ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

Optimizing Near-Infrared Spectroscopy Models for Rapid and Green Detection of Crude Protein and Fat in Crop Grains Using Sample Set Division

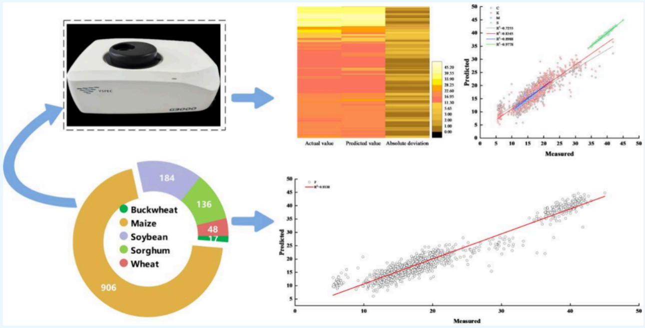
Qing Yang,[#] Yujiao Li,[#] Jie Li, Zhiyou Zhang, Qiqi Liu, Ge Guo, Shuang Wang, Xiaoyu Wang,* and Guanghui Xie

Cite This: *ACS Omega* 2025, 10, 14755−14769

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Rapid detection of crop grain components is crucial for effective production and energy conversion. We used the sample set division method to divide multiple sample sets and optimize NIRS models for rapid prediction of protein and fat content. 1243 and 415 crop grain samples were screened and divided into 5 and 4 sets, respectively. The aim was to establish NIRS models for protein and fat content prediction. The best modeling methods for protein were N (Norris Derivative)+D (detrending)-C (CARS)-P (PLS) and N+M (MC-UVE)-C-P, while those for fat were N+M-C-P and N+S (Savitzky-Golay)-C-P. The SS (Soybean Set), KS (Sorghum Set), and FS (Full Samples Set) data sets provided accurate protein content analysis, while the FS and SS data sets were suitable for both protein content prediction and evaluation. For fat, the FS, SS, and CS (Cereal Set) models met content analysis requirements, with the FS model suitable for external validation. It compared and analyzed the fitness, robustness, and accuracy of different NIRS set models, employing various division methods in this study, which provided a new idea of green method theoretical and technical support for major component rapid detection of biomass raw materials.



1. INTRODUCTION

Rapid development of near-infrared spectroscopy (NIRS) has been promoted by computer technology and modern mathematics. The combination of NIRS and chemometrics has become the fastest-growing and most eye-catching chemical analysis technology nowadays.[1−3] NIRS has been widely used in agriculture, petrochemicals, pharmaceuticals, food, and medicines since the 1990s. In agriculture, NIRS is mainly used for the determination of protein and water in agricultural products, as well as the analysis of soil moisture and crop residues.[4,5] By scanning the near-infrared spectrum of samples, NIRS can provide information about the molecular structure, composition, and physical state of the samples. It meets the demands for the detection of major components in crop grain and biomass, serving as an important tool for nondestructive and rapid determination of complex natural samples.[6,7]

The crude protein content of crop grain is an important index for evaluating grain quality, making the determination of the main components of crop grain highly significant for crop production, market transaction, and food processing. Conventional wet chemical analysis and near-infrared analysis are commonly used methods for detecting crop grain content,[8] NIRS method demonstrates greater efficiency compared to traditional detection methods[9] for analysis of crop grain samples.[10] Chemical analysis methods have drawbacks such as low sensitivity, long measurement time, complex procedures, and high cost. The NIRS analysis method is a nondestructive detection technique that allows for the avoidance of tedious preprocessing steps typically associated with chemical analysis, thereby reducing testing costs[11] Establishing an NIRS quantitative model for crop grain content analysis would enable rapid, accurate, and quantitative analysis of samples, providing a valuable tool for efficient crop production and energy conversion.

Currently, machine vision, hyperspectral imaging, imaging technologies, and electronic nose technology are also used for crop grain content analysis. However, NIRS technology stands out for its rapid field analysis, simple experimental procedures, higher efficiency, and the ability to simultaneously measure

Received: October 9, 2024
Revised: January 26, 2025
Accepted: March 31, 2025
Published: April 8, 2025

https://doi.org/10.1021/acsomega.4c09155

multiple components, enabling high-throughput automation.[12] An NIRS model using partial least-squares regression (PLSR) and support vector regression (SVR) was developed to predict protein content in single kernel maize seed with high accuracy.[13,14] The predicted values closely matched the reference values, indicating that NIRS technology is a feasible method for analyzing the protein content of crop grain.[15] Partial least-squares (PLS) is a linear correction method that demonstrates significant effectiveness in addressing the issue of multicollinearity among variables, exhibiting robust resistance to interference.[16] Li and Wang[17] compared NIRS and conventional wet chemical analysis for determining protein and fat content in soybeans. The results showed that the variations were less than 1.36 percentage points, with the smallest variation being only 0.19 percentage points.

The application of NIRS for measurement exhibits promising potential to supplant conventional wet analysis methods in crop grain protein and fat analysis.[18] Several studies have identified machine learning techniques such as PLSR and SVR as the optimal model for protein content detection in crop grain.[19,20] A successive projection algorithm (SPA)-principal component analysis (PCA)-support vector classification (SVC) method based on variable selection, feature extraction, and nonlinear modeling method was established to quickly identify contaminated rice, which showed that the recognition performance of the model constructed with the fused data was significantly improved.[21] In this study, spectral pretreatment methods, variable selection methods, and correction methods were combined for modeling. Norris derivative filter (ND) and Savitzky-Golay (SG) methods can effectively smooth and denoise the spectrum, making the peak of the spectrum clearer. Competitive adaptive reweighted algorithm (CARS) and Monte Carlo Uninformative Variable Elimination (MC-UVE) can eliminate irrelevant or uninformative variables and improve model prediction performance. This study aims to develop a technology for measuring the protein and fat content of crop grain by applying machine learning techniques. The near-infrared spectral characteristics of crop grain according to the protein and fat content were investigated, and the machine learning models were developed and their performance was compared.

There were collected crop grain samples from various regions, including wheat, maize, soybean, sorghum, and buckwheat, to develop NIRS models for crude protein and crude fat. Different modeling methods and sample set divisions were employed to assess model performance, accuracy, and robustness. Kennard-stone (KS) division method divided samples based on the Euclidean Distance difference of their spectra. The key of the KS method was to select samples based on the maximum distance between them, ensuring the training set represented the entire data set's spatial distribution, thereby improving the model's generalization ability and predictive performance. Both full sample set models and species-specific models were established to evaluate the impact of sample set division. The optimal sample set division method was identified, enabling rapid and accurate qualitative or quantitative analysis. This research presented innovations in various sample species and detection components, and different biomass sample set divisions were employed to establish models. Besides, a comparison of different NIRS modeling approaches and sample set models offers a more comprehensive and in-depth analysis than previous studies.

2. MATERIALS AND METHODS

2.1. Samples Preparation. The crop grain samples were divided into two separate sets: one for the analysis of crude protein content (comprising 1243 samples) and the other for the analysis of crude fat content (comprising 415 samples). The sources of these samples are presented in Table 1. The test samples were kindly provided by the National Energy Non-Grain Biomass Raw Materials R&D Center of China Agricultural University.

Table 1. Source and Quantity of Crop Grain Test Samples[a]

content	crop	sample source	sample size	total
crude protein	buckwheat	–	17	1243
	maize	Quzhou, Hebei/Zhuozhou, Hebei	906	
	soybean	Yicheng, Hubei	184	
	sorghum	Zhuozhou, Hebei	136	
crude fat	soybean	Yicheng, Hubei	184	415
	wheat	Zhuozhou, Hebei	48	
	maize	Zhuozhou, Hebei	48	
	sorghum	Zhuozhou, Hebei	118	
	buckwheat	–	17	

[a]Note: "–" means unknow source.

2.2. Chemical Values of Crop Grain Samples Determination. The National Energy Non-Grain Biomass Raw Materials R&D Center of China Agricultural University provided the chemical reference values and spectral information on crop grain. The total nitrogen content of crop grain samples was determined according to the standard NY/T 2419-2013 "Determination of total nitrogen in plant-Automatic Kjeldahl apparatus method."[22] Based on this standard, the crude protein content value (CP) of the test samples was calculated as

$$CP = N \times K \tag{1}$$

where CP denotes crude protein content in samples (g/kg), N denotes total nitrogen content in samples (g/kg) and K denotes protein conversion factor (6.25). The crude fat content of crop grain was determined by the Soxhlet extraction method in the standard GB 5009.6—2016 "Determination of fat in foods-National food safety standard".[23]

2.3. NIRS Scanning and Pretreatments. A Scientific-Antaris II Fourier Near-Infrared Spectrometer (Thermo Fisher USA) was used, equipped with a tungsten halogen lamp as the light source, a fixed mirror, a moving mirror, and a beam splitter as the interferometer. The laser was a HeNe laser with a laser wavelength of 632.8 nm, and the detector was a rated InGaAs detector. The environmental temperature and humidity were critical factors that could influence the performance of the spectrometer and the scanned spectrum. In this study, the temperature during the acquisition of near-infrared spectrum information on crop grain samples ranged from 19.3 to 24.5 °C, and the humidity was between 28 and 35%.

Before collecting the near-infrared spectrum information on the samples, the spectrometer was preheated for 30 min. The samples were poured into the sample cup, compressed using a sample press to prevent external light from penetrating the sample and placed on the rotary stage of the spectrometer for integrating sphere sampling. The spectral scan had a

Table 2. Sample Size and Crude Protein Value Range of Different Sample Divisions for Calibration and Prediction of Near-Infrared Spectroscopy Modeling[a]

	calibration					prediction				
sample set division	SS (n)	max (%)	min (%)	mean (%)	SD (%)	SS (n)	max (%)	min (%)	mean (%)	SD (%)
One Set										
1. Full samples (F)	994	45.04	5.50	19.73	8.62	249	42.07	5.56	21.15	10.68
Four Subsets										
1. Cereal (C)	833	41.84	5.50	16.98	4.82	209	30.46	5.56	15.42	3.98
2. Sorghum (K)	108	22.00	11.40	14.68	2.12	28	20.52	10.55	14.10	2.46
3. Maize (M)	724	41.84	5.50	17.38	5.03	182	28.78	5.65	15.42	3.96
4. Soybean (S)	147	45.04	34.06	39.15	1.99	37	42.07	35.16	38.90	1.76

[a]Note: SD stands for Standard deviation; SS means sample size in this table.

Table 3. Sample Size and Crude Fat Value Range of Different Sample Divisions for Calibration and Prediction of Near-Infrared Spectroscopy Modeling[a]

	calibration					prediction				
sample set division	SS (n)	max (%)	min (%)	mean (%)	SD (%)	SS (n)	max (%)	min (%)	mean (%)	SD (%)
One Set										
1. Full samples (F)	332	25.41	1.54	10.79	8.35	83	24.37	1.67	14.09	8.86
Three Subsets										
1. Cereal (C)	171	6.28	1.57	3.89	1.07	43	5.13	1.54	3.20	0.96
2. Soybean (S)	147	25.41	18.08	20.88	1.38	37	24.40	18.41	21.00	1.32
3. Sorghum (K)	94	6.28	2.20	4.33	0.79	24	5.67	3.31	4.46	0.67

[a]Note: SD stands for Standard deviation; SS means sample size in this table.

wavelength range of 4000−10,000 cm^{-1}, a resolution of 8 cm^{-1}, and a wavenumber accuracy of ± 0.03 cm^{-1}. Each sample was scanned 64 times, resulting in spectra containing 1557 spectral variables.

2.4. Modeling of Near-Infrared Spectra.

2.4.1. Modeling Methods. In this study, the combination of spectral pretreatment methods, variable selection methods, and correction methods was used to establish the model, and the optimal combination method was selected. In pretreatment methods, detrending (DT), SG, ND filter, and multiplicative scatter correction (MSC) were used to reduce or eliminate the impact of nontarget factors on the spectrum. Removing irrelevant information variables can improve spectral resolution and sensitivity. CARS and MC-UVE were employed in variable selection methods. The PLS correction method was a linear correction method based on the improved principal component regression analysis.

2.4.2. Crude Protein NIRS Model Establishment. Using two distinct approaches, we separated 1243 crop grain samples into five distinct sets: one set (Full samples) and four subclass sets (cereal: maize+sorghum, maize, sorghum, soybean). The Kennard-Stone technique was applied to allocate samples into a modeling set and a prediction set in a 4:1 ratio. The reference value content ranges for both sets are detailed in Table 2. Chem Data Solution 3.1.0 software was used to build an NIRS model for the estimation of crude protein and crude fat content in crop grains.

In Table 2, the modeling set and prediction set models for the five sets were present. The soybean set (SS) had the highest mean crude protein content at 39.15 and 38.90% for the modeling and prediction sets, respectively. In contrast, the Sorghum set (KS) had the lowest crude protein content at 14.68 and 14.10% for both sets. The standard deviation was lowest for SS at 1.99 and 1.76% for modeling and prediction sets. On the other hand, the full samples set (FS) had the highest standard deviation with 8.62 and 10.68% for the modeling and prediction sets, respectively. For FS, the modeling set's crude protein content ranged from 5.50 to 45.04%, while the prediction set's range was from 5.56 to 42.07%. This indicated that the division results were reasonable as the distribution range was consistent, and the extreme values at both ends of the prediction set were encompassed within the modeling set's range.

2.4.3. Crude Fat NIRS Model Establishment. 415 crop grain samples were divided into four sample sets: one set (Full samples) and three subclass sample sets (cereal: wheat+maize +sorghum, sorghum, soybean). Utilizing the Kennard-Stone method, we allocated samples into a modeling set and a prediction set in a 4:1 ratio to establish an NIRS model for crude fat content estimation. The division of crop grain sample sets and the range of values for the modeling set and prediction set are summarized in Table 3.

In the modeling set and prediction set of four sample sets, the mean crude fat content of SS was the highest, at 20.88 and 21.00% respectively. Conversely, the Cereal set (CS) was the lowest crude fat content at 3.89 and 3.20%. FS had the highest standard deviation at 8.35 and 8.86% in modeling set and prediction set. On the other hand, the Kennard-Stone set had the lowest standard deviation at 0.79 and 0.67% for the modeling and prediction sets, respectively. For FS, the modeling set's crude fat content ranged from 1.54 to 25.41%, while the prediction set's range was from 1.67 to 24.37%. This indicates that the division results were reasonable as the distribution range was consistent and the extreme values at both ends of the prediction set were encompassed within the modeling set's range.

2.5. Model Evaluation.

The evaluation criteria for the NIRS model were determined by several indices: the coefficient of determination (R^2), root-mean-square error (RMSE), relative percent deviation (RPD), and average relative deviation (ARD). R^2 represents the proportion of the variance in the dependent variable that is explained by the

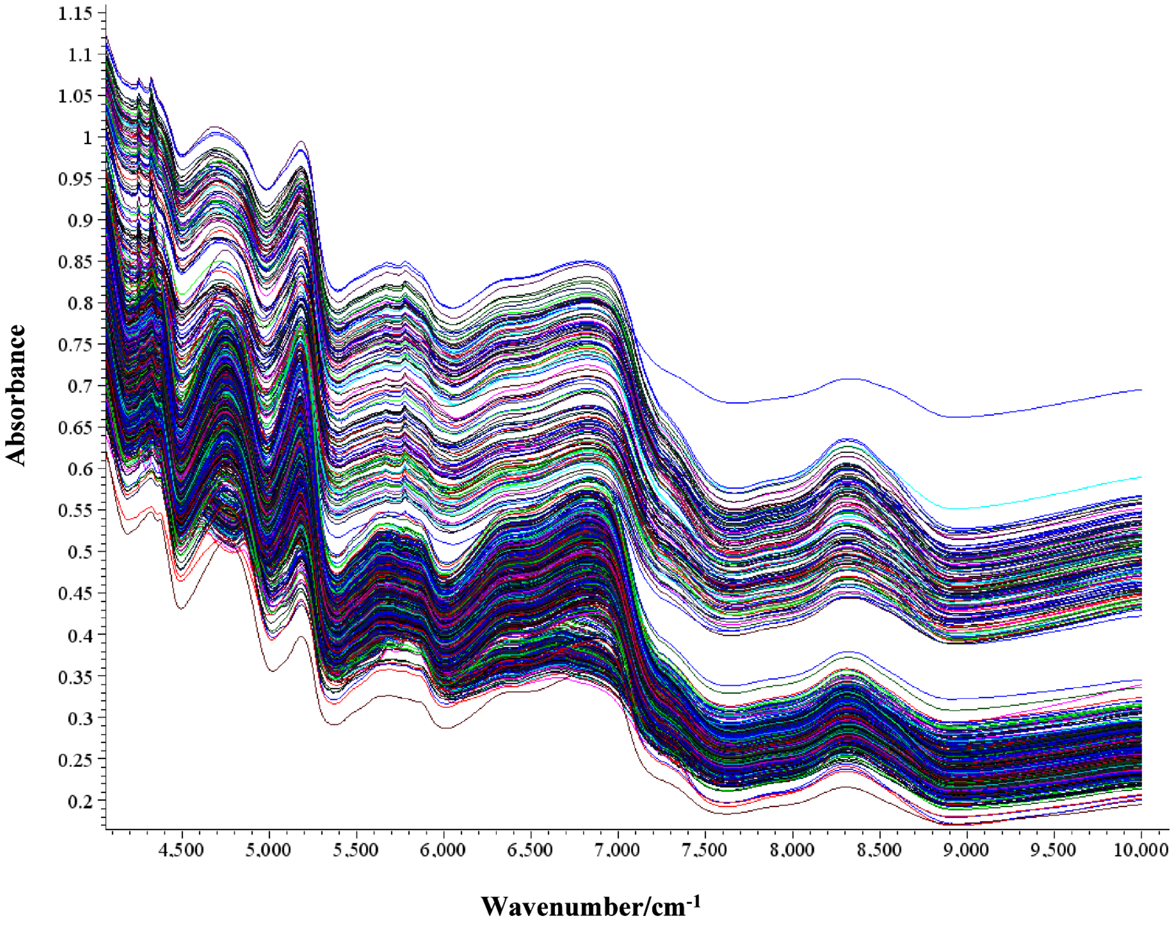


Figure 1. Original near-infrared spectrum of crop grain samples for crude protein.

model, with a value closer to 1 indicating better predictive performance. RMSE measures the standard error of the differences between NIRS results and reference values. The RMSE of cross validation (RMSECV) is a crucial index for assessing the stability and internal predictive ability of the model. The RMSE of prediction (RMSEP) is also significant for evaluating the actual predictive capability of the model through external validation. It calculates the variance between predicted and reference values to assess prediction accuracy. Therefore, smaller values of RMSECV and RMSEP indicate better model fitting.

RPD could be used to validate the stability and predictive ability of the model. An RPD value greater than 3 indicates a high prediction accuracy, making the model suitable for predicting the relevant components of samples. An RPD between 2.5 and 3 suggests the model can be used for quantitative analysis, but further improvement is needed. An RPD below 2.5 indicates that the NIRS model requires optimization, as it may not be suitable for analysis or detection.

The ARD serves as an important index for evaluating the robustness of the model. A smaller ARD value indicates smaller prediction errors and higher prediction accuracy of the model.[24,25]

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} \qquad (2)$$

where R^2: coefficient of determination; $\hat{y}_i$: predicted value of the ith sample; y_i: reference value of the ith sample; $\overline{y}$: mean value of sample reference value; n: sample size.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}} \qquad (3)$$

$$\mathrm{RPD} = \mathrm{SD}/\mathrm{RMSE} \qquad (4)$$

where RMSE: root-mean-square error; RPD: relative percent deviation; SD: standard deviation of reference value for the modeling or prediction set.

$$\mathrm{ARD} = \frac{\sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|}{n} \qquad (5)$$

where ARD: average relative deviation.

3. RESULTS AND DISCUSSION

3.1. Crude Protein NIRS Model.

3.1.1. Original Spectral Characteristics of Samples and Pretreatment for Crude Protein. A total of 1243 samples were scanned to obtain their near-infrared original spectra (Figure 1). The observed trends in the near-infrared spectral curves were predominantly similar, reflecting the likeness in genetic information and chemical composition among various crops. Variations in the content of specific chemical constituents across different samples led to disparities in the range of spectral absorbance. Each absorption peak in the near-infrared spectrum occurred within the range

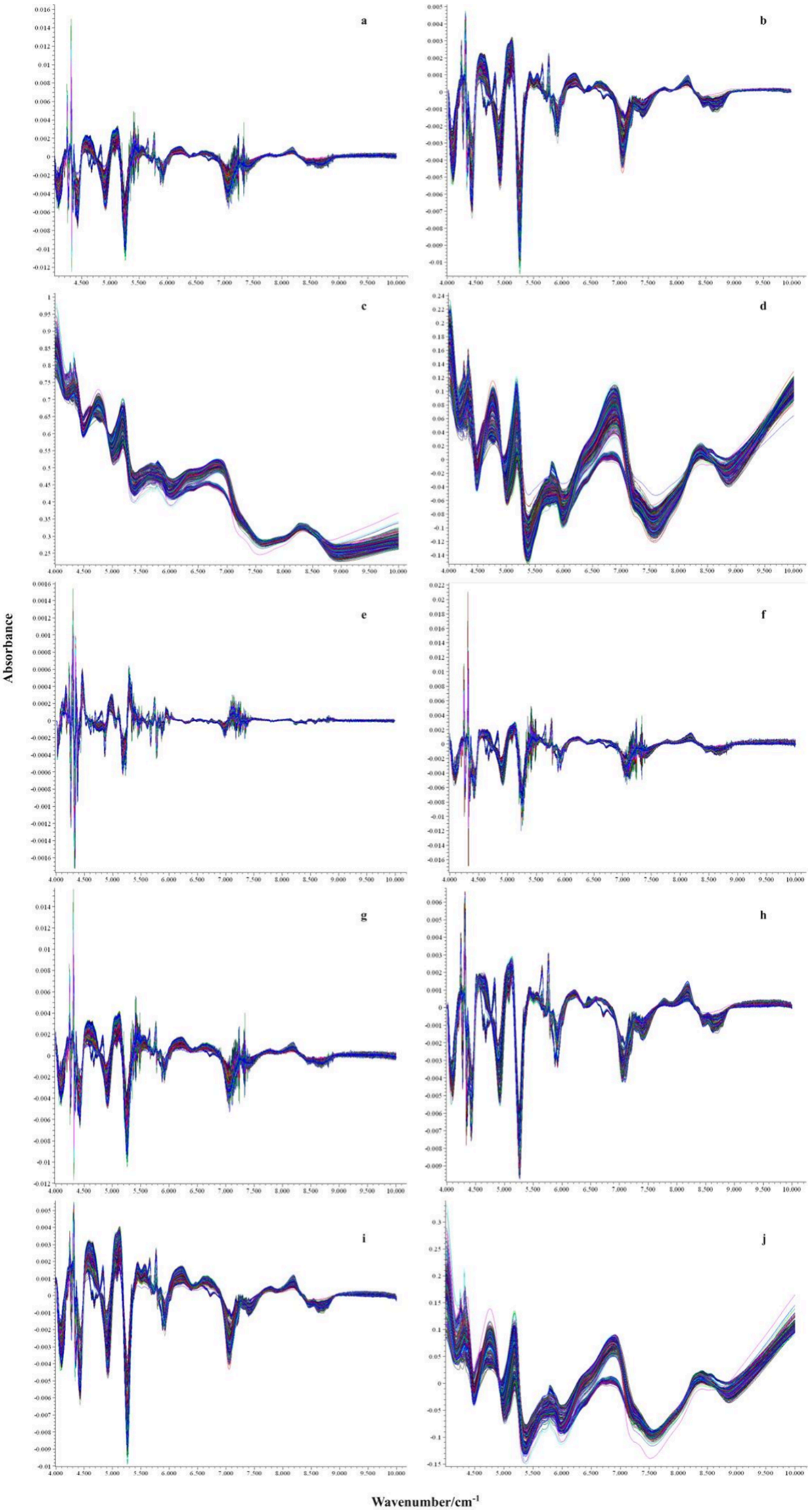


Figure 2. Near-infrared spectrum of crop grain samples for crude protein after pretreatment (a: spectrum after ND treatment; b: spectrum after SG treatment; c: spectrum after MSC treatment; d: spectrum after DT treatment; e: spectrum after ND+SG combination treatment; f: spectrum after ND+MSC combination treatment; g: spectrum after ND+DT combination treatment; h: spectrum after SG+MSC combination treatment; i: spectrum after SG+DT combination treatment; j: spectrum after MSC+DT combination treatment).

Table 4. Establishment and Prediction of Crude Protein for Crop Grain in PLS Models[a]

sample set	pretreatment	variable selection	factors	R^2_{cv}	R^2_p	RMSECV (%)	RMSEP (%)	RMSECV/ RMSEP	RPD_{cv}	RPD_P	ratio (%)	optimal model
FS	ND+MSC (N+M)	CARS (C)	8	0.863	0.944	3.195	2.515	1.270	2.698	4.247	12.5	N+D-C-P
		MC-UVE (M)	10	0.905	0.963	2.654	2.061	1.288	3.248	5.182	37.5	
	ND+DT (N+D)	**CARS (C)**	**10**	**0.919**	**0.958**	**2.454**	**2.193**	**1.119**	**3.512**	**4.869**	**50**	
		MC-UVE (M)	10	0.898	0.958	2.747	2.189	1.255	3.138	4.880	0	
CS	ND+MSC (N+M)	CARS (C)	9	0.704	0.721	2.626	2.097	1.252	1.836	1.898	50	N+D-C-P
		MC-UVE (M)	10	0.665	0.712	2.793	2.132	1.310	1.726	1.867	0	
	ND+DT (N+D)	**CARS (C)**	**9**	**0.717**	**0.612**	**2.567**	**2.476**	**1.037**	**1.877**	**1.607**	**62.5**	
		MC-UVE (M)	10	0.633	0.551	2.921	2.661	1.097	1.650	1.495	0	
MS	**ND+MSC (N+M)**	**CARS (C)**	**8**	**0.821**	**0.753**	**2.131**	**1.961**	**1.087**	**2.361**	**2.020**	**62.5**	N+M-C-P
		MC-UVE (M)	8	0.792	0.806	2.295	1.740	1.319	2.191	2.276	50	
	ND+DT (N+D)	CARS (C)	8	0.789	0.756	2.311	1.950	1.185	2.177	2.030	12.5	
		MC-UVE (M)	9	0.784	0.798	2.339	1.776	1.317	2.151	2.230	0	
KS	**ND+MSC (N+M)**	**CARS (C)**	**8**	**0.929**	**0.703**	**0.569**	**1.317**	**0.432**	**3.727**	**1.868**	**50**	N+M-C-P
		MC-UVE (M)	8	0.875	0.687	0.754	1.351	0.558	2.812	1.820	12.5	
	ND+DT (N+D)	CARS (C)	8	0.924	0.618	0.589	1.493	0.394	3.601	1.648	12.5	
		MC-UVE (M)	9	0.872	0.704	0.763	1.315	0.580	2.779	1.871	50	
SS	**ND+MSC (N+M)**	**CARS (C)**	**8**	**0.975**	**0.928**	**0.318**	**0.467**	**0.681**	**6.261**	**3.768**	**50**	N+M-C-P
		MC-UVE (M)	9	0.963	0.914	0.386	0.511	0.756	5.155	3.446	12.5	
	ND+DT (N+D)	CARS (C)	9	0.977	0.914	0.305	0.511	0.597	6.525	3.445	37.5	
		MC-UVE (M)	10	0.964	0.916	0.380	0.505	0.752	5.242	3.487	0	

[a] Note: FS, CS, MS, KS, SS: full samples, cereal set, maize set, sorghum set, soybean set; R^2_{cv}, R^2_p, RMSECV, RMSEP, RPD_{cv}, RPD_P: R-square of cross validation, R-square of prediction set, root-mean-square error of cross validation, root-mean-square error of prediction, relative percent deviation of cross-validation set, Relative percent deviation of prediction set; Factors means the number of principal factors; ratio means optimal weight ratio of data.

of 4000−9000 cm^{-1}. Specifically, prominent absorption peaks in the range of 4000−5000 cm^{-1} were attributed to C−H and O−H functional groups.[26,27] In the range of 5000−5500 cm^{-1}, significant absorption peaks were linked to C═O and free O−H functional groups. Within 5500−6500 cm^{-1}, primary absorption peaks were ascribed to C−H and S−H functional groups. In the 6500−7500 cm^{-1} range, main peaks resulted from the absorption of C−H, N−H, and O−H functional groups. Additionally, in the range of 7500−9000 cm^{-1}, primary absorption peaks were associated with C−H functional groups, while spectral absorption was relatively subdued in the 9000−10,000 cm^{-1} range.

The near-infrared spectral data were susceptible to interference from multiple noise sources during the acquisition process. These noises not only reduce the data quality but also affect the accuracy and reliability of subsequent modeling.[28] To address this issue, this study emphasized the significance of combined application for data pretreatment, in alignment with the research conducted by Jokin et al.[29] Techniques such as smoothing, differentiation, and scattering correction have proven to be effective in significantly reducing or eliminating noise and enhancing data quality. In this study, ten pretreatment methods were utilized, including ND, SG derivation,[30] MSC, DT, ND+SG, ND+MSC, ND+DT, SG +MSC, SG+DT, and MSC+DT.

The near-infrared spectrum of crop grain utilized the ten pretreatment combinations for the entire sample sets, as illustrated in Figure 2. Following several screening attempts, it was determined that the ND+MSC and ND+DT combinations were suitable for spectral preprocessing. The spectra processed with these two combinations were clearer and more accurate, displaying less noise. The curves were smoother with minimal fluctuations, and the absorption peaks were more prominent. Variable selection methods utilized CARS and MC-UVE. The PLS method was selected as the correction method to establish the model. The study then compared the model performance based on the combinations of pretreatment methods, variable selection methods, and correction methods.

3.1.2. Screening of the Optimal Near-Infrared Modeling Combination Method for Crude Protein. In this study, 20 NIRS models were developed for five different crop grain sample sets. Each index was calculated and reported, and the optimal data were selected for comprehensive evaluation and weight determination. Subsequently, the optimal near-infrared application model was chosen for each crop grain sample set, as summarized in Table 4.

To establish NIRS models for the FS, a combination of four different modeling methods was utilized. The table included eight evaluation metrics, such as R^2 and RPD factors, with a total score of 100 points and 12.5 per metric. The score of each modeling combination divided by 100 represented its optimal weight ratio. Among them, the practical application models were developed by selecting the ND+DT (N+D)-CARS (C)-PLS (P) combination. This modeling combination method had the best evaluation metrics and the highest optimal weight ratio (50%), making it the optimal modeling choice. DT was employed to eliminate the base drift caused by diffuse reflection, similar to the preprocessing method used in chemical components modeling in buckwheat based on NIRS.[31]

For the CS, four NIRS models were established and based on the optimal weight ratios of data. The selected modeling method combination was ND-C-P (62.5%) for near-infrared modeling. MS selected the modeling method combination of ND+MSC (N+M)-C-P (62.5%) to establish NIRS model. MSC effectively eliminated the impact of uneven particle size or inconsistent sample containers on the spectrum, consistent with the pretreatment method used for reducing sugar detection by NIRS.[32]

Table 5. Evaluation Model of Crude Protein for Crop Grain by NIRS[a]

sample set	factors	R^2_{cv}	R^2_P	RMSECV (%)	RMSEP (%)	RMSECV/RMSEP	RPD_{cv}	RPD_P	ARD_{cv} (%)	ARD_P (%)
FS	10	0.919	0.958	2.454	2.193	1.119	3.512	4.869	10.9	12.6
CS	9	0.717	0.612	2.567	2.476	1.037	1.877	1.607	12.1	16.7
MS	8	0.821	0.753	2.131	1.961	1.087	2.361	2.02	9.5	11.9
KS	8	0.929	0.703	0.569	1.317	0.432	3.727	1.868	2.3	7.7
SS	8	0.975	0.928	0.318	0.467	0.681	6.261	3.768	0.4	1.0

[a] Note: ARD_{cv} and ARD_P stand for average relative deviation of cross-validation set and average relative deviation of prediction set; Factors means the number of principal factors.

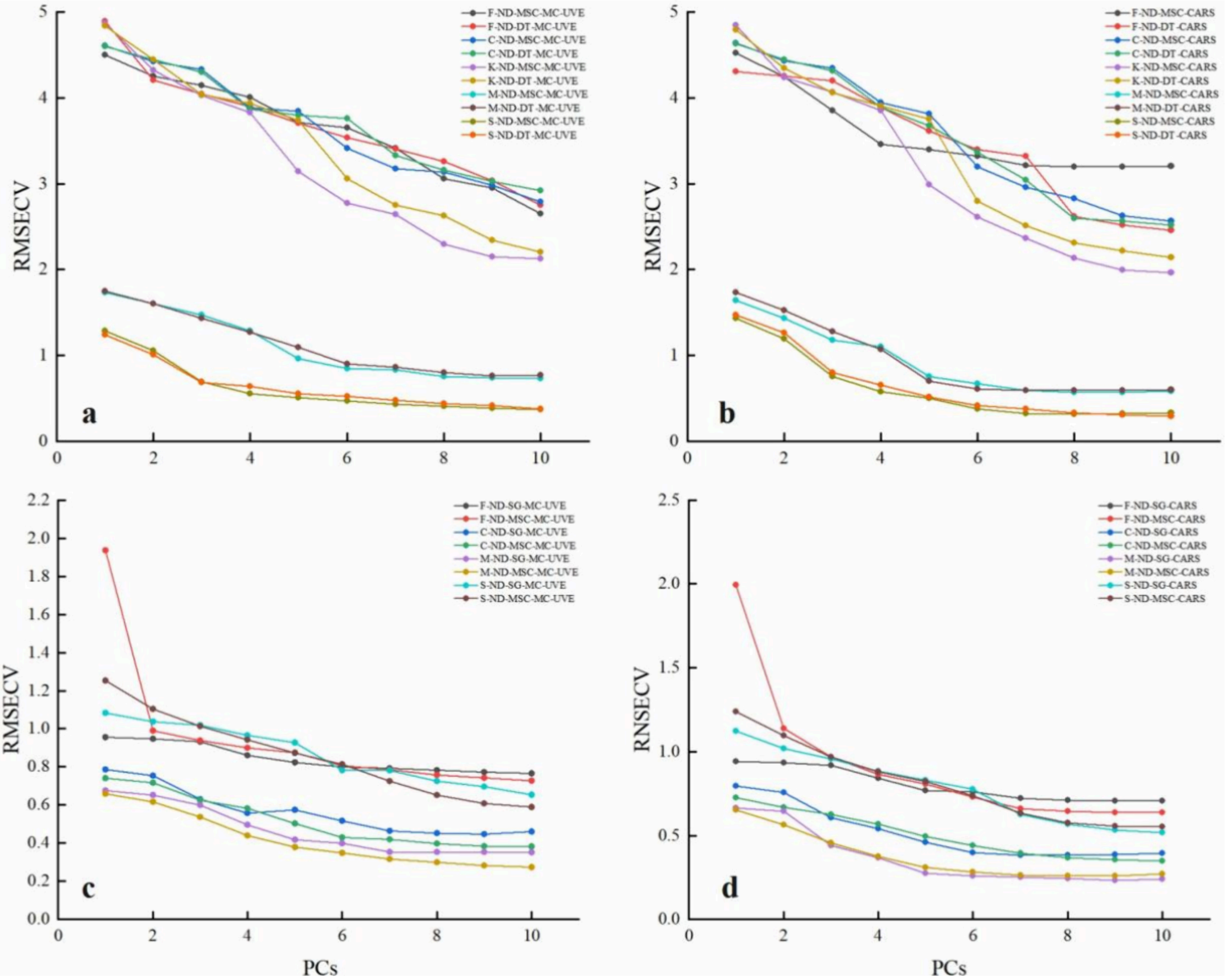


Figure 3. Trend plot of the relationship between the number of principal components (PCs) and RMSECV for crude protein (a,b) and crude fat (c,d).

The KS selected the NIRS model established using the N+M-C-P (50%) modeling method combination. Similarly, SS also selected the N+M-C-P (50%) combination to establish their NIRS model. In summary, the above analysis demonstrates that optimal modeling method combinations for establishing NIRS models for crop grain sample sets were N+D-C-P and N+M-C-P.

3.1.3. Fitting of Different Sample Set Models. The results in Table 5 were obtained through the arrangement and computational analysis of Table 4. Significant variations in crude protein content among different types of crop grain samples, leading to notable differences in the ensemble modeling outcomes. To establish NIRS models for crude protein, 1243 samples were categorized into five sets. The optimal number of principal factors (Factors) for establishing these models ranged from 8 to 10. It was critical to emphasize that the selection of Factors played an important role in determining the fitting of a model. A few Factors can result in underfitting, as the model failed to capture sufficient spectral information, while an excessive Factors may introduce noise and lead to overfitting. The optimal Factors was determined as a balance between these two extremes. For the FS models, this balance was reached at 10, where RMSECV values reached

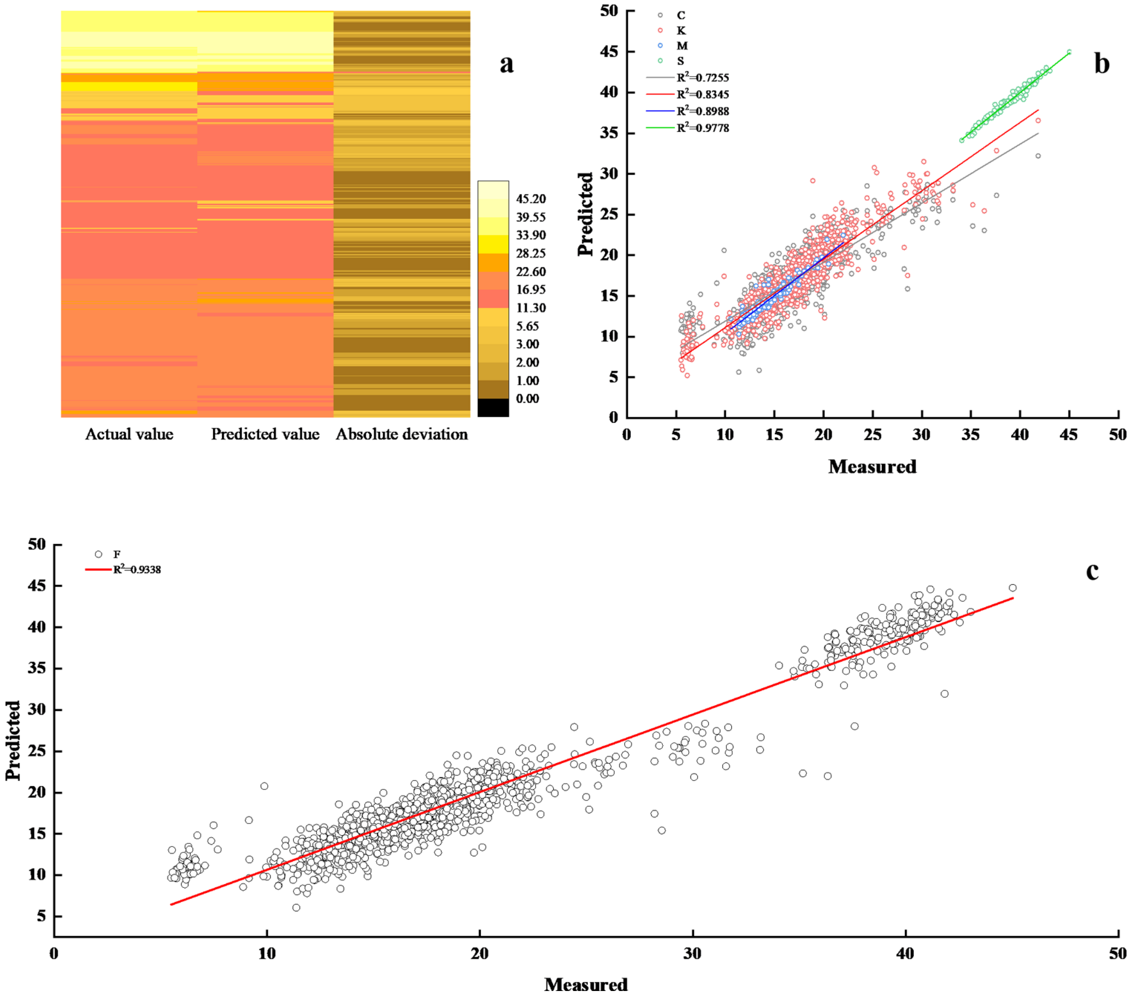


Figure 4. Scatter relationship between the true value and predicted value of the crude protein model in the near-infrared spectrum of different crop grain sample sets (a: Actual value, predicted value, and absolute deviation for FS; b: R^2 value for CS, KS, MS, and SS; c: R^2 value for FS).

their minimum, too many or too few Factors would reduce the models' fit.

In contrast, crude protein NIRS models based on the division of samples into four subclass sets showed Factors of 8, except for the CS, which had a Factor of 9. This reduced number of Factors suggested that the spectral information for the subclass sets was simpler and more homogeneous, requiring fewer components to achieve an optimal fit. This observation highlighted how sample division impacted the complexity of the resulting model. Nonetheless, relying solely on the number of Factors to evaluate model performance across different sets was methodologically unsound, as it only reflected model complexity rather than predictive accuracy. A more robust evaluation required considering performance metrics such as RMSECV, RPD, and R^2. To substantiate the choice of Factors, additional visualizations such as trend plots showing the relationship between the number of Factors and performance metrics (e.g., RMSECV) should be included to demonstrate how model performance changed with varying Factor counts (Figure 3).

The R-Square of cross validation (R^2_{cv}) for the NIRS models of five sets established by the set partitioning method ranged from 0.717 to 0.975. Among the four subclasses of sample crude protein NIRS models, SS and KS exhibited the best internal cross-validation performance, with R^2_{cv} values of 0.975 and 0.929, respectively. These were superior to MS (R^2_{cv} = 0.821) and FS (R^2_{cv} = 0.919), while CS had a lower R^2_{cv} of 0.717. The R-Square of prediction set (R^2_{p}) for five sets of crude protein models varied widely, ranging from 0.612 to 0.958. The R^2_{p} of the SS was 0.928, which was higher than KS (R^2_{p} = 0.703) and MS (R^2_{p} = 0.753). The highest R^2_{p} was observed in the FS at 0.958, while the CS exhibited the lowest R^2_{p} at 0.612. The notable difference in sample size between the prediction set and calibration set of sorghum grain (n = 136) and the sample size of maize grain (n = 906) may contribute to the lower R^2_{p}. Overall, the fitting results of the models established by the subclasses of sample sets were superior to that of the FS, consistent with the fitting effect of the protein content model of corn samples and R^2 value of 0.95.[33]

3.1.4. Robustness and Accuracy of Different Sample Set Models. Four sets were divided to establish crude protein NIRS models for crop grain with the CS exhibiting optimal RMSECV/RMSEP at 1.037, indicating the best model robustness. The RMSECV/RMSEP of MS (1.087) was superior to that of the KS (RMSECV/RMSEP = 0.432) and SS (RMSECV/RMSEP = 0.681). In contrast, when creating a single subset model for crude protein content in crop grain (FS), the RMSECV/RMSEP was 1.119 which the robustness was not as good as the CS and MS.

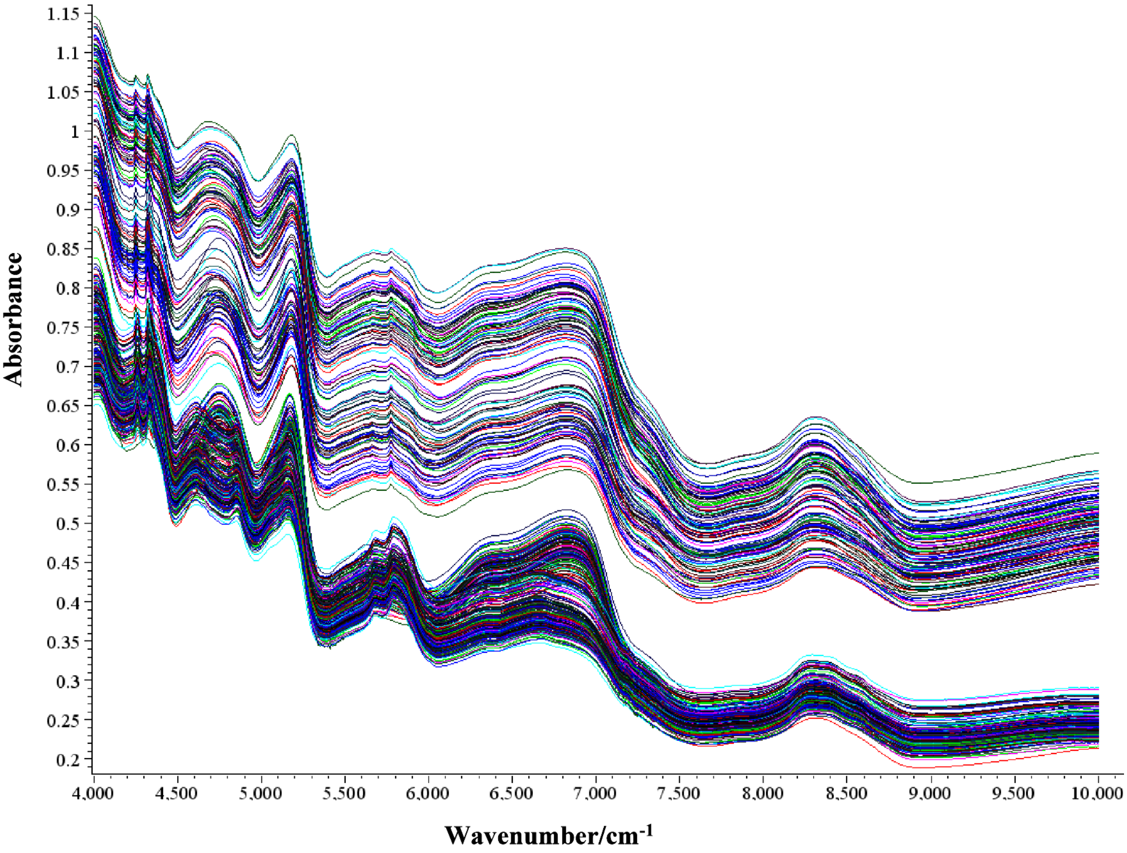


Figure 5. Original near-infrared spectrum of crop grain samples for crude fat.

The RMSECV and the RMSEP for the crude protein NIRS models of the five crop grain sample sets ranged from 0.318 to 2.567% and 0.467 to 2.476%, respectively. The SS model exhibited the smallest RMSECV (0.318%) and RMSEP (0.467%), indicating the highest fitting accuracy. This result was close to the RMSE (0.4823%) obtained for the buckwheat grain crude protein NIRS model.[34] However, RMSECV and RMSEP values alone may not provide a comprehensive picture due to the variability in sample size and distribution. Thus, we incorporated dimensionless metrics, such as ARD and RPD, to provide a more robust and comparable evaluation of model performance. These indicators enhance the reliability of cross-sample comparisons and mitigate the influence of data set variability.

The average relative deviation of cross-validation set (ARD_{cv}) and average relative deviation of prediction set (ARD_p) for the SS, KS and MS among five sample set models were relatively small. The ARD_{cv} values for these sets were 0.4, 2.3, and 9.5% respectively, indicating a higher internal prediction accuracy than FS.[24,25] The ARD_p values for these sets were 1.0, 7.7, and 11.9%, respectively, indicating a higher external prediction accuracy than FS. Therefore, the predictive accuracy of the models established by partitioning the crop grain samples was higher than that of a single comprehensive crop grain sample set.

The relative percent deviation of cross-validation set (RPD_{cv}) for the crude protein NIRS model established by different crop grain sample sets ranged from 1.877 to 6.261. To further standardize the interpretation, we clarified in the methods section that an RPD value ≥ 3 indicated strong predictive ability, while RPD < 2.5 suggested weaker predictive performance. These definitions enhanced the interpretability of RPD results and provided a standardized framework for evaluating model performance across data sets. Among them, the three set models of SS (RPD_{cv} = 6.261), KS (RPD_{cv} = 3.727), and FS (RPD_{cv} = 3.512) in the crop grain crude protein model established by subsets were capable of meeting the quantitative analysis of crude protein content in crop grain samples. These models can be directly applied to near-infrared analysis and detection of crude protein in biomass samples. The enhanced analysis confirmed that the individual subsets (e.g., SS) outperformed the single comprehensive FS model in terms of both robustness and predictive power. However, RPD_{cv} values of CS and MS were both less than 2.5, indicating that these models needed to be optimized.

The relative percent deviation of prediction set (RPD_p) for the crude protein NIRS models of the five crop grain sample sets ranged from 1.607 to 4.869. The FS (RPD_p = 4.869) and SS (RPD_p = 3.768) models were suitable for prediction and evaluation of crude protein content external validation in crop grain samples, while the RPD_p values of CS, MS, and KS were less than 2.5, indicating the need for model optimization. It was highlighted that RPD_p, in conjunction with R_p^2 and ARD_p, provided a more balanced assessment of external validation accuracy across sample sets.

3.1.5. Correlation between the Modeling Set and External Prediction Set of Different Sample Set Models.

The correlation between the reference value and predicted value of crude protein content models for the five crop grain sample sets was shown in Figure 4. The R-square (R^2) value, which measures the goodness of fit of the model, ranged from 0.7255 to 0.9778. This indicated a strong linear relationship with

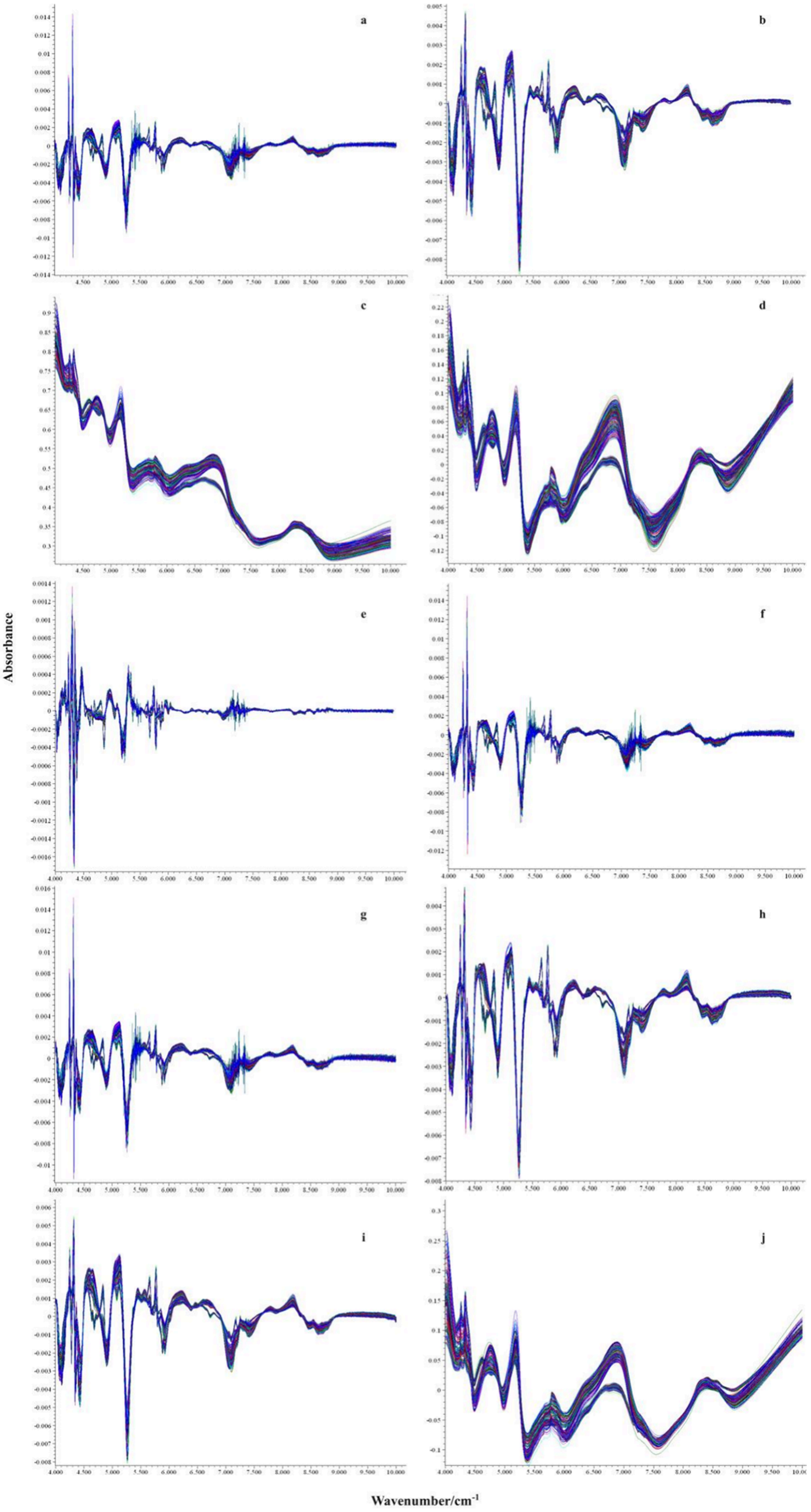


Figure 6. Near-infrared spectrum of crop grain samples for crude fat after pretreatment (a: spectrum after ND treatment; b: spectrum after SG treatment; c: spectrum after MSC treatment; d: spectrum after DT treatment; e: spectrum after ND+SG combination treatment; f: spectrum after ND+MSC combination treatment; g: spectrum after ND+DT combination treatment; h: spectrum after SG+MSC combination treatment; i: spectrum after SG+DT combination treatment; j: spectrum after MSC+DT combination treatment).

Table 6. Establishment and Prediction of Crude Fat for Crop Grain in PLS Models[a]

sample set	pretreatment	variable selection	factors	R^2_{cv}	R^2_p	RMSECV (%)	RMSEP (%)	RMSECV/RMSEP	RPD_{cv}	RPD_P	ratio (%)	optimal model
FS	ND+SG (N+S)	CARS (C)	7	0.993	0.990	0.721	0.895	0.806	11.581	9.899	12.5	N+M-C-P
		MC-UVE (M)	7	0.991	0.990	0.792	0.901	0.879	10.543	9.834	25	
	ND+MSC (N+M)	**CARS (C)**	**7**	**0.994**	**0.991**	**0.660**	**0.830**	**0.795**	**12.652**	**10.675**	**75**	
		MC-UVE (M)	9	0.992	0.990	0.741	0.887	0.836	11.269	9.989	0	
CS	ND+SG (N+S)	CARS (C)	7	0.874	0.575	0.382	0.621	0.615	2.801	1.546	12.5	N+M-C-P
		MC-UVE (M)	7	0.815	0.690	0.463	0.531	0.873	2.311	1.808	25	
	ND+MSC (N+M)	**CARS (C)**	**9**	**0.891**	**0.722**	**0.356**	**0.503**	**0.709**	**3.006**	**1.909**	**37.5**	
		MC-UVE (M)	9	0.865	0.736	0.396	0.490	0.810	2.702	1.959	25	
KS	**ND+SG (N+S)**	**CARS (C)**	**9**	**0.851**	**-0.360**	**0.532**	**1.522**	**0.350**	**1.485**	**0.440**	**37.5**	N+S-C-P
		MC-UVE (M)	10	0.776	0.011	0.654	1.297	0.504	1.208	0.517	0	
	ND+MSC (N+M)	CARS (C)	8	0.827	−0.078	0.574	1.354	0.424	1.376	0.495	12.5	
		MC-UVE (M)	9	0.807	0.227	0.607	1.147	0.529	1.301	0.584	37.5	
SS	**ND+SG (N+S)**	**CARS (C)**	**8**	**0.906**	**0.258**	**0.243**	**0.566**	**0.428**	**5.679**	**2.332**	**37.5**	N+S-C-P
		MC-UVE (M)	7	0.801	0.476	0.353	0.476	0.741	3.909	2.773	12.5	
	ND+MSC (N+M)	CARS (C)	7	0.888	0.412	0.265	0.504	0.525	5.208	2.619	12.5	
		MC-UVE (M)	9	0.873	0.673	0.282	0.376	0.750	4.894	3.511	37.5	

[a]Note: FS, CS, KS, SS: full samples, cereal set, sorghum set, soybean set; R^2_{cv}, R^2_p, RMSECV, RMSEP, RPD_{cv}, RPD_P: R-square of cross validation, R-square of prediction set, root-mean-square error of cross validation, root-mean-square error of prediction, Relative percent deviation of cross-validation set, relative percent deviation of prediction set; Factors means the number of principal factors; Ratio means optimal weight ratio of data.

Table 7. Evaluation Model of Crude Fat for Crop Grain by NIRS[a]

sample set	factors	R^2_{cv}	R^2_p	RMSECV (%)	RMSEP (%)	RMSECV/RMSEP	RPD_{cv}	RPD_P	ARD_{cv} (%)	ARD_P (%)
FS	7	0.994	0.991	0.660	0.830	0.795	12.652	10.675	7	7
CS	9	0.891	0.722	0.356	0.503	0.709	3.006	1.909	6	13
KS	9	0.851	−0.360	0.532	1.522	0.350	1.485	0.440	1	6
SS	8	0.906	0.258	0.243	0.566	0.428	5.679	2.332	3	11

[a]Note: ARD_{cv} and ARD_P stand for average relative deviation of cross-validation set and average relative deviation of prediction set; factors means the number of principal factors.

minimal deviations between the modeling and prediction sets. When considering the four-group division, the SS model exhibited an R^2 value of 0.9778, indicating an optimal fit and linear relationship. The FS model (R^2 = 0.9338) also demonstrated a satisfactory fit and linear relationship compared to the KS (R^2 = 0.8988), MS (R^2 = 0.8345) and CS (R^2 = 0.7255) models (Figure 4). The R^2 served as a critical metric for assessing model fit quality, while other metrics provided complementary insights into predictive performance. This multidimensional approach enhanced the robustness of our conclusions.

3.2. Crude Fat NIRS Model.

3.2.1. Original Spectral Characteristics of Samples and Pretreatment. 415 crop grain samples were scanned for near-infrared original spectrum (Figure 5). Each near-infrared spectrum exhibited absorption peaks between 4000 and 9000 cm^{-1}. Within the range of 4000−5000 cm^{-1}, the main absorption peaks were associated with C−H and O−H functional groups. The absorption peaks within the 5500−7500 cm^{-1} range were primarily due to C−H, S−H, N−H, and O−H functional groups. The primary absorption peaks in the 7500−9000 cm^{-1} range were associated with C═H functional groups. The spectra contained a wealth of information on H-containing groups such as C−H bonds, O−H bonds, and N−H bonds, making it suitable for establishing a prediction model for crude fat content.

In this study, ten different pretreatment approaches were applied to the near-infrared spectra of crop grain, such as ND, SG, MSC, DT, and their pairwise combinations. The pretreatment results were illustrated in Figure 6, highlighting the ND+SG and ND+MSC combinations were more suitable for preprocessing. Variable selection methods and correction methods were the same as in the previous study.

3.2.2. Screening of the Optimal Near-Infrared Modeling Combination Method for Crude Fat. Based on 8 near-infrared model evaluation indexes and chemometrics knowledge, comprehensive evaluation and weighted optimization were conducted for the optimal data of each index of 16 NIRS models from 4 crop grain sets. Optimal near-infrared application models were selected for each crop grain sample set (Table 6).

Four different modeling methods were combined to establish NIRS models for the FS. The optimal weight ratios of data were as follows: N+M-C-P (75%) > ND+SG (N+S)-MC-UVE (M)-P (25%) > N+S-C-P (12.5%) > N+M-M-P (0). Therefore, the NIRS model of N+M−C-P for FS was selected as the optimal modeling method combination.

Four NIRS models were established for the CS, and according to the optimal weight ratio of data, the NIRS modeling method combination with the largest ratio was N+M-C-P (37.5%). For KS, the NIRS model was established using the N+S-C-P (37.5%) modeling method combination. SS selected the modeling method combination of N+S-C-P (37.5%) to establish the NIRS model.

In summary, the above analysis showed that the optimal modeling method combinations for establishing NIRS models for crop grain sample sets were N+M-C-P and N+S-C-P. SG derivation eliminated the baseline drift in the spectrum and improved the resolution of the spectrum, which was similar to

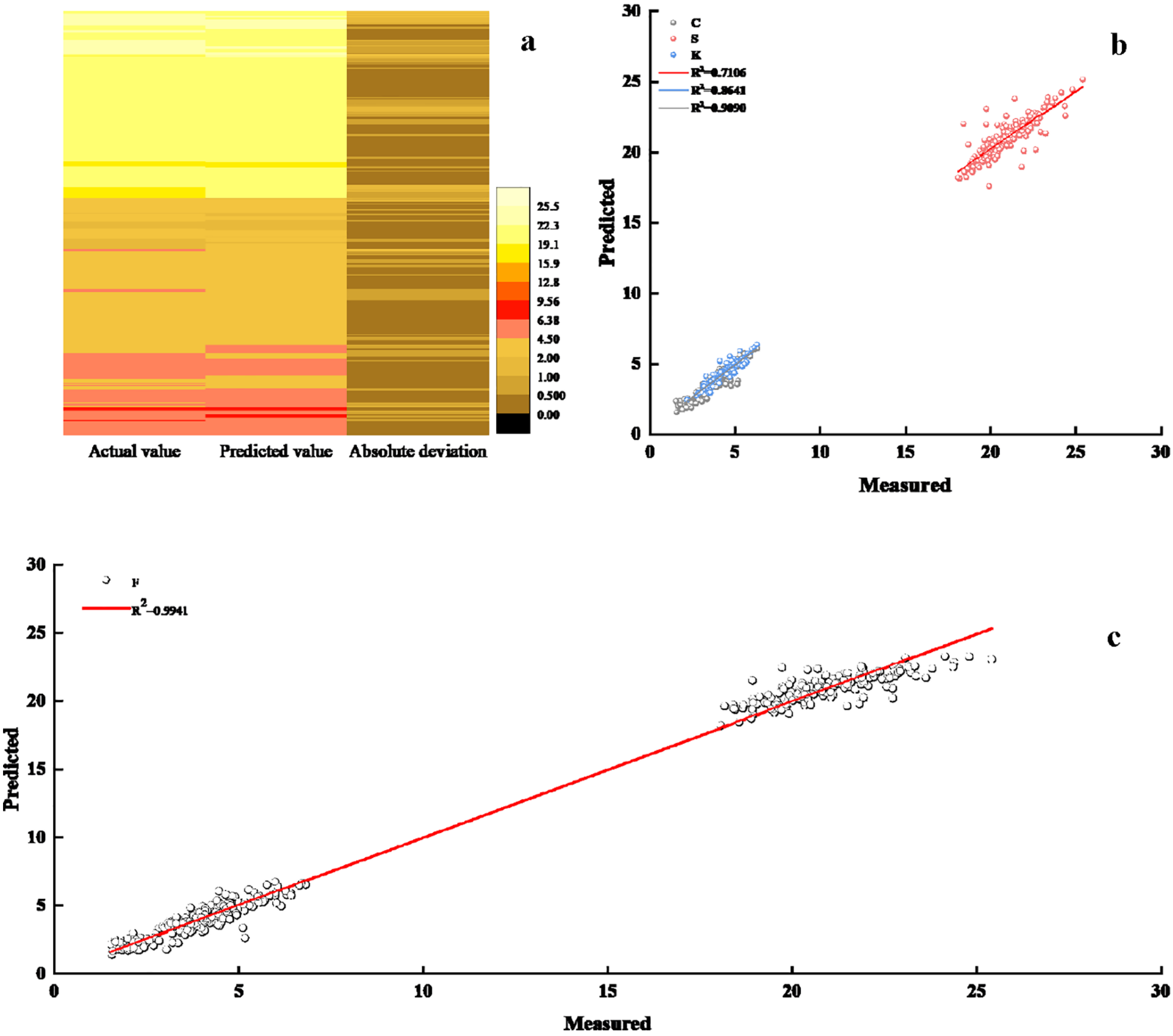


Figure 7. Scatter relationship between the true value and predicted value of the crude fat model in the near-infrared spectrum of different crop grain samples (a: Actual value, predicted value, and absolute deviation for FS; b: R^2 value for CS, SS, and KS; c: R^2 value for FS).

the preprocessing method used for the detection of rice amylose content by NIRS.[35,36]

3.2.3. Fitting of Different Sample Set Models. As Table 7 indicates, the number of Factors in the NIRS models for crude fat varied between 7 and 9, depending on the crop grain set. Specifically, the NIRS models for FS and SS had 7 and 8 Factors, respectively. However, the models for KS and CS had a relatively high number of Factors (9) (Figure 3c,d).

The coefficient of determination for internal cross-validation R^2_{cv} of the NIRS models established using four sets of samples ranged from 0.851 to 0.994. Among them, the internal cross validation performance of the FS and SS models was the most impressive, with R^2_{cv} values of 0.994 and 0.906, respectively. These results outperformed the CS (R^2_{cv} = 0.891), while the KS had a lower R^2_{cv} of 0.851.

R^2_p also exhibited a wide range, varying from −0.360 to 0.991 across the four sets of models. The FS model demonstrated the best external prediction performance, with an R^2_p of 0.991, surpassing the CS (R^2_p = 0.722) and SS (R^2_p = 0.258). The KS model had a poorer external predictive fitting with an R^2_p of −0.360.

The consistency of the fitting effect with the model established for protein and fat content in rice by NIRS was noteworthy,[37] as the coefficient of determination R^2 value exceeded 0.9, indicating similar results. The notable difference in sample size between the prediction set and calibration set of sorghum grain (n = 118) may have contributed to the lower R^2_p. Furthermore, dividing the samples into one set for establishing the NIRS crop grain crude fat model provided better fitness compared to dividing them into three sets.

3.2.4. Robustness and Accuracy of Different Sample Set Models. RMSECV and RMSEP values of the NIRS models established using the four sample sets ranged from 0.350 to 0.795. The optimal RMSECV/RMSEP of 0.795 was achieved when a single sample set (FS) was used to establish the crude fat NIRS model. In contrast, when three sets were utilized for model establishment, the robustness of the CS (RMSECV/RMSEP = 0.709) was superior to that of SS (RMSECV/

RMSEP = 0.428) and KS (RMSECV/RMSEP = 0.350). ARD_{cv} of the KS and SS models were 1 and 3%, respectively, indicating high internal prediction accuracy. ARD_p values for KS and FS were 6 and 7% respectively, suggesting high accuracy in external prediction.[24,25] Models established using three crop grain sample sets exhibited lower average relative deviations compared to the FS.

RMSECV and RMSEP of the NIRS models for crude fat in four crop grain sample sets varied from 0.243 to 0.660% and 0.503 to 1.522%, respectively. Among them, the lowest RMSECV belonged to the SS model at 0.243%, indicating the highest accuracy. The RMSECV values for the KS and CS models were 0.532 and 0.356%, respectively, which were both lower than that of FS (RMSECV = 0.660%). The lowest RMSEP was observed in the CS model at 0.503%, while the SS model had an RMSEP of 0.566%, which was lower than that of the FS (RMSEP = 0.830%) and KS (RMSEP = 1.522%).

RPD_{cv} range of the NIRS models for crude fat established using different crop grain sample sets was 1.485−12.652. The FS (RPD_{cv} = 12.652), SS (RPD_{cv} = 5.679), and CS (RPD_{cv} = 3.006) models were capable of meeting the requirements for quantitative analysis of crop grain crude fat content. However, the RPD_{cv} of the KS model was less than 2.5, indicating that further optimization of the model was necessary. RPD_p of NIRS models for crude fat in the four crop grain sample sets was 0.440−10.675. The FS (RPD_p = 10.675) model was suitable for external validation of crude fat content in crop grain samples, while the RPD_p values for the CS, KS, and SS were less than 2.5, indicating the need for further optimization of these models.

3.2.5. Correlation between Modeling Set and External Prediction Set of Different Sample Set Models. The correlation between reference value and predicted value of the crude fat content models for the four crop grain sample sets was presented in Figure 7. The R^2 value, which represents the goodness of fit, ranged from 0.7106 to 0.9941. The modeling and prediction sets exhibited small deviations, indicating a strong linear relationship. When a single set (FS) was used for partitioning, the R^2 value was 0.9941, indicating the optimal fitting and linear relationship. The fitting and linear relationship of the CS (R^2 = 0.9090) was better than that of the KS (R^2 = 0.8641) and SS (R^2 = 0.7106).

3.3. Discussion. In this study, the reproducibility factors included (1) the selection of near-infrared spectral pretreatment methods. We used multiple pretreatment methods, such as ND, SG, MSC, DT, and their combinations. (2) The sample set division method. Such as the crude protein study, we applied five different sample set division methods, including FS and four subclass sets (CS/MS/KS/SS), which led to differences in model performance. (3) NIRS modeling methods. Different pretreatment methods were combined with various variable selection techniques, and each set had four different modeling combinations, from which the optimal method was selected based on evaluation.

NIRS utilized specific wavelength bands and spectral absorption information from functional groups such as N−H, C−H, and O−H bonds to enable rapid and nondestructive quantitative analysis of crude protein and crude fat. This approach ensured the reliability of the models and the accuracy of the measurements, thereby providing technical support for the detection of crop grain composition. In crude protein NIRS models, the Factors of MS, KS, and SS models were 8, and the internal cross-validation results of SS and KS models demonstrated the best performance, with R^2_{cv} values of 0.975 and 0.929. FS model exhibited the highest R^2_p of 0.958 for external prediction performance. The models established for the division of subclass sample sets demonstrate the highest fitting degree, which was consistent with the fitting effect (R^2 = 0.95) for the protein content model of corn samples.[25] The optimal RMSECV/RMSEP for CS model was 1.037, the ARD_{cv} and ARD_p for the models of SS, MS, and KS were relatively lower than FS, which exhibited higher robustness than FS model. SS model exhibited the lowest RMSECV (0.318%) and RMSEP (0.467%), indicating the highest level of fitting accuracy for this model. This highlighted the capability of NIRS to simultaneously determine the crude protein of crop grain and handle complex interactions and variability across subclass divisions. This finding was consistent the data obtained by Zhao et al.,[34] who established an NIRS model for the crude protein content of quinoa grains (RMSE = 0.4823).

In crude fat NIRS models, the Factors of FS were 7 and FS model exhibited the best fit (R^2_{cv} = 0.994, R^2_p = 0.991). The external predictive fit of the KS was relatively poor. The fit of the model established for the division of one sample set was superior to that for the division of three sample sets, which was consistent with the finding results of Lu et al.,[37] who utilized NIRS to establish models for indicators such as protein and fat in rice. The ARD_{cv} values of KS and SS were relatively low at 1 and 3%, respectively. The ARD_p of KS was 6% and it demonstrated the highest level of accuracy. The minimum RMSECV of SS model was 0.243, and the minimum RMSEP of CS model was 0.503%. Therefore, the fitting accuracy of the crude fat NIRS models established by the division into three sets was the highest.

Overall, the crude protein models established by dividing the samples into 4 subclass sets had a higher degree of fitting, accuracy and correlation. But, the FS model for crude fat exhibited significant advantages in terms of fit, robustness, accuracy, and correlation. The success of these NIRS models lied in their ability to quickly process spectral data and extract relevant chemical information, enabling precise quantification of protein and fat components across diverse sample divisions. Different model performance can be obtained by using different sample set division methods.

4. CONCLUSIONS

In this study, a total of 1243 and 415 crop grain samples were screened and classified into 5 sets and 4 sets, respectively, using sample set division method. The aim was to establish NIRS models for the prediction of crude protein and crude fat content in crop grain samples. The main findings of the crude protein NIRS models are as follows: The optimal modeling method combinations were N+D-C-P and N+M-C-P. ND was effective in eliminating noise, while MSC effectively mitigated the impact of uneven particle size or inconsistent sample containers on the spectrum. DT was used to eliminate base drift caused by diffuse reflection. The internal cross-validation performance of the SS and KS was the best, with R^2_{cv} values of 0.975 and 0.929, respectively. The RMSECV/RMSEP ratio of CS was optimal at 1.037, and the ARD_{cv} and ARD_p values of the SS, MS, and KS were smaller, the 4 subsets showed the highest robustness. The SS model had the lowest RMSECV (0.318%) and RMSEP (0.467%), demonstrating the highest fitting accuracy. The SS (RPD_{cv} = 6.261), KS (RPD_{cv} = 3.727), and FS (RPD_{cv} = 3.512) data sets provided sufficient accuracy for the quantitative analysis of crop grain crude protein content. FS (RPD_p = 4.869) and SS (RPD_p = 3.768) data sets were suitable for both prediction and evaluation of external validation of the crude protein content of crop grain samples. The model established using samples divided into 4 subclasses exhibited the highest degree of model fitting, accuracy, and correlation. Therefore, it is feasible to employ near-infrared diffuse reflectance spectroscopy for nondestructive determination of the crude protein content in crop grain. The obtained results meet the requirements of quantitative and qualitative detection.

The main results for crude fat NIRS models were as follows: The optimal modeling method combinations were N+M-C-P and N+S-C-P. The SG derivative can eliminate baseline drift in spectra and enhance the spectral resolution. KS and SS had the smaller ARD_{cv} values, at 1% and 3%, respectively. The SS and CS had the smallest RMSECV (0.243%) and RMSEP (0.503), respectively. This showed the crude fat NIRS model after dividing the 3 subclasses achieved better fitting accuracy. The FS exhibited the best fit (R^2_{cv} = 0.994, R^2_p = 0.991), the highest RMSDCV/RMSEP (0.795) and RPD values. The FS model could be used for predicting and evaluating the crude fat content of crop grain samples for external validation. Furthermore, the FS model established using one sample set demonstrated clear advantages in terms of fitting, robustness, accuracy, and correlation. The FS, SS, and CS models were suitable for quantitative analysis of crop grain crude fat content. This finding provides reliable theoretical and technical support for near-infrared detection of crude fat content in crop grain.

AUTHOR INFORMATION

Corresponding Author

Xiaoyu Wang − *Agricultural Equipment Institute of Hunan/Hunan Intelligent Agriculture Engineering Technology Research Center/Hunan Branch Center of National Energy R&D Center for Non-Food Biomass, Changsha 410125, China;* orcid.org/0000-0003-4786-7368; Email: xiao_yu_100@163.com

Authors

Qing Yang − *Agricultural Equipment Institute of Hunan/Hunan Intelligent Agriculture Engineering Technology Research Center/Hunan Branch Center of National Energy R&D Center for Non-Food Biomass, Changsha 410125, China*

Yujiao Li − *Agricultural Equipment Institute of Hunan/Hunan Intelligent Agriculture Engineering Technology Research Center/Hunan Branch Center of National Energy R&D Center for Non-Food Biomass, Changsha 410125, China*

Jie Li − *Agricultural Equipment Institute of Hunan/Hunan Intelligent Agriculture Engineering Technology Research Center/Hunan Branch Center of National Energy R&D Center for Non-Food Biomass, Changsha 410125, China*

Zhiyou Zhang − *Agricultural Equipment Institute of Hunan/Hunan Intelligent Agriculture Engineering Technology Research Center/Hunan Branch Center of National Energy R&D Center for Non-Food Biomass, Changsha 410125, China*

Qiqi Liu − *Yueyang Academy of Agriculture Sciences and Researches, Yueyang 414022, China*

Ge Guo − *College of Agronomy and Biotechnology/National Energy R&D Center for Non-Food Biomass, China Agricultural University, Beijing 100193, China*

Shuang Wang − *College of Computer and Mathematics, Central South University of Forestry and Technology, Changsha 410004, China*

Guanghui Xie − *College of Agronomy and Biotechnology/National Energy R&D Center for Non-Food Biomass, China Agricultural University, Beijing 100193, China*

Complete contact information is available at:
https://pubs.acs.org/10.1021/acsomega.4c09155

Author Contributions

[#]Q.Y. and Y.L. are cofirst authors of the article.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This research was supported by Key Research and Development Program of Hunan Province of China (No. 2024JK2159/2023NK2011).

REFERENCES

(1) Yang, Q. Study on near infrared spectrometry for quantitative analysis and its analytical application. *Southwest Univ.* **2009**, .

(2) Xu, G. T.; Yuan, H. F.; Lu, W. Z. Modern near-infrared spectroscopy technology and application progress. *Spectrosc. Spectral Anal.* **2000**, *20* (2), 134−142.

(3) Wang, J. W.; Ye, S. Research progress on detection of food ingredients by near infrared spectroscopy. *China Condiment* **2021**, *46* (9), 6.

(4) Chu, X. L.; Shi, Y. P.; Chen, P.; Li, J. Y.; Xu, Y. P. Research and application progresses of near infrared spectroscopy analytical technique in China in past five years. *J. Instrum. Anal.* **2019**, *38* (5), 603−611.

(5) Filgueiras, R.; Almeida, T. S.; Mantovani, E. C.; Dias, S. H. B.; Fernandes-Filho, E. I.; da Cunha, F. F.; Venancio, L. P. Soil water content and actual evapotranspiration predictions using regression algorithms and remote sensing data. *Agric. Water Manag.* **2020**, *241*, No. 106346.

(6) Okere, E. E.; Arendse, E.; Nieuwoudt, H.; Perold, W. J.; Opara, U. L. Non-destructive evaluation of the quality characteristics of pomegranate kernel oil by fourier transform near-infrared and mid-infrared spectroscopy. *Front. Plant Sci.* **2022**, *13*, No. 867555.

(7) Zhang, Y.; Cong, Q.; Xie, Y. F.; Zhao, B. Progress in application of near infrared spectroscopy technology in agriculture. *Trans. CSAE* **2007**, *23* (10), 285−290.

(8) Greenberg, I.; Linsler, D.; Vohland, M.; Ludwig, B. Robustness of visible near-infrared and mid-infrared spectroscopic models to changes in the quantity and quality of crop residues in soil. *Soil Sci. Soc. Am. J.* **2020**, *84* (3), 963−977.

(9) Šramková, Z.; Gregová, E.; Šturdík, E. Chemical composition and nutritional quality of wheat grain. *Acta Chim. Slovaca* **2009**, *2* (1), 115−138.

(10) Yang, Z. Y.; Cheng, Z.; Su, P. Y.; Wang, C.; Qin, M. X.; Song, X. Y.; Xiao, L. J.; Yang, W. D.; Feng, M. C.; Zhang, M. J. A model for the detection of *β*-glucan content in oat grain based on near infrared spectroscopy. *J. Food Compos. Anal.* **2024**, *129*, No. 106105.

(11) Liu, J. M.; Luo, X.; Zhang, D. J.; Wang, C. Q.; Chen, Z. G.; Zhao, X. Y. Rapid determination of rice protein content using near-infrared spectroscopy coupled with feature wavelength selection. *Infrared Phys. Technol.* **2023**, *135*, No. 104969.

(12) Rahman, A.; Cho, B. K. Assessment of seed quality using non-destructive measurement techniques: a review. *Seed Sci. Res.* **2016**, *26* (4), 285−305.

(13) Mehmood, T.; Liland, K. H.; Snipen, L.; Sæbø, S. A review of variable selection methods in partial least squares regression. *Chemom. Intell. Lab. Syst.* **2012**, *118*, 62−69.

(14) Cai, W.; Li, Y.; Shao, X. A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra. *Chemom. Intell. Lab. Syst.* **2008**, *90* (2), 188−194.

(15) Zhang, X. Y.; Wang, Y. J.; Liu, R. X.; Shen, B. H.; Wang, J. Y.; Yan, Y. L.; Kang, D. M. Application of near-infrared spectroscopy technology to analyze protein content in single kernel maize seed. *J. China Agric. Univ.* **2017**, *22* (05), 25−31.

(16) Xu, L. L.; Liu, J. M.; Wang, C. Q.; Li, Z. J.; Zhang, D. J. Rapid determination of the main components of corn based on near-infrared spectroscopy and a BiPLS-PCA-ELM model. *Appl. Opt.* **2023**, *62* (11), 2756−2765.

(17) Li, D. P.; Wang, G. Y. Comparative Study on Testing Protein and Oil Contents of Soybean by Spectroscopy and Chemical Analysis Method. *Mod. Agric. Sci. Technol.* **2015**, *9*, 295−302.

(18) Yang, H. E.; Kim, N. W.; Lee, H. G.; Kim, M. J.; Sang, W. G.; Yang, C.; Mo, C. Prediction of protein content in paddy rice (Oryza sativa L.) combining near-infrared spectroscopy and deep-learning algorithm. *Front. Plant Sci.* **2024**, *15*, No. 1398762.

(19) Lin, L. H.; Lu, F. M.; Chang, Y. C. Prediction of protein content in rice using a near-infrared imaging system as diagnostic technique. *Int. J. Agric. Biol. Eng.* **2019**, *12* (2), 195−200.

(20) Kamboj, U.; Guha, P.; Mishra, S. Comparison of PLSR, MLR, SVM regression methods for determination of crude protein and carbohydrate content in stored wheat using near Infrared spectroscopy. *Mater. Today: Proc.* **2022**, *48*, 576−582.

(21) Song, C. X.; Liu, J. M.; Wang, C. Q.; Li, Z. J.; Zhang, D. J.; Li, P. F. Rapid identification of adulterated rice based on data fusion of near-infrared spectroscopy and machine vision. *J. Food Meas. Charact.* **2024**, *18* (5), 3881−3892.

(22) NY/T 2419-2013. *Determination of total nitrogen in plant-Automatic kjeldahl apparatus method*, 2013. https://www.chinesestandard.net/PDF/English.aspx/NYT2419-2013?Redirect.

(23) GB 5009.6-2016. *National food safety standard-Determination of fat in foods*, 2016. https://www.svscr.cz/wp-content/files/obchodovani/GB_5009.6-2016_Fat-in-foods.pdf.

(24) Feng, Y. C.; Zhang, Q.; Hu, C. Q. Study on the selection of parameters for evaluating drug NIR universal quantitative models. *Spectrosc. Spectral Anal.* **2016**, *36* (8), 2447−2454.

(25) Cozzolino, D.; Kwiatkowski, M. J.; Parker, M.; Cynkar, W. U.; Dambergs, R. G.; Gishen, M.; Herderich, M. J. Prediction of phenolic compounds in red wine fermentations by visible and near infrared spectroscopy. *Anal. Chim. Acta* **2004**, *513* (1), 73−80.

(26) Pasquini, C. Near infrared spectroscopy: A mature analytical technique with new perspectives−A review. *Anal. Chim. Acta* **2018**, *1026*, 8−36.

(27) Xu, F.; Yu, J.; Tesso, T.; Dowell, F.; Wang, D. Qualitative and quantitative analysis of lignocellulosic biomass using infrared techniques: a mini-review. *Appl. Energy* **2013**, *104*, 801−809.

(28) Wang, Y. J.; Li, M. H.; Li, L. Q.; Ning, J. M.; Zhang, Z. Z. Green analytical assay for the quality assessment of tea by using pocket-sized NIR spectrometer. *Food Chem.* **2021**, *345*, No. 128816.

(29) Ezenarro, J.; Riu, J.; Ahmed, H. J.; Busto, O.; Giussani, B.; Boqué, R. Measurement errors and implications for preprocessing in miniaturised near-infrared spectrometers: Classification of sweet and bitter almonds as a case of study. *Talanta* **2024**, *276*, No. 126271.

(30) Savitzky, A.; Golay, M. J. E. Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* **1964**, *36* (8), 1627−1639.

(31) Zhang, J.; Guo, J.; Zhang, M. L.; Zhang, X. Establishment of rapid detection model of buckwheat nutritional components based on near infrared spectroscopy. *J. Chin. Cereals Oils* **2020**, *35* (6), 151−158.

(32) Zhang, X.; Shan, Y.; Li, S. F. Near-infrared determination of reducing sugar content in honey based on multiplicative scatter correction partial least square (MSC/PLS) method. *Food Mach.* **2009**, *25* (6), 109−112.

(33) Fontaine, J.; Schirmer, B.; Hörr, J. Near-infrared reflectance spectroscopy (NIRS) enables the fast and accurate prediction of essential amino acid contents. 2. Results for wheat, barley, corn, triticale, wheat bran/middlings, rice bran, and sorghum. *J. Agric. Food Chem.* **2002**, *50* (14), 3902−3911.

(34) Zhao, L. H.; Gong, Y. Y.; Zhang, J.; Lin, C. B.; Wang, Y.; Li, X. W.; Dai, X.; Jiang, Y. Rapid determination of quinoa seeds crude protein content using near infrared spectroscopy. *Sci. Technol. Food Ind.* **2020**, *41* (15), 233−236.

(35) Liu, Y. C.; Li, Y. Y.; Peng, Y. K.; Wang, F.; Yan, S.; Ding, J. G. Non-destructive Rapid Detection of Rice Amylose Content by Near-Infrared Diffuse Transmission Optical Compensation Method. *Chin. J. Anal. Chem.* **2019**, *47* (5), 785−793.

(36) Engel, J.; Gerretzen, J.; Szymańska, E.; Jansen, J. J.; Downey, G.; Blanchet, L.; Buydens, L. M. Breaking with trends in pre-processing? *TrAC. Trends Anal. Chem.* **2013**, *50*, 96−106.

(37) Lu, H.; Peng, B. Q.; Feng, X. Y.; Shen, X. F. Model optimization for determination of amylose, protein, fat and moisture content in rice by near-infrared spectroscopy. *China Rice* **2020**, *26* (6), 55−59.