Liu, "Quantitative NIR determination of isoflavone and saponin content of ground soybeans," *Food Chemistry*, vol. 317, 2020.
- [20] L. Zheng, M. Zhao, J. Zhu et al., "Fusion of hyperspectral imaging (HSI) and RGB for identification of soybean kernel damages using ShuffleNet with convolutional optimization and cross stage partial architecture," *Frontiers in Plant Science*, vol. 13, 2022.
- [21] W. Wang, W. Yang, P. Zhou, Y. Cui, D. Wang, and M. Li, "Development and performance test of a vehicle-mounted total nitrogen content prediction system based on the fusion of near-infrared spectroscopy and image information," *Computers and Electronics in Agriculture*, vol. 192, 2022.
- [22] S. Gao and J. Xu, "Hyperspectral image information fusion-based detection of soluble solids content in red globe grapes," *Computers and Electronics in Agriculture*, vol. 196, 2022.
- [23] S. Xu, X. Xu, C. Blacker et al., "Estimation of leaf nitrogen content in rice using vegetation indices and feature variable optimization with information fusion of multiple-sensor images from UAV," *Remote Sensing*, vol. 15, no. 3, p. 854, 2023.
- [24] D. Fu, J. Zhou, A. M. Scaboo, and X. Niu, "Nondestructive phenotyping fatty acid trait of single soy-bean seeds using reflective hyperspectral imagery," *Journal of Food Process Engineering*, vol. 44, no. 8, 2021.
- [25] S. Kukunin, "The application of morphological algorithms for the restoration and segmentation of graphic data for the machine vision system," *Global Prosperity*, vol. 3, no. 1, pp. 49–56, 2023.
- [26] Y. Peng, C. Sun, and M. Zhao, "Dynamic nondestructive sensing and grading manipulator system for apple quality," *Transactions of the Chinese Society of Agricultural Engineering*, vol. 38, no. 16, pp. 293–303, 2022.
- [27] W. Zhu, Z. Feng, S. Wu et al., "Development of an airborne non-contact near-infrared soil moisture detection system," *Transactions of the Chinese Society of Agricultural Engineering*, vol. 38, no. 9, pp. 73–80, 2022.
- [28] Y. Li, G. Liu, N. Fan et al., "A combination of hyperspectral imaging with two-dimensional correlation spectroscopy for monitoring the hemicellulose content in lingwu long jujube," *Spectroscopy and Spectral Analysis*, vol. 42, no. 12, pp. 3935–3940, 2022.
- [29] S. Zhao, H. Yu, G. Gao et al., "Rapid determination of protein components and their subunits in peanut based on near infrared technology," *Spectroscopy and Spectral Analysis*, vol. 41, no. 03, pp. 912–917, 2021.
- [30] Y. Zhu, X. Zou, J. Shi, J. Zhao, and H. Lin, "Rapidly detecting total acid distribution of vinegar culture based on hyperspectral imaging technology," *Transactions of the Chinese Society of Agricultural Engineering*, vol. 30, no. 16, pp. 320–327, 2014.
- [31] Y. Xu, H. Zhang, C. Zhang et al., "Rapid prediction and visualization of moisture content in single cucumber (*Cucumis sativus* L.) seed using hyperspectral imaging technology," *Infrared Physics & Technology*, vol. 102, 2019.
- [32] W. Wang, W. Yang, P. Zhou, Y. Cui, D. Wang, and M. Li, "Development and performance test of a vehicle-mounted total nitrogen content prediction system based on the fusion of near-infrared spectroscopy and image information," *Computers and Electronics in Agriculture*, vol. 192, 2022.
- [33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.
- [34] J. Mercat, T. Gilles, N. El Zoghby, G. Sandou, D. Beauvois, and G. P. Gil, "Multi-head attention for multi-modal joint vehicle motion forecasting," *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9638–9644, 2020.
- [35] R. K. H. Galvao, M. C. U. Araujo, G. E. Jose, M. J. C. Pontes, E. C. Silva, and T. C. B. Saldanha, "A method for calibration and validation subset partitioning," *Talanta*, vol. 67, no. 4, pp. 736–740, 2005.