

Using VIS-NIR hyperspectral imaging and machine learning for non-destructive detection of nutrient contents in sorghum

Kai Wu

College of Agricultural Engineering, Shanxi Agricultural University

Zilin Zhang

College of Agricultural Engineering, Shanxi Agricultural University

Xiuhan He

College of Agricultural Engineering, Shanxi Agricultural University

Gangao Li

College of Agricultural Engineering, Shanxi Agricultural University

Decong Zheng

College of Agricultural Engineering, Shanxi Agricultural University

Zhiwei Li

l i z w @ s x a u . e d u . c n

College of Information Science and Engineering, Shanxi Agricultural University

Article

Keywords: Sorghum nutrient contents, Hyperspectral, Non-destructive, Characteristic extraction, Detection model, Machine learning

Posted Date: November 18th, 2024

DOI: <https://doi.org/10.21203/rs.3.rs-5136525/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on February 19th, 2025.
See the published version at <https://doi.org/10.1038/s41598-025-90892-6>.

Using VIS-NIR hyperspectral imaging and machine learning for non-destructive detection of nutrient contents in sorghum

Kai Wu^{a,c}, Zilin Zhang^{a,c}, Xiuhan He^{a,c}, Gangao Li^{a,c}, Decong Zheng^{a,c,*}, and Zhiwei Li^{b,c,*}

^aCollege of Agricultural Engineering, Shanxi Agricultural University, Jinzhong 030801, China; (ORCID: 0009-0007-3174-0617, K.W.)

^bCollege of Information Science and Engineering, Shanxi Agricultural University, Jinzhong 030801, China;

^cDryland Farm Machinery Key Technology and Equipment Key Laboratory of Shanxi Province, Jinzhong 030801, China;

*Corresponding author: Decong Zheng, Email: zhengdecong@sxau.edu.cn (D.Z.);

Zhiwei Li, Email: lizw@sxau.edu.cn (ORCID:0000-0001-6803-2349, Z.L.).

Abstract:

Non-destructive, rapid, and accurate detection of the nutritional compositions in sorghum is of great significance to the application of sorghum in agricultural production and food industry. In the process of sorghum nutrition detection, it can obtain good effect by extracting the corresponding characteristic wavelengths and selecting the suitable detection model for different nutrients. In this study, the crude protein, tannin, and crude fat contents of sorghum variety samples were taken as the research object. Firstly, the visible near-infrared(Vis-NIR) hyperspectral curves of sorghum were measured by the Starter Kit indoor mobile scanning platform (Starter Kit, Headwall Photonics, USA). Secondly, the nutritional components were determined using chemical methods in order to analyze the differences in nutritional composition among different varieties. Thirdly, the original spectral curves were de-noised by Standard normal variate(SNV), Detrending, and Multiplicative Scatter Correction (MSC) algorithms, and the Competitive adaptive reweighted sampling (CARS) and Bootstrapping soft shrinkage (BOSS) algorithms were used to coarse extract the characteristic variables, then Iteratively retains informative variables (IRIV) was used to judge the importance of the characteristic variables, and the optimal wavelength sets of crude protein, tannin and crude fat were obtained respectively. Finally, Partial least squares(PLS), Back propagation(BP) and Extreme learning machine(ELM) were used to establish the non-destructive detection models of crude protein, tannin and crude fat content respectively. The results showed the following: (1) The optimal variable sets of crude protein, tannin and crude fat contain 41, 38 and 22 wavelength variables,

respectively. (2) The CARS-IRIV-PLS model was suitable for detecting crude protein, the prediction set exhibits R^2 , RMSE and RPD values of 0.6913, 0.7996 % and 1.7998. The BOSS-IRIV-PLS model achieved good results in tannin detection, the prediction set exhibits R^2 , RMSE and RPD values of 0.8760, 0.2169 % and 2.8398. The BOSS-IRIV-ELM model achieved the best results in crude fat detection, the prediction set exhibits R^2 , RMSE and RPD values of 0.6145, 0.3208 % and 1.6106. (3) Linear PLS model is suitable for crude protein and tannin detection, and nonlinear ELM model is suitable for crude fat detection. These detection models can be used for the effective estimation of the nutritional compositions in sorghum with Vis-NIR spectral data, and can provide an important basis for the application of food nutrition assessment.

Key Words: Sorghum nutrient contents, Hyperspectral, Non-destructive, Characteristic extraction, Detection model, Machine learning

1 Introduction

Sorghum, a member of the Poaceae family, ranks as the fifth largest cereal crop globally and is predominantly cultivated in arid and semi-arid regions worldwide [1], its grains are rich in nutrients, including starch, protein, tannin, fat, and amino acids[2, 3]. Different varieties of sorghum grains have significantly different nutritional content, these differences lead to different application fields, such as food manufacturing, feed processing and brewing industries[4-6]. Therefore, the accurate acquisition of nutritional components in sorghum grains plays an important role for its application.

Despite the quantification results of traditional chemical detection methods are accurate, but it has numerous shortcomings, including complex sample preparation, high costs, time-consuming procedures, destructiveness to the samples, and need to use special equipment and chemical reagents. Meanwhile, the traditional chemical method quantifies a single nutrient component and cannot quantify multiple nutrients at the same time. All of these limit its widespread application[7, 8]. As a result, a low-cost, non-destructive and efficient detection method is required.

Hyperspectral technology, utilizes specific wavelengths of light absorption, scattering, or reflection to determine the nutrient content without compromising their original structure, which has significant advantages such as being fast and non-destructive[9,10]. It has been recognized in the industry as a new, revolutionary technology recently. Increasingly, studies have shown that can build models with hyperspectral data to predict a variety of nutrient contents. A large number of studies on the application

of spectroscopy to the detection of grain nutritional quality mainly focus on the near-infrared(NIR) spectroscopic data. Fatemi et al. employed NIR spectroscopy to identify and predict the spectral range of major chemical components in maize, thereby assessing its quality [11]. Zareef et al. employed mid-infrared and NIR spectroscopic techniques to analyze crops, including cereals, legumes, and grains [12]. Chen et al. used NIR diffuse reflectance spectroscopy combined with partial least squares discriminant analysis technology to establish a multi-grain rice seed variety recognition method, and the proposed model can provide a reference for special spectrometers [13]. Zhu et al. used KS algorithm to extract characteristic spectra and develop NIR spectroscopy model. The proposed NIR spectroscopy prediction model can meet the requirements of rapid determination of Tartary buckwheat starch[14]. Fan et al. proposed two methods for the detection of amylose content (AC) and fat content (FC) in NIRS single rice grains. The proposed method is suitable for rapid and non-destructive detection and sorting of rice seeds with different AC and FC, and can meet the requirements of coarse screening of rice seed varieties[15]. Their study demonstrated the applicability of mid-infrared spectroscopy in characterizing active compounds present in minor cereals and their processing, while near-infrared spectroscopy was found suitable for quantitative analysis and identification of quality indicators in minor cereals. Huang et al. employed hyperspectral technology for the rapid determination of amylose and amylopectin contents in sorghum. They investigated the impact of various preprocessing methods on Visible (VIS) and NIR spectroscopy data, while also comparing the predictive accuracy of these spectroscopic measurements. By combining principal component analysis and the successive projections algorithm to extract characteristic wavelengths, and using both full-spectrum and characteristic-band spectra, Partial Least Squares Regression (PLSR) and Random Forest (RF) models were established to predict the contents of amylose and amylopectin in different sorghum varieties. The study confirmed the application of spectral technology for rapid and non-destructive prediction of amylose and amylopectin contents in various sorghum varieties[10]. The spectral characteristics of the visible light band (400 to 700 nm) are closely related to a variety of chemical components in the grain. For example, chlorophyll, lutein, carotene and other pigments have specific absorption peaks in the visible light band, and the content of these pigments is directly related to the nutritional quality of cereals. At the same time, other nutrients in grains such as protein, starch, cellulose, etc., may also exhibit specific spectral characteristics in the visible light band. Therefore, the nutritional quality of cereals can be assessed indirectly through spectral analysis in the visible band. Caporaso et al. pointed out that HSI proved to be a multifunctional tool for

studying grains, allowing rapid analysis of individual grains as well as multi-component analysis. HSI is one of the most promising non-contact technologies that can quickly provide chemical information and gain insight into the spatial distribution of chemical properties [16]. Almoujahed et al. combined VIS-NIR and MIR spectroscopy with machine learning models to provide a simple, suitable and non-destructive method to distinguish fusarium ear blight FHB infection in wheat and flour samples at the post-harvest stage [17]. Shi et al. proposed SLR model based on VIS-NIR hyperspectral data, which can accurately predict hundreds of nutrients in wheat grains [18]. Shuai et al. use visible hyperspectral images to capture spatial and spectral features of object surfaces. The combination of deep learning and hyperspectral imaging has shown good results in multi-scale agricultural management [19]. However, there is a limited availability of rapid detection methods for sorghum components such as crude protein, tannin, and crude fat content. Therefore, this work proposes a rapid and non-destructive detection method for sorghum based on VIS-NIR spectral technology to achieve accurate determination of nutrient contents in different sorghum varieties.

This work focused on a total of 93 sorghum varieties, three samples were collected from each variety, comprising 279 samples. First, the spectral information of the samples was obtained using VIS-NIR hyperspectral detection technology, and chemical analysis methods were employed to measure the nutrient contents, including crude protein, tannin, and crude fat. Next, the preprocessing of the data was performed using Standard Normal Variate (SNV), Detrending, and Multiplicative Scatter Correction (MSC) algorithms to enhance the signal-to-noise ratio and data reliability between samples. Then, the Competitive Adaptive Reweighted Sampling (CARS) and Bootstrapping Soft Shrinkage (BOSS) algorithms were utilized for crude feature extraction to identify spectral variable combinations that are highly correlated with the corresponding nutritional components. The Iteratively Retains Informative Variables (IRIV) algorithm was then utilized to further refine the initially extracted feature variables, accurately extracting interpretable variables that are strongly correlated with the nutritional components. These processes effectively reduce the dimensionality of the dataset and highlight important characteristic variable bands. Finally, synchronous detection models for determining crude protein, tannin, and crude fat content in sorghum were constructed using Partial Least Squares (PLS), Back Propagation (BP), and Extreme Learning Machine (ELM) methods, achieving rapid and non-destructive determination of nutrient contents. These methods offer theoretical support for the application of spectral technology combined with machine learning in the detection of sorghum component contents, thereby

enhancing the efficiency and precision of product quality assessment. This advancement caters to market demands and fosters sustainable development in agriculture and food processing technologies.

2 Material and method

2.1 Sorghum samples and growth conditions

A total of 93 sorghum varieties were used in this study, all from the Sorghum Research Institute of Shanxi Agricultural University. The plantation is located in Yuci District, Jinzhong City, Shanxi Province (112°34' ~ 113°8' E, 37°23' ~ 37°54' N). The region has a temperate continental arid climate, with an average altitude of 800 m, an average annual temperature of 9.8°C, an average annual precipitation of 440.7 mm, an annual sunshine duration of 2662 h, and a frost-free period lasting for 158 days.

In this experiment, we selected sorghum grains characterized by plump and uniform, three samples were collected as a weight of 300g from each variety, and a total of 279 samples were collected. All samples stored in wide-mouth bottles under hermetically sealed conditions. The experiment began with the capture of hyperspectral images, followed by the determination of component content in the second step. This sequence aimed to investigate the nutritional composition of different sorghum varieties using various analytical methods, including spectroscopy and chemical analysis.

2.2 Hyperspectral image acquisition and spectral data extraction

2.2.1 Hyperspectral image acquisition

This experiment employed a visible-near infrared hyperspectral (Starter Kit, Headwall Photonics, USA) scanning platform to capture hyperspectral images of sorghum samples. The components of the platform included a visible near-infrared light hyperspectral camera (Headwall Photonics, USA), a 1000W external light source produced by Philips (6995P, Philips, Netherlands), a transmission control platform, and a computer, located in a darkroom. The acquisition parameters are set as follows: the working distance is 240 mm, the push-scan travel is 100 mm, and the platform movement speed is 2.94 mm/s. The lens has a spectral range of 380~1000 nm, a spectral resolution of 0.727 nm, and a total of 856 bands were collected. Due to oscillations in spectral reflectance at the extremes of the spectrometer's range, which introduce significant errors, the selected spectral range for modeling purposes is 430~900 nm, totaling 646 bands.

Sorghum grain samples were loaded into a sample tray with a diameter of 50 mm and a depth of 15 mm, ensuring that it was smooth and compacted, moved by the transmission control platform to facilitate

line scanning of the hyperspectral camera. The computer was used to control the transmission platform and save the raw data of collected from samples.

The process of hyperspectral acquisition was as follows:

(1) Power on the computer, camera and light source, and preheat for 15 min.

(2) Collect current black-and-white board correction data, to eliminate the influence of environmental light, noise, and camera wavelength deviation. The black-and-white board correction formula is as formula 1.

$$R_0 = \frac{R_{raw} - R_{dark}}{R_{white} - R_{dark}} \quad (1)$$

Where, R_0 represents the value of the corrected hyperspectral data, R_{raw} represents the original data, R_{dark} represents the dark current data, and R_{white} represents the whiteboard data.

(3) Hyperspectral image acquisition and data storage.

2.2.2 Spectral data extraction

To simplify the extraction process of hyperspectral image data and achieve batch processing of hyperspectral data, this work developed a preprocessing platform for hyperspectral data using Visual Basic (Version 6.0, Microsoft, USA). By loading the generated sampling point files into Spectral View software (Headwall Photonics, USA), it becomes possible to efficiently collect spectral data from each pixel within the region of interest (ROI) of the samples under examination, while also facilitating preprocessing of the acquired data.

The ellipse model is used to determine the coordinates of the ROI center, the length of the X/Y half-axis (a, b), and the distance between the X/Y axes (Δx and Δy are set to 1). In accordance with the principle of "left to right, top to bottom" in the target image, pixel points in ROI are sequentially collected according to formula 2 to generate ROI coordinate matrix. The ROI region selected for this experiment contains more than 15,000 pixels.

$$\frac{x_i^2}{a^2} + \frac{y_i^2}{b^2} \leq 1 \quad (2)$$

SpectralView software was used to import images and actively extract the reflectivity information according to the coordinate matrix. The average reflectance of all pixels is calculated by formula 3.

$$A_i = \frac{1}{n}(A_1 + A_2 + A_3 + \dots + A_n) = \frac{1}{n} \sum_{i=1}^n A_{i,k} \quad (3)$$

Where, A_i represents the average reflectance, and n denotes the number of pixels.

The calculated average reflectance will serve as the foundational dataset for subsequent data processing.

2.3 Chemical determination of nutrient contents

After the completion of spectral data collection, the 279 samples will undergo pulverization using a high-speed pulverizer (750T, Bo ou Hardware Products Co., Ltd., Yongkang City, China), followed by passage through an 80-mesh sieve, sequentially sealed and stored as raw material samples for component content analysis. The contents of the crude protein, tannin and crude fat were quantitatively determined in this experiment. To minimize potential measurement errors associated with chemical analysis methods, all samples were processed by a single experimenter to ensure consistency. At the same time, each sample was repeated 3 times to take the mean value. In addition, laboratory temperature stability and instrument calibration methods are strictly controlled throughout the testing procedure.

2.3.1 Crude protein contents determination

The crude protein content was determined using the Kjeldahl method [20, 21], employing an automated Kjeldahl nitrogen analyzer (K1100, Shandong Hanon Scientific Instrument Co., Ltd., China) for analysis. The digestion of organic nitrogen compounds involves the utilization of concentrated sulfuric acid to synthesize ammonium sulfate. Subsequently, ammonium sulfate undergoes a reaction with a strong base, resulting in the liberation of ammonia, which is then distilled and absorbed into an excess of standard inorganic solution. The quantification of ammonia is achieved through titration using standard alkali, enabling the calculation of the crude protein content present in the sample. The operational steps are as follows:

Step 1: Place 0.3 g of dried and defatted sorghum sample into a 100 mL Kjeldahl flask, add 5 mL of concentrated sulfuric acid, gently shake to mix, let it stand for 10 h, and place a bent-neck funnel at the mouth of the flask.

Step 2: Add 1 mL of 30% hydrogen peroxide dropwise, raise the temperature of the digestion furnace to 370 °C, and place the Kjeldahl flask into the digestion furnace for digestion.

Step 3: When the solution turns uniformly brown-black, remove the Kjeldahl flask, allow it to cool

down, and add 1 mL of 30% hydrogen peroxide drop by drop while shaking. Continue heating until boiling for 15 min, then remove and let it cool. Add a small amount of 30% hydrogen peroxide again and continue digestion until the liquid becomes colorless or clear bright. To completely remove excess hydrogen peroxide, heat the solution again for 5 to 10 min. Remove the Kjeldahl flask and rinse the bent-neck funnel with distilled water to wash off any remaining excess. Transfer the digestion solution to a volumetric flask and dilute it with distilled water up to 100 mL. Take 25 mL of the clear supernatant for further use.

Step 4: After completing the digestion process of the sample, add 25 mL of the clear supernatant to the reaction chamber of the Kjeldahl nitrogen analyzer. Distill for 5 min, collecting the distillate in a receiving bottle, and then add a mixture of methyl red and bromothymol blue indicators to the outlet of the receiving bottle. Standard hydrochloric acid titration, when the solution changes from bluish-green to purplish-red and remains unchanged for 30 s, this indicates the endpoint of titration. Record the volume of hydrochloric acid consumed during titration and calculate the nitrogen content of the sample. The calculation formula is as formula 4.

$$P = \frac{(A - B) \times C \times 14 \times V_1}{M \times V_2} \times 6.25 \quad (4)$$

Where, P represents the protein content of the sample (mg). A represents the volume of hydrochloric acid consumed during sample titration (mL). B represents the volume of hydrochloric acid consumed during control titration (mL). C represents the molar concentration of standard hydrochloric acid; 14 is the relative molecular mass of nitrogen. V1 represents the total volume of digestion solution (mL). M represents the sample mass (mg). V2 represents the volume of digestion solution used during distillation (mL). 6.25 is a conversion factor for total crude protein content.

2.3.2 Tannin contents determination

Tannin determination involves the extraction of sorghum tannins with dimethyl sulfoxide, followed by measurement using the ferric ammonium citrate colorimetric method and analysis performed using a UV-visible spectrophotometer (752N, Shanghai Yidi Analytical Instrument Co., Ltd., China) [22,23]. The operational steps are as follows:

Step 1: Accurately measure out a sample weighing 1.0 g with a precision of 1 mg and transfer it into the centrifuge tube.

Step 2: Transfer 20 mL of dimethylformamide solution into a hermetically sealed centrifuge tube

containing the sample using a pipette, and agitate for 60 min. Subsequently, subject the mixture to centrifugation at a speed of 3000 g for 10 min.

Step 3: Add 6 mL of distilled water and 1 mL of ammonia solution, respectively. Vigorously agitate the mixture using a oscillator for a duration of 30 s.

Step 4: Add 5 mL of distilled water and 1 mL of ferric ammonium citrate solution, respectively. Vigorously agitate the mixture using a oscillator for a duration of 30 s.

Step 5: After 10 minutes, pour the solution into a cuvette and measure the absorbance at 525 nm using a spectrophotometer, with water as the blank control.

Step 6: Using a tannic acid solution to construct a standard curve enables the determination of tannin content. The calculation formula is as formula 5.

$$D = \frac{2c}{m} \times \frac{100}{100 - H} \quad (5)$$

Where, D represents the tannin content (%), c represents the tannic acid content obtained from the calculation of the standard curve (mg/mL); m represents the sample mass (mg); H represents the moisture content of the sample (%).

2.3.3 Crude fat contents determination

The crude fat content determination is performed using the JTONE series fat analyzer (Soxhlet extractor), which also includes a defatting step for the sample [24]. The working process of the Soxhlet extractor mainly consists of three parts: heating extraction, solvent recovery, and cooling. To improve defatting efficiency, it is necessary to repeatedly soak and extract the sample during the extraction process. During actual operation, adjusting the heating temperature in real-time is possible based on changes in both the reagent's boiling point and ambient temperature. The operational steps are as follows:

Step 1: Fold the skimmed filter paper into a packet with one end, mark it, weigh it, and record its weight as W1.

Step 2: Take a sample weighing 3 ± 0.5 g, place it into a labeled paper bag according to the experiment number, seal it tightly, and put it in an oven set at 105°C . After baking for 3 h, allow it to cool naturally in a desiccator, record the weight, and label it as W2.

Step 3: Place the paper bag into the extractor and add petroleum ether reagent to completely submerge the paper bag, ensuring that some reagent flows into the flask in the water bath.

Step 4: After assembling the equipment, fill the water bath with water, ensuring that there are no air leaks at the interfaces. Begin heating the flask in the water bath and set the temperature of the water bath to 45 ° C. Connect the water source for condensation reflux and reflux for 8 h at a rate of 10 cycles per hour.

Step 5: After the experiment, remove the paper bag and place it in the fume hood for 30 min. Then, set the oven temperature to 105 ° C and bake for 3 h. After cooling in a desiccator, record its weight and label it as W3. The calculation formula is as formula 6.

$$T = \frac{W_2 - W_3}{W_2 - W_1} \times 100\% \quad (6)$$

Where, T represents the fat content (%). W₁ is the weight of the defatted filter paper (g). W₂ is the weight before defatting (g). W₃ is the weight after defatting (g).

2.4 Spectral data preprocessing and data set division

2.4.1 Spectral data preprocessing

To eliminate errors introduced during spectral measurement, enhance the signal-to-noise ratio of the data further, and improve the correlation between reflectance data and physicochemical indicators, preprocessing of the spectral data is conducted. When preprocessing spectral data, the procedures include consecutive application of Standard Normal Variate (SNV), Detrending, and Multiplicative Scatter Correction (MSC). SNV is used to eliminate the effects of grain glossiness, surface scattering, and background interference on reflectance spectra. Detrending is employed to remove baseline drift caused by diffuse reflection. MSC is utilized to eliminate spectral variations due to environmental light scattering during the measurement process, enhancing spectral information related to chemical composition. This correction adjusts baseline shifts and offsets in spectral data, thereby improving the correlation between spectra and data [25-28].

By combining these three preprocessing methods, the differences in spectral intensity between different samples are eliminated, intrinsic sample characteristics are highlighted, differentiation among various features is enhanced, data stability and reliability are improved, data resolution is increased, and spectral data clarity is enhanced. This lays the foundation for establishing detection models in subsequent steps.

2.4.2 Data set division

This work adopts the SPXY algorithm to partition the sample set based on joint x-y distances,

dividing it into calibration and prediction sets in a 3:1 ratio. The calibration set is used for building the model and cross-validation in spectral data modeling with component content, while the prediction set is used to evaluate the predictive performance of the model.

2.5 Characteristic variables extraction

2.5.1 Algorithm of characteristic variables extraction

An extraction algorithm can be employed to transform the original data into a feature set that possesses enhanced representativeness and separability, thereby reducing data complexity and improving model performance. This experiment utilized three extraction algorithms, namely Competitive Adaptive Reweighted Sampling (CARS), Bootstrapping Soft Shrinkage (BOSS), and Iteratively Retains Informative Variables (IRIV).

The CARS method simulates Darwin's theory of "survival of the fittest" by employing adaptive re-weighted sampling techniques [29, 30]. It strategically selects wavelengths with relatively large absolute regression coefficients, which are calculated using partial least squares analysis. Additionally, it utilizes cross-validation to identify the combination set that yields the lowest root mean square error. Due to the inherent stochastic nature of algorithmic computations, it is imperative to conduct multiple iterations to retain bands that exhibit higher selection frequencies, thereby augmenting the stability of subsequent regression models. So, the Monte Carlo sampling procedure was repeated 100 times.

BOSS combines the bootstrap method and flexible shrinkage to address model instability or inaccurate prediction results caused by collinear variables. By statistically analyzing the regression coefficients of multiple sub-models, variable importance is determined based on the absolute value of the coefficients. Variables with larger proportions have a higher probability of selection, thereby evaluating the importance of each variable. Using Weighted Boot Sampling (WBS) for gradual correction and optimization of variable proportions simplifies and reduces the variable space, effectively selecting the variables that have the greatest impact on the predictive performance of the model. The optimal sub-model, which has the smallest root mean square error in the cross-validation set, is selected from the cluster of sub-models to ensure that the model possesses good generalization ability and prediction accuracy, thus obtaining an optimal variable set [31, 32]. The optimal model ratio is set at 0.1, with 1000 iterations of WBS.

IRIV follows the concept of model clustering analysis and fully considers the joint effects among variables. It randomly combines variables to generate a binary matrix, where rows represent variable

combinations and columns represent the number of variables. Based on each row, separate partial least squares models are established, and the effectiveness of these combined models is evaluated using cross-validation root mean square error [26, 33]. Finally, the contributions are ultimately categorized into four distinct groups: strong information variables, weak information variables, no information variables, and interference information variables. The selection of retained variables is determined by the Difference of Mean Values (DMEAN) and P-value. The former calculates the root-mean-square error when the subset contains a certain wavelength variable and when it does not, while the latter is the P-value obtained by Mann-Whitney U test. After executing the algorithm through multiple iterations, ineffective and interfering variables are excluded. Then, reverse elimination is performed to extract the best feature bands from strong and weak information variables. The Mann-Whitney U test is conducted with a P-value of 0.05.

2.5.2. Combined use of characteristic variables extraction algorithms

By employing a combination of extraction algorithms, the dimensionality of the dataset can be further reduced, thereby facilitating the accentuation of feature bands' significance, alleviating overfitting concerns, and providing a more comprehensive and accurate depiction of data characteristics. Consequently, this augmentation enhances model accuracy, computational efficiency, and generalization capability, rendering it better suited for accommodating novel data.

In this work, a two-step strategy was adopted for characteristic wavelength extraction. Firstly, the CARS and BOSS algorithms were used to extract correlation variables corresponding to crude protein, tannin, and crude fat from the entire spectrum, forming a coarse subset. Then, the IRIV algorithm was applied to further refine the optimized variable subset into the selected subset, forming a refined subset. The CARS and BOSS models are based on the concept of model ensemble, which involves selecting the optimal subset of characteristic wavelengths based on RMSECV values. By excluding irrelevant wavelengths related to sorghum chemical composition, the RMSECV value decreases. Ultimately, the subset with the minimum RMSECV value is chosen as the final result for the variable selection of feature wavelengths. By iteratively eliminating variables from the preliminary screening, a finely selected set of feature wavelengths exhibiting strong correlations was identified, thereby enhancing the model's robustness and stability.

2.6 Nutrient contents detection models and evaluation indexes

2.6.1 Nutrient contents detection models

Partial Least Squares (PLS) is a chemometric method that combines multiple linear regression and principal component analysis. Its essence lies in performing regression after eliminating collinearity in spectra, making it suitable for handling linear relationships between multiple variables, including cases with both independent and dependent variables [34, 35]. Its main objective is to identify the principal component direction, which represents the maximum covariance between the predictor variables and the response variables and optimizes the correlations between predictor variables and response variables to obtain new composite variables.

Back Propagation (BP) neural network is a multi-layer feedforward network that utilizes the error backpropagation algorithm for training. It comprises an input layer, hidden layers, and an output layer. Adjacent layers are interconnected through weighted connections, with the input layer receiving external input data, the hidden layers processing data to extract features, and the output layer generating final prediction results. Each node possesses corresponding weight and bias parameters, while nodes within each layer operate independently [36]. The network iteratively performs forward propagation and backpropagation on training samples, gradually adjusting the weights and biases to minimize the discrepancy between the network's output results and actual values. This approach enables efficient data modeling and prediction, while also showcasing strong capabilities in non-linear mapping, self-learning, generalization, and fault tolerance.

Extreme Learning Machine (ELM) is a feedforward neural network-based machine learning algorithm, comprising an input layer, hidden layers, and an output layer. It exhibits applicability to both supervised and unsupervised learning problems. The ELM model has the advantages of faster training speed and better generalization performance compared to traditional neural network models [37]. Optimize the parameters n_{hid} , $actfun$, and $init_weights$ based on the principle of minimizing the Root Mean Square Error (RMSE).

2.6.2 Evaluation indexes

These models' performance was evaluated using the coefficient of determination (R^2), Root Mean Squared Error and relative analysis error (RPD). The calculation formula is as formula 7-9.

$$R^2 = 1 - \frac{\sum_{i=1}^n (Y_{Pi} - Y_i)^2}{\sum_{i=1}^n (Y_i - \hat{Y})^2} \quad (7)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (Y_{Pi} - Y_i)^2}{n}} \quad (8)$$

$$RPD = \frac{1}{\sqrt{1 - R^2}} \quad (9)$$

Generally, prediction ability was defined as:

If the R^2 value is < 0.6 , the models' performing is poor, R^2 value is $0.6-0.8$, which indicates that the model is good, R^2 value > 0.8 , which indicates that the model is very high. A lower RMSE value indicate that the model perform better in terms of prediction accuracy. If the RPD value is < 1.5 , the models' performing is not well; RPD is $1.5-2$, which indicates that the model is helpful for estimation; RPD value is > 2 indicate excellent prediction[38-40].

When extracting characteristic variables and constructing detection models, each algorithm is executed 50 times to ensure robustness, and the optimal set of feature variables and model results are selected based on rigorous evaluation metrics.

3 Results

3.1 Nutrient contents analysis

A total of 279 samples were collected from 93 sorghum varieties, the contents of crude protein, tannin, and crude fat were obtained by calculating the average index for each variety. The results of the chemical analysis are presented in Fig 1, which showed that the average contents of crude protein, tannin, and crude fat were 9.45%, 1.15%, and 3.88% respectively. These findings are consistent with previous reports on the composition of sorghum. Specifically, the crude protein content ranged from 7.21% to 14.39%, the tannin content ranged from 0.04% to 2.56%, and the crude fat content ranged from 2.67% to 5.93%. These differences indicate significant variations in the nutritional content among different varieties of sorghum. Furthermore, the standard deviation (SD) values of crude protein, tannin, and crude fat content were 1.53, 0.72, and 0.56, respectively[34,41]. This reflects the stability and representativeness of the sample data selected for this work. In summary, the 279 samples selected in this experiment exhibit excellent representativeness, thereby establishing a robust data foundation for further investigation into the characteristics and nutritional components of sorghum varieties.

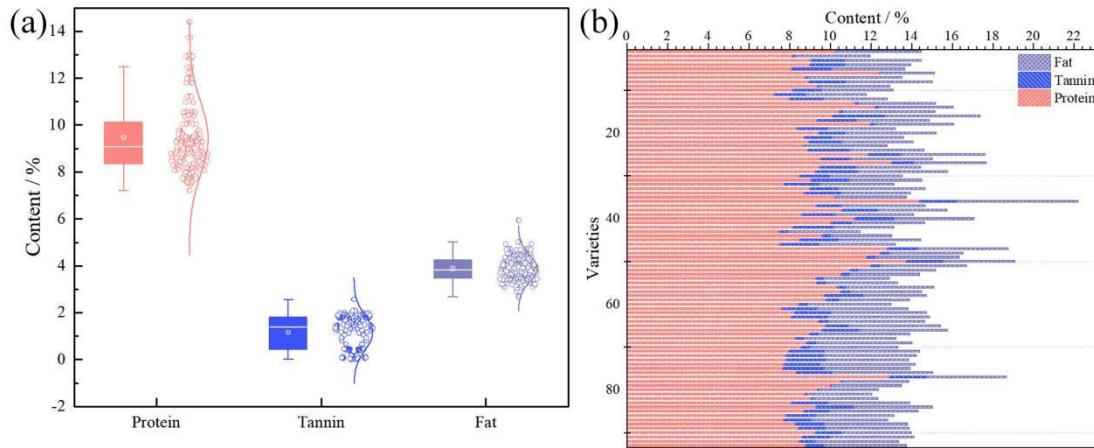


Fig.1 The chemical analysis results of crude protein, tannin, and crude fat content in 93 sorghum varieties.

(a) The box plots and normal distribution of crude protein, tannin, and crude fat in 279 sorghum samples.

(b) The differences in crude protein, tannin, and crude fat content among 93 varieties of sorghum grains.

3.2 Spectral characteristic analysis

Fig 2(a) describes the spectral curves of 279 sorghum samples in the VIS-NIR spectral range. Different sorghum varieties exhibit significant differences in spectral reflectance in the VIS range, mainly due to variations in seed color, which significantly alter the absorption and reflection characteristics of light. For instance, certain sorghum samples exhibit distinctive spectral troughs within the range of 650-680 nm, which can be attributed to spectral variations arising from the presence of dark chromophores in these grains. In the NIR range, the spectral shapes of various sorghum varieties exhibit a high degree of similarity, while their reflectance values differ. These differences in spectral characteristics reflect variations in nutrient composition among different sorghum varieties. Therefore, the spectral differences provide an important foundation for analyzing and detecting sorghum components, offering strong support and guidance for further exploration of its chemical composition and potential applications. The spectral analysis technology can effectively distinguish and evaluate the nutritional characteristics of different sorghum varieties, which is of great significance for further research on and utilization of the characteristics of sorghum varieties. Through the overall observation of the spectrum of all samples, the spectral curves showed a consistent trend of initially increasing and then decreasing, with no apparent abnormal samples. Then, the spectral curves of the original data were processed using SNV, Detrending, and MSC preprocessing algorithms, as shown in Fig 2(b), (c), and (d).

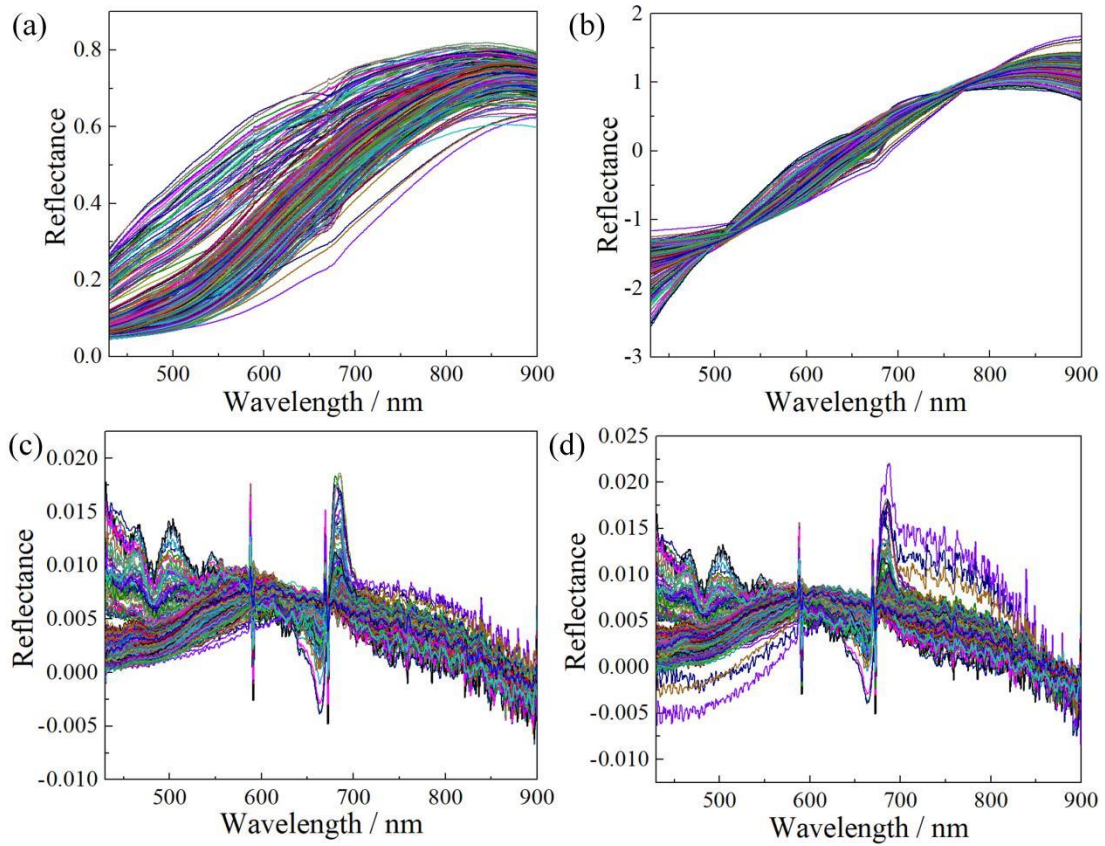


Fig.2 The pretreatment of spectral data for sorghum samples; (a) Raw spectra. (b) Spectral pre-processing by SNV; (c) Spectral pre-processing by SNV and Detrending; (d) Spectral pre-processing by SNV, Detrending and MSC.

Taking the tannin as an example, the PLS model was established by comparing three pretreatment methods, as shown in Table 1. The results show that the combination of the three methods can effectively improve the prediction accuracy of the model. By combining these three preprocessing methods, the spectral differences caused by grain glossiness, surface scattering, background interference, baseline drift, and environmental light scattering were effectively eliminated, while enhancing the spectral information related to chemical composition content. This preprocessing enhances the quality and interpretability of spectral data, thereby contributing to improved accuracy in subsequent modeling.

Table 1 Results of the PLS model established for tannin content by spectral pretreatment method

Preprocessing	Calibration			Prediction		
	R ²	RMSE / %	RPD	R ²	RMSE / %	RPD
RAW	0.7681	0.3257	2.0766	0.7472	0.3666	1.9889
SNV	0.7230	0.3893	1.9000	0.7122	0.4047	1.8640
Detrending	0.8003	0.3002	2.2377	0.7824	0.3082	2.1437
MSC	0.7341	0.3810	1.9393	0.6915	0.4276	1.8004

SNV-Detrending	0.8435	0.2944	2.5278	0.7736	0.3175	2.1017
SNV-MSC	0.8523	0.2861	2.6020	0.8222	0.2595	2.3716
Detrending-MSC	0.7343	0.3801	1.9400	0.6912	0.4280	1.7995
SNV-Detrending-MSC	0.8565	0.2814	2.6398	0.8457	0.2419	2.5458

Additionally, employing the SPXY algorithm, a total of 279 samples were partitioned into two subsets at a ratio of 3:1. One subset was designated as the calibration set, comprising 209 samples (75% of the total), while the remaining subset served as the prediction set, consisting of 70 samples (25% of the total). The data distribution of both the calibration set and prediction set exhibits a relatively uniform pattern, characterized by closely aligned mean values. This indicates that the partitioning of this dataset is reasonable and effective, thereby mitigating potential biases in data distribution, as shown in Table 2. By conducting data preprocessing and implementing a reasonable division of the dataset, it establishes a solid foundation for analyzing and predicting the chemical composition of sorghum, ensuring the reliability and accuracy of the work.

Table 2 The results of the content (%) set partitioning

Items	Calibration					Prediction				
	NS ¹	Mean	Max	Min	SD	NS	Mean	Max	Min	SD
Crude protein	209	9.56	14.39	7.21	1.54	70	9.11	14.39	7.21	1.43
Tannin	209	1.11	2.56	0.06	0.74	70	1.27	2.09	0.04	0.62
Crude fat	209	3.85	5.93	2.67	0.53	70	3.99	5.93	2.89	0.61

1.NS: number of samples

3.3 Characteristic variables analysis

Fig 3 illustrates the variation in $RMSE_{CV}$ as the number of samples increases during the process of extracting feature wavelengths for coarse protein, tannin, and crude fat using the CARS algorithm. When using the CARS algorithm for characteristic wavelength extraction, as the number of Monte Carlo random sampling runs increases, the $RMSE_{CV}$ values exhibit a decreasing trend followed by an increasing trend. During the cross-validation process, the $RMSE_{CV}$ for the crude protein reached its minimum value in the validation set after 41 iterations. A total of 63 wavelengths were selected, accounting for 9.75 % of all bands. At this point, the R_{CV}^2 and $RMSE_{CV}$ were 0.7338 and 0.7961 %, respectively. For the tannin, the $RMSE_{CV}$ reached its minimum value in the validation set after 45 iterations. A total of 50 wavelengths were selected, accounting for 7.74% of all bands. At this point, the R_{CV}^2 and $RMSE_{CV}$ were 0.8517 and 0.2861 %, respectively. For the crude fat, the $RMSE_{CV}$ reached its minimum value in the validation set after 51 iterations. A total of 35 wavelengths were selected, accounting for 5.42% of all bands. At this point, the R_{CV}^2 and $RMSE_{CV}$ were 0.4276 and 0.4038 %, respectively.

respectively.

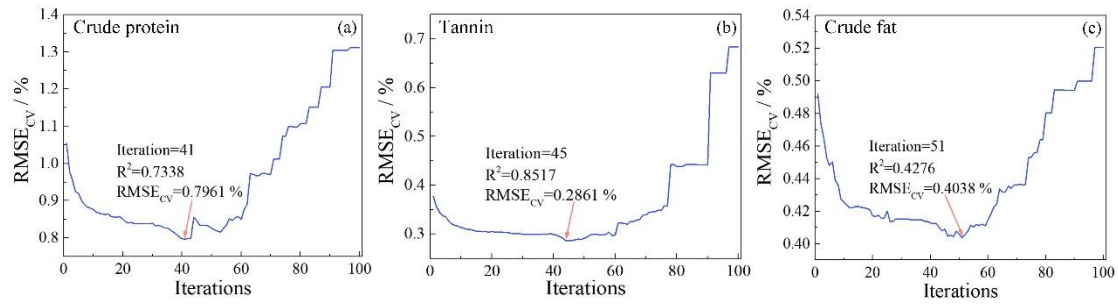


Fig.3 The change of $RMSE_{CV}$ value with the increase of iterations during the key wavelengths extraction process of crude protein, tannin, and crude fat using the CARS algorithm. (a) Crude protein; (b) Tannin; (c) Crude fat.

Fig 4 illustrates the variation in $RMSE_{CV}$ as the number of samples increases during the process of extracting feature wavelengths for crude protein, tannin, and crude fat using the BOSS algorithm. When using the BOSS algorithm for characteristic wavelength extraction, as the number of iterations runs increases, the $RMSE_{CV}$ values exhibit a decreasing trend followed by an increasing trend. During the cross-validation process, the $RMSE_{CV}$ for crude protein reached its minimum value in the validation set after 6 iterations. A total of 73 wavelengths were selected, accounting for 11.30% of all bands. At this point, the R_{CV}^2 and $RMSE_{CV}$ were 0.7001 and 0.8450 %, respectively. For tannin, the $RMSE_{CV}$ reached its minimum value in the validation set after 6 iterations. A total of 61 wavelengths were selected, accounting for 9.44% of all bands. At this point, the R_{CV}^2 and $RMSE_{CV}$ were 0.8506 and 0.2872 %, respectively. For crude fat, the $RMSE_{CV}$ reached its minimum value in the validation set after 9 iterations. A total of 30 wavelengths were selected, accounting for 4.64% of all bands. At this point, the R_{CV}^2 and $RMSE_{CV}$ were 0.4131 and 0.4088 %, respectively.

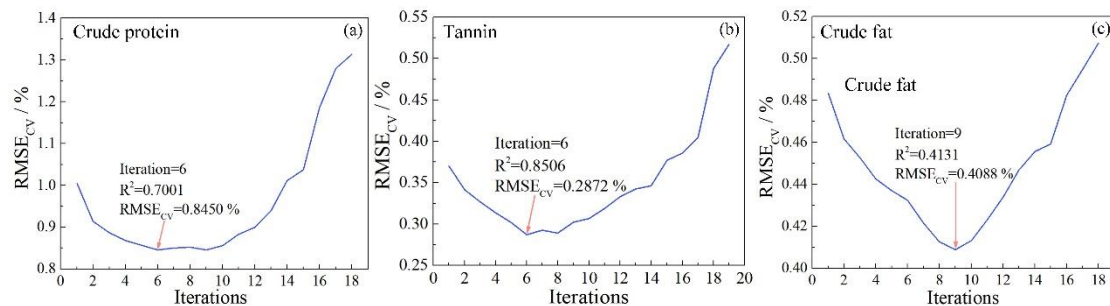


Fig.4 The change of $RMSE_{CV}$ value with the increase of iterations during the key wavelengths extraction process of crude protein, tannin, and crude fat using the BOSS algorithm. (a) Crude protein; (b) Tannin;

(c) Crude fat.

The aforementioned steps finalized the establishment of the coarse subset, upon which, the IRIV algorithm was applied to form a refined subset. The mean root-mean-square error is initially calculated with and without a specific wavelength, followed by the performance of a Mann-Whitney U test ($P=0.05$). Secondly, non-informative variables were eliminated while retaining informative variables of both strong and weak nature. Finally, the optimal subset of variables was determined through cross-validation using the reverse elimination method. The results of the variable classification for the subset were depicted in Fig 5. The optimized results of the variable subsets are as follows.

The crude protein subset of variables derived from CARS and BOSS underwent further refinement, resulting in 41 and 36 variables respectively, which accounted for 6.35% and 5.57% of the total wavelength bands. The selected characteristic wavelengths were primarily concentrated within the ranges of 435-445 nm, 570-575 nm, 687-695 nm, 704-716 nm, 790-810 nm, and 884-900 nm.

The tannin subset of variables derived from CARS and BOSS underwent further refinement, resulting in 44 and 38 variables respectively, which accounted for 6.81% and 5.88% of the total wavelength bands. The selected characteristic wavelengths were primarily concentrated within the ranges of 540-550 nm, 570-600 nm, 628-645 nm, 660-690 nm, 745-765 nm, 820-850 nm, and 880-900 nm.

The crude fat subset of variables derived from CARS and BOSS underwent further refinement, resulting in 27 and 22 variables respectively, which accounted for 4.18% and 3.41% of the total wavelength bands. The selected characteristic wavelengths were primarily concentrated within the ranges of 545-585 nm, 570-600 nm, 630-650 nm, 668-672 nm, 720-750 nm, 785-815 nm, 830-865 nm, and 880-900 nm.

The adopted coarse-refined characteristic variable extraction strategy effectively reduces the number of variables while preserving crucial information wavelengths associated with the crude protein, tannin, and crude fat content. This approach provides an optimized foundation for subsequent analysis and modeling endeavors.

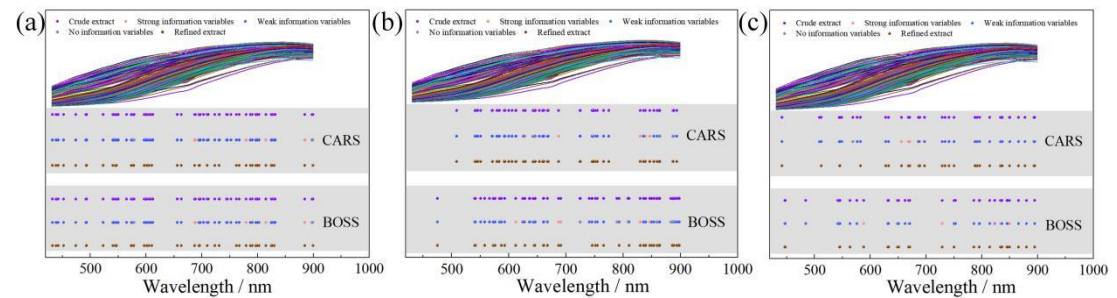


Fig.5 Distribution of the strong information, weak information, no information and interference information wavelengths of key wavelengths of the crude protein, tannin and crude fat in grain spectra of sorghum. (a) Crude protein, (b) Tannin, (c) Crude fat.

3.4 Comparison of detection models

3.4.1 Performances of detection models based on whole spectral data and characteristic spectral data

The PLS, BP, and ELM modeling methods were employed to establish detection models for crude protein, tannin, and crude fat based on whole spectral data initially. Subsequently, we proceed to establish separate models based on the coarse subset and the refined subset, respectively. The results of these detection models are presented in Table 3.

Table 3 The results of detection model for quantifying crude protein, tannin, and crude fat content in sorghum samples.

Items	Characteristic extraction	NVs ¹	Model	Calibration			Prediction		
				R ²	RMSE / %	RPD	R ²	RMSE / %	RPD
Crude protein	Raw	646	PLS	0.7665	0.7456	2.0695	0.6670	0.8273	1.7329
	CARS	63		0.7888	0.7091	2.1760	0.6892	0.7992	1.7937
	CARS-IRIV²	41		0.7927	0.7026	2.1963	0.6913	0.7969	1.7998
	BOSS	73		0.7614	0.7537	2.0472	0.6861	0.8003	1.7849
	BOSS-IRIV	36		0.7702	0.7397	2.0861	0.6619	0.8336	1.7198
	Raw	646	BP	0.6967	0.8498	1.8158	0.5875	0.9208	1.5570
	CARS	63		0.7271	0.8112	1.9142	0.6192	0.8956	1.6205
	CARS-IRIV	41		0.7473	0.7867	1.9893	0.6482	0.8682	1.6860
	BOSS	73		0.7076	0.8309	1.8493	0.6020	0.9099	1.5851
	BOSS-IRIV	36		0.7279	0.8108	1.9171	0.6269	0.8847	1.6371
	Raw	646	ELM	0.6544	0.8735	1.7010	0.6096	0.8993	1.6005
	CARS	63		0.7026	0.8399	1.8337	0.6222	0.8894	1.6269
	CARS-IRIV	41		0.7049	0.8353	1.8408	0.6429	0.8770	1.6734
	BOSS	73		0.6947	0.8536	1.8098	0.6163	0.8997	1.6144
	BOSS-IRIV	36		0.7031	0.8374	1.8352	0.6362	0.8806	1.6579
Tannin	Raw	646	PLS	0.8565	0.2414	2.6398	0.8457	0.2419	2.5458
	CARS	50		0.8882	0.2484	2.9907	0.8338	0.2511	2.4529
	CARS-IRIV	44		0.8823	0.2549	2.9148	0.8508	0.2379	2.5889
	BOSS	61		0.8944	0.2414	3.0773	0.8440	0.2433	2.5318
	BOSS-IRIV²	38		0.8975	0.2378	3.1235	0.8760	0.2169	2.8398
	Raw	646	BP	0.8595	0.2785	2.6679	0.8508	0.2379	2.5889
	CARS	50		0.8850	0.2505	2.9488	0.8647	0.2251	2.7186
	CARS-IRIV	44		0.8903	0.2451	3.0192	0.8681	0.2284	2.7535
	BOSS	61		0.8682	0.2653	2.7545	0.8563	0.2327	2.6380
	BOSS-IRIV	38		0.8838	0.2475	2.9336	0.8613	0.2302	2.6851

	Raw	646		0.8406	0.2928	2.5047	0.8147	0.2793	2.3231
	CARS	50		0.8587	0.2795	2.6603	0.8456	0.2420	2.5449
	CARS-IRIV	44	ELM	0.8797	0.2559	2.8831	0.8502	0.2381	2.5837
	BOSS	61		0.8687	0.2648	2.7597	0.8304	0.2586	2.4282
	BOSS-IRIV	38		0.8808	0.2513	2.8964	0.8609	0.2326	2.6812
Crude fat	Raw	646		0.4400	0.3993	1.3363	0.2331	0.5368	1.1419
	CARS	35		0.5272	0.3670	1.4543	0.3436	0.4996	1.2343
	CARS-IRIV	27	PLS	0.5377	0.3629	1.4707	0.3666	0.4879	1.2565
	BOSS	30		0.5357	0.3637	1.4676	0.3651	0.4885	1.2550
	BOSS-IRIV	22		0.5389	0.3624	1.4727	0.3733	0.4853	1.2632
	Raw	646		0.4197	0.4065	1.3127	0.2503	0.5308	1.1549
	CARS	35		0.4351	0.4022	1.3305	0.2873	0.5269	1.1845
	CARS-IRIV	27	BP	0.4556	0.3964	1.3553	0.3013	0.5197	1.1963
	BOSS	30		0.4267	0.4058	1.3207	0.2741	0.5284	1.1737
	BOSS-IRIV	22		0.4392	0.4011	1.3305	0.2844	0.5280	1.1821
	Raw	646		0.5681	0.3483	1.5216	0.5789	0.3365	1.5410
	CARS	35		0.5982	0.3267	1.5776	0.5285	0.3616	1.4563
	CARS-IRIV	27	ELM	0.6213	0.3046	1.6250	0.5907	0.3294	1.5631
	BOSS	30		0.6182	0.3194	1.6184	0.5819	0.3349	1.5465
	BOSS-IRIV²	22		0.6354	0.3006	1.6561	0.6145	0.3208	1.6106

1NVs: number of variables.

2Bold: optimal models.

The results obtained from the CARS-PLS synchronous detection model are presented as an illustrative example. In terms of the number of variables employed in the modeling process, the variables of the CARS subset decreased by 90.25%, 92.26%, and 94.58% than full spectrum set (Raw) for the crude protein, tannin, and crude fat, respectively. And then, in terms of the result of the detection model, the R^2 of the prediction set for the crude protein and crude fat increased by 0.0222 and 0.0608, respectively. The RPD increased by 0.1105 and 0.0924, respectively. And the RMSE decreased by 0.0281% and 0.0372%, respectively. Its indicates that the performance of the characteristic-spectrum model was superior to that of the full-spectrum model. However, for the tannin model, the R^2 and RPD decreased by 0.0119 and 0.0929, the RMSE increased by 0.0092, and the characteristic-spectrum model performance exhibited a slight decline. The reason may be that the variable strongly correlated with tannin was excluded during the process of extracting characteristic variables from the CARS. Furthermore, upon application of the IRIV algorithm for variable extraction, a further reduction in modeling variables was observed, and the model performance was further enhanced. The number of variables employed in the modeling process, the variables of CARS-IRVI subset of the crude protein, tannin, crude fat decreased by 34.92%, 12.00%, and 22.86% than the variables of the CARS subset. The

R^2 and RPD increased by 0.0021, 0.0061 and 0.0170, and 0.1360, 0.0230 and 0.0222, respectively. And the RMSE decreased by 0.0023 %, 0.0132 %, 0.0117 %. It indicates that the detection model after post fine extraction is more lightweight and has better detection accuracy. The detection models established using both BP and ELM also reach the same conclusion. In summary, employing a combined strategy of extracting characteristic wavelengths can reduce variable redundancy, lighten the model, and enhance model performance.

3.4.2 Optimal models of crude protein, tannin, and crude fat

The optimal results were achieved by employing the CARS-IRIV characteristic extraction algorithm in combination with PLS to establish the detection model for the crude protein content determination. The prediction set exhibits R^2 , RMSE, and RPD values of 0.6913, 0.7996 %, and 1.7998, respectively. For the detection of tannin, the optimal results were achieved by employing the BOSS-IRIV characteristic extraction algorithm in combination with the PLS algorithm. The prediction set exhibits R^2 , RMSE, and RPD values of 0.8760, 0.2169 %, and 2.8398 respectively. For the detection of crude fat content, the optimal results were achieved by employing the BOSS-IRIV characteristic extraction algorithm in combination with the ELM algorithm. The prediction set exhibits R^2 , RMSE, and RPD values of 0.6145, 0.3208 %, and 1.6106 respectively. The results show that the tannin detection model is very high, crude protein and crude fat detection model is good. Fig 6(a)-(c) illustrates the graph depicting the fitting of the model, it showed that these models had excellent performance.

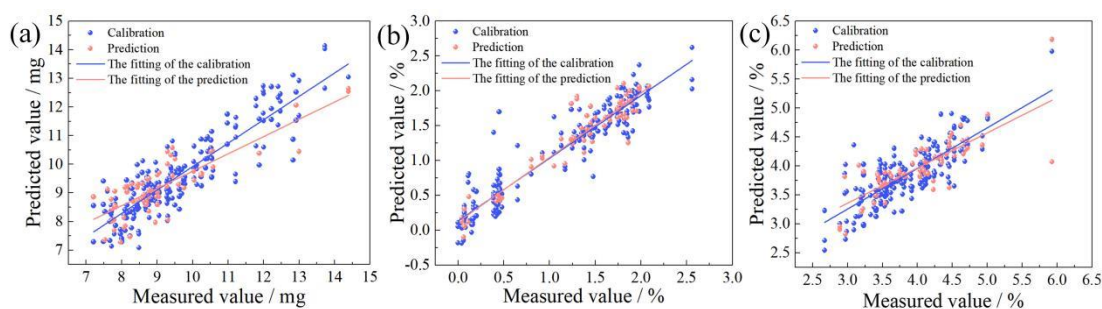


Fig.6 Fitting results of the calibration set and prediction set for each index in the optimal simultaneous detection model. (a) Crude protein; (b) Tannin; (c) Crude fat.

4 Discussion

Direct at the sensitive wavelength screening for crude protein components, the range of 434~522 nm may be attributed to the interaction between light and the π -electron system present in aromatic amino acids (e.g., phenylalanine, tryptophan) within protein molecules. The aromatic ring structures of these

amino acids are sensitive to light within this wavelength range. The range of 540~546 nm may be associated with resonance absorption related to certain π -electron conjugated structures present in proteins, potentially linked to aromatic amino acids. The range of 573~601 nm exhibits sensitivity towards the secondary structures of proteins, such as α -helices and β -sheets, wherein light at this specific wavelength interacts with vibrations or electronic transitions occurring within these structural elements. The range of 610~730 nm can be ascribed to the interaction between vibrational modes of amino acid residues within protein molecules, specifically involving the vibrations of amine (NH) and carboxyl (COOH) groups. The range of 761~899 nm can be ascribed to the interaction between light and electronic transitions of larger molecules, along with the molecular vibrational modes within protein molecules at longer wavelengths. These modes may involve long-chain amino acid residues or overall conformational changes of the protein. Qiao et al. study found that the corn is sensitive to light at 400~700nm and 910nm, related to the C-H vibrations of proteins [42]. Fatemi et al. study indicated that crude protein is sensitive to light at 760 nm, which is related to the vibration of nitrogen elements [43]. Wang et al. study found characteristic wavelengths of rice is nearby 500 nm [44]. These results confirm the accuracy of the characteristic wavelength selected for sorghum crude protein in this work.

Direct at the sensitive wavelengths of tannin, the range of 475~600 nm may match the π - π or n - π transitions of tannin molecules, resulting in related absorption phenomena. The wavelength range of 612~690 nm may interact with the C-H or C-O vibration modes in tannin molecules, thereby triggering associated absorption phenomena. The wavelength range of 745~810 nm may be attributed to interactions such as hydrogen bonding and π - π stacking within tannin molecules, which regulate electron distribution and the charge state of the molecules, thereby influencing their absorption characteristics. The wavelength range of 829~897 nm may be due to tannin being a polyphenolic compounds, with chemical structures that include functional groups such as hydroxyl, carboxyl, and acyl groups. These structural features result in varying absorption characteristics across different wavelengths of light. In their investigation of plant tannin, Zhang et al. identified characteristic spectral ranges at approximately 500-510 nm and around 680-690 nm [45], which closely resemble the selected tannin characteristic wavelengths in this work.

Direct at the sensitive wavelengths of crude fat, the wavelength of 447 nm may be associated with the shorter π - π transitions in fatty acids, particularly near the non-conjugated structures adjacent to double bonds. The absorption at 545 nm and 563 nm may be attributed to the π - π^* transitions in fatty

acid molecules and the absorption associated with conjugated double bonds. The absorption related to the long-chain structure and conjugated system in fatty acids may be associated with a wavelength of 588 nm. 631 nm and 632 nm may be associated with the C-H stretching vibrations in the fatty acid chains, the wavelengths of 649 nm and 650 nm may be associated with the vibrational absorption of C=O stretching in fatty molecules or other functional groups, 668 nm and 671 nm may be related to the spectral features associated with aromatic rings or other conjugated structures in fatty molecules. The wavelengths of 729 nm and 750 nm may be associated with the resonant absorption of long-chain structures in fatty acid molecules, 785 nm and 795 nm may be associated with features such as C-H bending vibrations in fatty acid molecules. Wavelengths in the range of 800-900 nm may exhibit complex absorption patterns in fatty acid chains, including conjugation and long-chain effects. Xue et al. focused on extracting characteristic spectra of crude fat primarily at wavelengths around 450, 500, 680, 800, and 1000 nm [46], which aligns with the findings of our work.

In summary, the coarse subset obtained by the CARS and BOSS algorithms is devoid of any interference variables, however, according to the IRIV selection process, it is possible that certain variables may not have associated messages. It demonstrates the advantages of CARS and BOSS in the characteristic extraction process. The final variable set is obtained through the reverse elimination of the coarse subset, named the refined subset. The refined subset includes only variables with strong and weak characteristic wavelengths. According to DMEAN and P-value in IRIV algorithm, the strong and weak characteristics of variables are judged. A weak feature wavelength is a wavelength point in spectral data or any other form of high-dimensional data that has a weak signal strength and limited direct prediction of the target variable compared to other wavelengths. Although the predictive power of these wavelength points alone may not be strong, they often contain important information that is indirectly related to the target variable or interacts with other strongly characteristic wavelengths. Therefore, ignoring these weak feature wavelengths can lead to information loss, which in turn affects the comprehensiveness and accuracy of the model. On one hand, it demonstrates that the precision extraction strategy can further reduce the number of variables and enhance the accuracy of the detection model. On the other hand, it underscores the crucial role of weak characteristic wavelength in constructing a robust detection model. The findings also demonstrate that the linear PLS detection model is well-suited for detecting crude protein and tannin, whereas the nonlinear ELM detection model is more appropriate for detecting crude fat.

5 Conclusion

In this work, VIS-NIR spectroscopy combined with chemical determination methods are employed to investigate the key nutrient components (crude protein, tannin, and crude fat) of 93 sorghum varieties. The objective was to identify interpretable wavelengths associated with these constituents and establish detection models. The proposed joint strategy incorporates the CARS-IRIV and BOSS-IRIV feature wavelength extraction algorithms, and a significant amount of redundant data is eliminated, and wavelengths with both strong and weak interpretations are extracted. The results show that 41, 38 and 22 characteristic wavelengths have been extracted for crude protein, tannin, and crude fat, respectively, and these characteristic wavelengths have been interpreted and analyzed based on electronic transitions, chemical bond stretching vibrations, molecular structure, etc., demonstrated the reliability of wavelength extraction. In the meantime, the PLS, BP, and ELM modeling methods are employed to establish simultaneous detection models for crude protein, tannin, and crude fat. The results show that the CARS-IRIV-PLSR model is suitable for detecting crude protein, while the BOSS-IRIV-PLSR model achieves good results in tannin detection, and the BOSS-IRIV-ELM model achieved the best results in fat detection. It indicates that the linear model is more suitable for detecting protein and tannin, the nonlinear detection model is more appropriate for detecting crude fat, and all the best models exhibit high R^2 and RPD values, as well as low RMSE.

In summary, simultaneous detection models effectively achieved real-time and non-destructive detection of crude protein, tannin, and crude fat content in sorghum grains. It provides crucial theoretical support and guidance for the application of spectral techniques in detecting other grain components. By combining big data analysis and artificial intelligence technology, an intelligent detection system can be developed for sorghum grains under different growing environments and conditions, promoting its wide application in agricultural production and food processing. In subsequent studies, multiple independent models can be integrated into a comprehensive system to achieve synchronous detection of crude protein, tannin and crude fat of cereals such as sorghum, which is very prospective in order to predict multiple parameters simultaneously in practical applications. This model not only improves predictive efficiency, but also provides more comprehensive quality monitoring in more complex industrial processes. With the continuous advancement of technology, it is expected that these methods will provide important support for achieving food security and high-quality agricultural production, as well as offering theoretical references for research and application in related fields.

Author Contributions: Conceptualization, K.W. and Z.L.; methodology, K.W. and T.Z.; analysis, K.W., Z.Z. and X.H.; investigation, Z.Z. and G.L.; resources, D.Z., and Z.L.; data curation, K.W. and D.Z.; writing—original draft preparation, K.W.; writing—review and editing, D.Z., and Z.L.; supervision, T.Z. and Z.Z.. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Key Research and Development Program Project of Shanxi Province, grant number 202102140601013; Central Government Guides Local Funds for Scientific and Technological Development, grant number YDZJSX20231C009; School academic recovery project, grant number 2023XSHF2; Construction Project of Shanxi Modern Agricultural Industry Technology System, grant number 24142C0207001.

Data Availability: All data generated or analysed during this study are included in this published article.

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

References

1. Khoddami, A., Messina, V., Vadabali Venkata, K., Farahnaky, A., Blanchard, C. L., & Roberts, T. H. (2023). Sorghum in foods: Functionality and potential in innovative products. *Critical reviews In Food Science and Nutrition*, 63(9): 1170-1186. <https://doi.org/10.1080/10408398.2021.1960793>
2. Carcedo, A. J., Mayor, L., Demarco, P., Morris, G. P., Lingenfelser, J., Messina, C. D., & Ciampitti, I. A. (2022). Environment characterization in Sorghum (*Sorghum bicolor* L.) by modeling water-deficit and heat patterns in the Great Plains Region, United States. *Frontiers In Plant Science*, 13: 768610. <https://doi.org/10.3389/fpls.2022.768610>
3. Bakari, H., Djomdi, Ruben, Z. F., Roger, D. D., Cedric, D., Guillaume, P., ... & Gwendoline, C. (2023). Sorghum (*Sorghum bicolor* L. Moench) and its main parts (by-products) as promising sustainable sources of value-added ingredients. *Waste and Biomass Valorization*, 14(4): 1023-1044. <https://doi.org/10.1007/s12649-022-01992-7>
4. Wang, H.; Zhang, X.; Zhang, G.; Fang, Y.; Hou, H.; Lei, K.; Ma, Y. (2023) Regulation of density and fertilization on crude protein synthesis in forage maize in a Semiarid Rain-Fed area. *Agriculture*, 13(3): 715. <https://doi.org/10.3390/agriculture13030715>
5. Zeng, X.; Jiang, W.; Du, Z.; Kokini, J. L. (2023) Encapsulation of tannins and tannin-rich plant extracts by complex coacervation to improve their physicochemical properties and biological activities: A review. *Critical reviews in food science and nutrition*, 63: 3005-3018. <https://doi.org/10.1080/10408398.2022.2075313>
6. Kordan, B.; Nietupski, M.; Ludwiczak, E.; Gabryś, B.; Cabaj, R. (2023). Selected cultivar-specific parameters of wheat grain as factors influencing intensity of development of grain weevil *Sitophilus granarius* (L.). *Agriculture*, 13(8): 1492. <https://doi.org/10.3390/agriculture13081492>

7. Rizvi, N. B.; Aleem, S.; Khan, M. R.; Ashraf, S. Busquets, R. (2022). Quantitative estimation of protein in sprouts of vigna radiate (Mung beans), lens culinaris (Lentils), and cicer arietinum (Chickpeas) by kjeldahl and lowry methods. *Molecules*, 27(3): 814. <https://doi.org/10.3390/molecules27030814>
8. Zhao, H.; Zhan, Y.; Xu, Z.; Nduwamungu, J. J.; Zhou, Y.; Powers, R.; Xu, C. (2022). The application of machine-learning and Raman spectroscopy for the rapid detection of edible oils type and adulteration. *Food Chemistry*, 373: 131471. <https://doi.org/10.1016/j.foodchem.2021.131471>
9. Huang, L.; Luo, R.; Liu, X.; Hao, X. (2022). Spectral imaging with deep learning. *Light: Science & Applications*, 11(1), 61. <https://doi.org/10.1038/s41377-022-00743-6>
10. Huang, H.; Hu, X.; Tian, J.; Jiang, X.; Sun, T.; Luo, H.; Huang, D. (2021). Rapid and nondestructive prediction of amylose and amylopectin contents in sorghum based on hyperspectral imaging. *Food Chemistry*, 359: 129954. <https://doi.org/10.1016/j.foodchem.2021.129954>
11. Fatemi, A.; Singh, V.; Kamruzzaman, M. (2022). Identification of informative spectral ranges for predicting major chemical constituents in corn using NIR spectroscopy. *Food Chemistry*, 383, 132442. <https://doi.org/10.1016/j.foodchem.2022.132442>
12. Zareef, M.; Arslan, M.; Hassan, M. M.; Ahmad, W.; Ali, S.; Li, H.; Ouyang, Q; Wu, X; Hashim, M; Chen, Q. (2021). Recent advances in assessing qualitative and quantitative aspects of cereals using nondestructive techniques: A review. *Trends In Food Science & Technology*, 116: 815-828. <https://doi.org/10.1016/j.tifs.2021.08.012>
13. Chen, J., Li, M., Pan, T., Pang, L., Yao, L., & Zhang, J. (2019). Rapid and non-destructive analysis for the identification of multi-grain rice seeds with near-infrared spectroscopy. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 219: 179-185. <https://doi.org/10.1016/j.saa.2019.03.105>
14. Zhu, L., Liu, F., Du, Q., Shi, T., Deng, J., Li, H., Cai, F., Meng, Z., Chen, Q. Zhang, J. & Huang, J. (2024). Variation Analysis of Starch Properties in Tartary Buckwheat and Construction of Near-Infrared Models for Rapid Non-Destructive Detection. *Plants-Basel*, 13(15): 2155. <https://doi.org/10.3390/plants13152155>
15. Fan, S., Xu, Z., Cheng, W., Wang, Q., Yang, Y., Guo, J., Zhang, P. & Wu, Y. (2022). Establishment of non-destructive methods for the detection of amylose and fat content in single rice kernels using near-infrared spectroscopy. *Agriculture-Basel*, 12(8): 1258. <https://doi.org/10.3390/agriculture12081258>
16. Caporaso N, Whitworth M B, Fisk I D. (2018). Near-Infrared spectroscopy and hyperspectral imaging for non-destructive quality assessment of cereal grains. *Applied Spectroscopy Reviews* 53(8): 667-687. <https://doi.org/10.1080/05704928.2018.1425214>
17. Almoujahed, M. B., Rangarajan, A. K., Whetton, R. L., Vincke, D., Eylenbosch, D., Vermeulen, P., & Mouazen, A. M. (2024). Non-destructive detection of fusarium head blight in wheat kernels and flour using visible near-infrared and mid-infrared spectroscopy. *Chemometrics and Intelligent Laboratory Systems*, 245: 105050. <https://doi.org/10.1016/j.chemolab.2023.105050>
18. Shi, T., Gao, Y., Song, J., Ao, M., Hu, X., Yang, W., chen, W., Liu, Y. & Feng, H. (2024). Using VIS-NIR hyperspectral imaging and deep learning for non-destructive high-throughput quantification

- and visualization of nutrients in wheat grains. Food Chemistry, 461: 140651.
<https://doi.org/10.1016/j.foodchem.2024.140651>
19. Shuai, L., Li, Z., Chen, Z., Luo, D., & Mu, J. (2024). A research review on deep learning combined with hyperspectral Imaging in multiscale agricultural sensing. Computers And Electronics In Agriculture, 217: 108577. <https://doi.org/10.1016/j.compag.2023.108577>
20. Sáez-Plaza, P.; Michałowski, T.; Navas, M. J.; Asuero, A. G.; Wybraniec, S. (2013). An overview of the Kjeldahl method of nitrogen determination. Part I. Early history, chemistry of the procedure, and titrimetric finish. Critical Reviews in Analytical Chemistry, 43(4): 178-223.
<https://doi.org/10.1080/10408347.2012.751786>
21. Sáez-Plaza, P.; Navas, M. J.; Wybraniec, S.; Michałowski, T.; Asuero, A. G.(2013). An overview of the Kjeldahl method of nitrogen determination. Part II. Sample preparation, working scale, instrumental finish, and quality control. Critical Reviews in Analytical Chemistry, 43(4): 224-272.
<https://doi.org/10.1080/10408347.2012.751787>
22. Carmona, A.; Seidl, D. S.; Jaffe, W. G. (1991). Comparison of extraction methods and assay procedures for the determination of the apparent tannin content of common beans. Journal of the Science of Food and Agriculture, 56(3): 291-301. <https://doi.org/10.1002/jsfa.2740560305>
23. Palacios, C. E.; Nagai, A.; Torres, P.; Rodrigues, J. A.; Salatino, A.(2021). Contents of tannins of cultivars of sorghum cultivated in Brazil, as determined by four quantification methods. Food Chemistry, 337: 127970. <https://doi.org/10.1016/j.foodchem.2020.127970>
24. Liu, C.; Chen, F. S.; Xia, Y. M. (2022). Composition and structural characterization of peanut crude oil bodies extracted by aqueous enzymatic method. Journal of Food Composition and Analysis, 105: 104238. <https://doi.org/10.1016/j.jfca.2021.104238>
25. Munawar, A. A.; Meilina, H.; Pawelzik, E. (2022). Near infrared spectroscopy as a fast and non-destructive technique for total acidity prediction of intact mango: Comparison among regression approaches. Computers and Electronics in Agriculture, 193: 106657.
<https://doi.org/10.1016/j.compag.2021.106657>
26. Yao, K.; Sun, J.; Chen, C.; Xu, M.; Zhou, X.; Cao, Y.; Tian, Y. (2022). Non-destructive detection of egg qualities based on hyperspectral imaging. Journal of Food Engineering, 325: 111024.
<https://doi.org/10.1016/j.jfoodeng.2022.111024>
27. Lang, X.; Zhang, Y.; Xie, L.; Li, P.; Horch, A.; Su, H. (2022). Detrending and denoising of industrial oscillation data. IEEE Transactions on Industrial Informatics, 19(4): 5809-5820. <https://doi.org/10.1109/TII.2022.3188844>
28. Dhanoa, M. S.; López, S.; Sanderson, R.; Lister, S. J.; Barnes, R. J.; Ellis, J. L.; France, J. (2023). Methodology adjusting for least squares regression slope in the application of multiplicative scatter correction to near-infrared spectra of forage feed samples. Journal of Chemometrics, 37(11): e3511. <https://doi.org/10.1002/cem.3511>
29. Tang, G.; Huang, Y.; Tian, K.; Song, X.; Yan, H.; Hu, J.; Xiong, Y.; Min, S. (2014). A new spectral variable selection pattern using competitive adaptive reweighted sampling combined with successive projections algorithm. Analyst, 139(19): 4894-4902.
<https://doi.org/10.1039/C4AN00837E>

30. He, X.; Huanyu, E.; Ding, G. (2023). Development of a CH₂-dependent analytical method using near-infrared spectroscopy via the integration of two algorithms: non-dominated sorting genetic-II and competitive adaptive reweighted sampling (NSGAII-CARS). *Analytical Methods*, 15(10): 1286-1296. <https://doi.org/10.1039/D2AY02072F>
31. Yan, H.; Song, X.; Tian, K.; Gao, J.; Li, Q.; Xiong, Y.; Min, S. (2019). A modification of the bootstrapping soft shrinkage approach for spectral variable selection in the issue of over-fitting, model accuracy and variable selection credibility. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 210: 362-371. <https://doi.org/10.1016/j.saa.2018.10.034>
32. Zhang, P.; Xu, Z.; Ma, H.; Cheng, W.; Li, X.; Tang, L.; Zhao, G.; Wu, Y.; Liu, Z.; Wang, Q. Novel comprehensive variable selection algorithm based on multi-weight vector optimal selection and bootstrapping soft shrinkage. *INFRARED PHYS TECHN*, 2023, 133, 104800. <https://doi.org/10.1016/j.infrared.2023.104800>
33. Wang, F.; Wang, C.; Song, S. (2022). Rapid and low-cost detection of millet quality by miniature near-infrared spectroscopy and iteratively retaining informative variables. *Foods*, 11(13): 1841. <https://doi.org/10.3390/foods11131841>
34. Mateos-Aparicio, G. (2011). Partial least squares (PLS) methods: Origins, evolution, and application to social sciences. *Communications in Statistics-Theory and Methods*, 40(13): 2305-2317. <https://doi.org/10.1080/03610921003778225>
35. Liu, C.; Zhang, X.; Nguyen, T. T.; Liu, J.; Wu, T.; Lee, E.; Tu, X. M. (2022). Partial least squares regression and principal component analysis: similarity and differences between two popular variable reduction approaches. *General psychiatry*, 35(1): e100662. <http://doi:10.1136/gpsych-2021-100662>
36. Deng, Y.; Zhou, X.; Shen, J.; Xiao, G.; Hong, H.; Lin, H.; Wu, F.; Liao, B. Q. (2021). New methods based on back propagation (BP) and radial basis function (RBF) artificial neural networks (ANNs) for predicting the occurrence of halo ketones in tap water. *Science of The Total Environment*, 772: 145534. <https://doi.org/10.1016/j.scitotenv.2021.145534>
37. Zhou, H.; Ma, L.; Niu, X.; Xiang, Y.; Chen, J.; Su, Y.; Li, J.; Lu, S.; Chen, C.; Wu, Q. (2024). A novel hybrid model combined with ensemble embedded feature selection method for estimating reference evapotranspiration in the North China Plain. *Agricultural Water Management*, 296: 108807. <https://doi.org/10.1016/j.agwat.2024.108807>
38. Piepho, H. P. (2019). A coefficient of determination (R^2) for generalized linear mixed models. *Biometrical journal*, 61(4): 860-872. <https://doi.org/10.1002/bimj.201800270>
39. Chen, L.; Wu, X.; Lopes, A. M.; Yin, L.; Li, P. (2022). Adaptive state-of-charge estimation of lithium-ion batteries based on square-root unscented Kalman filter. *Energy*, 252: 123972. <https://doi.org/10.1016/j.energy.2022.123972>
40. Wang, G.; Wang, W.; Fang, Q.; Jiang, H.; Xin, Q.; Xue, B. (2018). The application of discrete wavelet transform with improved partial least-squares method for the estimation of soil properties with visible and near-infrared Spectral Data. *Remote Sensing*, 10(6): 867. <https://doi.org/10.3390/rs10060867>

41. Khoddami, A.; Messina, V.; Vadabali Venkata, K.; Farahnaky, A.; Blanchard, C. L.; Roberts, T. H. (2023). Sorghum in foods: Functionality and potential in innovative products. *Critical reviews in food science and nutrition*, 63(9): 1170-1186. <https://doi.org/10.1080/10408398.2021.1960793>
42. Qiao, M.; Cui, T.; Xia, G.; Xu, Y.; Li, Y.; Fan, C.; Han, S.; Dong, J. (2024). Integration of spectral and image features of hyperspectral imaging for quantitative determination of protein and starch contents in maize kernels. *Computers and Electronics in Agriculture*, 218: 108718. <https://doi.org/10.1016/j.compag.2024.108718>
43. Fatemi, A.; Singh, V.; Kamruzzaman, M. (2022). Identification of informative spectral ranges for predicting major chemical constituents in corn using NIR spectroscopy. *Food Chemistry*, 383: 132442. <https://doi.org/10.1016/j.foodchem.2022.132442>
44. Wang, Z.; Liu, J.; Zeng, C.; Bao, C.; Li, Z.; Zhang, D.; Zhen, F. (2023). Rapid detection of protein content in rice based on Raman and near-infrared spectroscopy fusion strategy combined with characteristic wavelength selection. *Infrared Physics & Technology*, 129: 104563. <https://doi.org/10.1016/j.infrared.2023.104563>
45. Zhang, P.; Wu, Q.; Wang, Y.; Huang, Y.; Xie, M.; Fan, L. (2024). Rapid Detection of Tannin Content in Wine Grapes Using Hyperspectral Technology. *Life*, 14(3): 416. <https://doi.org/10.3390/life14030416>
46. Fei, X.; Jiang, X.; Lei, Y.; Tian, J.; Hu, X.; Bu, Y.; Huang, D.; Luo, H. (2023). The rapid non-destructive detection of the protein and fat contents of sorghum based on hyperspectral imaging. *Food Analytical Methods*, 16(11): 1690-1701. <https://doi.org/10.1007/s12161-023-02529-x>