foods

MDPI

Article

Rapid Determination of Crude Protein Content in Alfalfa Based on Fourier Transform Infrared Spectroscopy

Haijun Du [1], Yaru Zhang [2], Yanhua Ma [1,*], Wei Jiao [3], Ting Lei [1] and He Su [1]

[1] College of Mechanical and Electrical Engineering, Inner Mongolia Agricultural University, No. 36 Zhaowuda Road, Hohhot 010018, China; du15563886115@163.com (H.D.); leiting0401@163.com (T.L.); suhe0826@126.com (H.S.)
[2] College of Horticulture and Plant Protection, Inner Mongolia Agricultural University, No. 36 Zhaowuda Road, Hohhot 010018, China; zhangyr0729@163.com
[3] The China Academy of Grassland Research, No. 120 Wulanchabu East Street, Saihan District, Hohhot 010018, China; jiaowei@caas.cn
* Correspondence: yanhuama@126.com

Abstract: The crude protein (CP) content is an important determining factor for the quality of alfalfa, and its accurate and rapid evaluation is a challenge for the industry. A model was developed by combining Fourier transform infrared spectroscopy (FTIS) and chemometric analysis. Fourier spectra were collected in the range of 4000~400 cm^{-1}. Adaptive iteratively reweighted penalized least squares (airPLS) and Savitzky–Golay (SG) were used for preprocessing the spectral data; competitive adaptive reweighted sampling (CARS) and the characteristic peaks of CP functional groups and moieties were used for feature selection; partial least squares regression (PLSR) and random forest regression (RFR) were used for quantitative prediction modelling. By comparing the combined prediction results of CP content, the predictive performance of airPLST-cars-PLSR-CV was the best, with an R_P^2 of 0.99 and an RMSEP of 0.053, which is suitable for establishing a small-sample prediction model. The research results show that the combination of the PLSR model can achieve an accurate prediction of the crude protein content of alfalfa forage, which can provide a reliable and effective new detection method for the crude protein content of alfalfa forage.

Keywords: alfalfa; crude protein; FTIS; PLSR-CV

check for updates

Citation: Du, H.; Zhang, Y.; Ma, Y.; Jiao, W.; Lei, T.; Su, H. Rapid Determination of Crude Protein Content in Alfalfa Based on Fourier Transform Infrared Spectroscopy. *Foods* **2024**, *13*, 2187. https://doi.org/10.3390/foods13142187

Academic Editor: Mourad Kharbach

Received: 14 June 2024
Revised: 8 July 2024
Accepted: 10 July 2024
Published: 11 July 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Alfalfa is regarded as the "king of forages" due to its high protein content, and the constantly evolving livestock industry has led to an increasing demand for high-quality alfalfa grass for animal feed [1]. In particular, Inner Mongolia is a major livestock farming base in China, which requires large quantities of high-quality forage to address the seasonal and regional shortages of feed resources. This has presented ongoing challenges for the quality assessment of forage grasses [2,3]. The feeding value of alfalfa is determined by the ratio of crude protein (CP) content to fiber content [4]. As a high-quality forage, the abundant protein in alfalfa can play an important role in improving the nutritional status, product quality, and yield of cattle and sheep [5]. However, the variety of alfalfa, the planting region, the harvesting time, the drying method, etc., all affect the CP content in alfalfa [6–8], which results in the inconsistent quality of alfalfa circulating in the market. Feeding livestock with alfalfa products low in CP content will prolong their growth cycle. Therefore, accurately evaluating the CP content of alfalfa is one way to assess the quality of forage products. The currently common chemical method for determining the nutritional value of forage is complex in operation, time-consuming, with certain risks to the operators, and requires the consumption of large amounts of reagents, leading to environmental pollution [9]. Therefore, finding a rapid, accurate, and efficient method to detect the

nutritional value of alfalfa is of great significance for promoting the rapid development of the alfalfa industry [10].

The Fourier transform infrared spectroscopy (FTIS) technology has been widely used in the quality detection of agricultural products [11]. This is a fast, efficient, and non-destructive detection method [12,13]. He et al. utilized the Fourier transform near-infrared spectroscopy (FT-NIRS) technology and the TQAnalyst v9 software to accurately determine the CP content in alfalfa hay [9]. Díaz et al. proposed the use of near-infrared spectroscopy (NIR) to non-destructively detect the protein, starch, and other contents in rice, thereby classifying the rice flavor [14]. Lu et al. utilized FT-NIRS and partial least squares regression (PLSR) combined with cross-validation to establish a quantitative model, which achieved excellent results in predicting the soluble solids and titratable acidity of pears [15]. Sahachairungrueng et al. used PLSR to fit the NIR information with chemical properties and digestibility, generating a set of trained NIR models that can be used for rapid biomass measurement [16].

The application of machine learning algorithms to the processing and prediction of infrared spectral data can improve data understanding and quantification, and achieve relatively high accuracy [17]. Mishra et al. used PLSR analysis and infrared spectroscopy to quantitatively detect the protein content in rice, with a detection error of 0.349% [18]. Baath et al.'s work using NIRS technology employed support vector machines (SVMs) and least-squares SVMs to predict the crude protein content of soybean and velvet bean feed samples. The reported R^2 values ranged from 0.92 to 0.98 [19]. Wang et al. used competitive adaptive reweighted sampling (CARS) to screen spectral features and used PLSR to establish prediction models for the non-destructive detection of protein and fat content in Chinese torreya, with coefficients of determination (R^2) of 0.977 and 0.984, respectively [20]. Magali used a random forest regression (RFR) model to predict and estimate the nitrogen content of feed corn, demonstrating good predictive performance [21]. Tian et al. used PLSR to establish quantitative analysis models for the nutritional components of fresh and dried oat samples, enabling the rapid detection of the nutritional quality of oat forage silage [22]. However, there are few reports on the use of Fourier transform infrared spectroscopy and machine learning to construct quantitative analysis models for the CP content of alfalfa.

The present study aims to identify alfalfa from different regions and varieties in Inner Mongolia within the spectral range of 400~4000 cm^{-1} using FTIS. It also combines chemometric methods to conduct a physicochemical analysis of the CP content in alfalfa, exploring the relationship between content variation and spectral information. This aims to provide technical guidance for the accurate, rapid, and non-destructive detection of the CP content in alfalfa.

2. Materials and Methods

2.1. Materials

The alfalfa grass came from the Inner Mongolia Agricultural University experimental base (located in Saihan District, Hohhot, Inner Mongolia, China) and the Hohhot alfalfa planting base (located in Hulin'ger County, Hohhot, Inner Mongolia, China). After artificial harvesting, the fresh alfalfa had a moisture content of approximately 70% to 80%; it was placed in a well-ventilated and shaded factory (temperature around 26 °C, relative humidity 12%), and turned twice a day. Using a portable moisture content meter (High-precision MS-H forage moisture meter, measurement range 0~84%, accuracy range 1%, Qingdao Tuoke Instrument Co., Ltd., Qingdao, China), the moisture content was determined, and when the moisture content of the alfalfa was less than 10%, the alfalfa grass from different areas, planting conditions, and varieties was crushed into granules to obtain the alfalfa grass samples. The prepared 90 grass samples were placed in a drying oven at 60 °C until they reached a constant weight.

2.2. Chemical Analysis Methods

The CP content in the alfalfa was determined using the Kjeldahl method according to the Chinese national standard GB 5009.5-2016 [23]. In total, 1 g of the thoroughly mixed solid sample, accurate to 0.001 g, was weighed and transferred to the digestion flask. Then, 0.4 g of copper sulfate, 6 g of potassium sulfate, and 20 mL of sulfuric acid were added and digested in the digestion furnace. The digestion temperature was 420 °C, and the duration was 1 h. After cooling, 50 mL of purified water was added, and then an automated Kjeldahl nitrogen analyzer (K9860 Fully Automatic Kjeldahl Nitrogen Analyzer—Hunan Honglin Scientific Instrument Co., Ltd., Changsha, China) was used to conduct the titration process and output the crude protein content result. The crude protein content range of alfalfa grass was 9.00% to 22.04%.

2.3. The Spectral Curve Measurement of the Sample

The analytical instrument used was the FTIR-1500 Fourier Transform Infrared Spectrometer (Zhengzhou oda Instrument Co., Ltd., Zhengzhou, China), which can collect spectra in the range of 4000~400 cm^{-1} with a resolution better than 1 cm^{-1} and 16 scans. Prior to sample scanning, the instrument was preheated for 0.5 h. Approximately 0.02 g of dry powder and 2.5 g of potassium bromide (KBr) dried at 80 °C for 10 h were ground for more than 10 min using a quartz mortar and pestle. Then, 0.18 g of prepared sample was pressed into a test sample using a press and placed in the spectrometer for scanning to obtain the spectrum. The accompanying Fourier transform infrared spectroscopy software was used to collect the raw spectral data. Before each sample spectrum collection, (air) background collection was performed. Each sample was randomly measured 10 times, and the spectral curve of the sample was constructed using the average method. The FTIS of the sample is shown in Figure 1. The spectral data have 10,329 characteristic points.

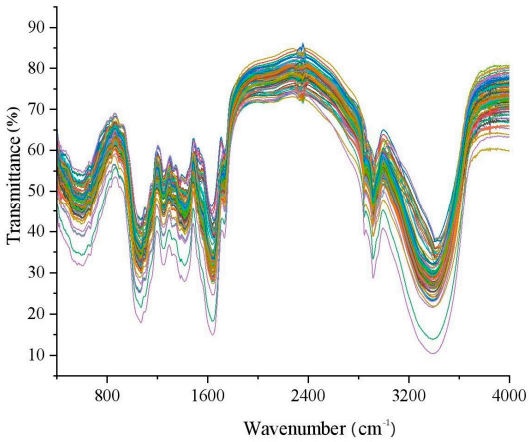


Figure 1. FTIS of alfalfa samples.

2.4. Spectrum Preprocessing

The FTIS data of alfalfa were preprocessed to eliminate noise and outliers in the spectral data [24]. The adaptive iteratively reweighted penalized least squares (airPLS) method was used to correct the entire spectrum [25]. AirPLS is an improved method based on weighted least squares to weigh the original model. AirPLS estimates the baseline of the original signal, controls the smoothness of the baseline, and eliminates the baseline drift in the spectral data, thereby improving the accuracy of baseline estimation [26,27]. In this study, the adaptive iteratively reweighted penalized least squares (airPLS) algorithm was used to preprocess the Fourier transform spectrum of alfalfa, with lambda set to 10×10^9, the order of the difference of penalties set to 2, the weight exception proportion at both the start and end set to 0.05, the asymmetry parameter for the start and end set to 0.5, and the maximum iteration times set to 20. The Savitzky–Golay (SG) algorithm was used to smooth and differentiate the signal data to improve the spectral quality and enable

accurate quantitative analysis [28,29]. After applying the smoothing algorithm, the min-max normalization method was further used to process the data, as these preprocessing steps are crucial for minimizing the impact of light scattering on the development and validation of classification models [30].

2.5. Spectral Feature Selection

Spectral data often require feature selection to overcome the challenge of a large number of variables in the data, and to remove wavelength features that are not related to the protein. Feature selection aims to identify the most relevant and informative features, which can improve model accuracy, reduce computational cost, and enhance interpretability [31]. Various feature selection techniques can be utilized, such as peak identification, cluster analysis, statistical methods, and a more advanced method, CARS [32]. The CARS algorithm can adaptively select the optimal subset from a large number of variables, significantly reducing model complexity while improving model prediction accuracy [33].

These techniques assess the correlation between each wavelength and the target variable, considering factors such as the signal-to-noise ratio, spectral characteristics, and relevance to potential physical or chemical processes. In the current research, CARS is used to select a subset of important features from the spectra, which can improve model accuracy by removing redundant or noisy information and prevent overfitting, especially in small data sets [34].

The PLS model [35] and the random forest model were constructed using a combination of preprocessing algorithms and different feature wavelength selection algorithms to estimate the crude protein content of alfalfa. The model prediction results were evaluated based on the coefficient of determination (R_c^2) and root mean square error (RMSEC) of the calibration set, as well as the coefficient of determination (R_p^2) and root mean square error (RMSEP) of the prediction set. An R_c^2 and R_p^2 closer to 1 indicate better model stability, goodness of fit, and accuracy, while RMSEC and RMSEP closer to 0 suggest smaller model errors and better predictive capability [36].

$$R_C^2 = 1 - \frac{\sum(C_i - \hat{C}_i)^2}{\sum(C_i - C_n)^2}, \quad (1)$$

$$R_P^2 = 1 - \frac{\sum(C_i - \hat{C}_i)^2}{\sum(C_i - C_m)^2}, \quad (2)$$

$$\text{RMSEC} = \sqrt{\frac{\sum(C_i - \hat{C}_i)^2}{\text{n}}}, \quad (3)$$

$$\text{RMSEP} = \sqrt{\frac{\sum(C_i - \hat{C}_i)^2}{\text{m}}} \quad (4)$$

where C_i is the i-th measured value of the reference method; C_i is the i-th predicted value of the model; C_n is the mean value of the measured values of n samples; C_m is the mean of the predicted values of m samples; m is the number of samples in the prediction sets; and n is the number of samples in the calibration sets.

In this work, all data analyses, including spectral pretreatment, wavelength selection, and PLS modeling, were performed in the Python 3.11 software.

3. Results and Discussion

3.1. Spectral Feature Analysis

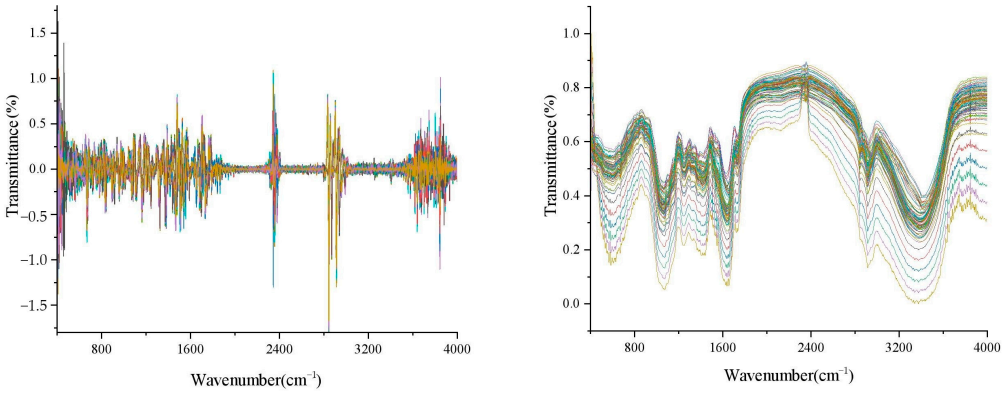
The Fourier transform infrared transmittance spectra of alfalfa samples in the range of 400–4000 cm^{-1} are shown in Figure 1. The figure depicts the average raw transmittance spectra of 90 alfalfa samples. The spectra exhibit a consistent pattern with essentially similar curve shapes. Relatively strong absorption peaks are observed around 593, 600,

665, 1053, 1102, 1246, 1381, 1417, 1630, 2915, and 3394 cm^{-1}. The absorption peak around 3394 cm^{-1} is associated with the asymmetric stretching frequency of the NH2 group, while the absorption peak around 1630 cm^{-1} is related to the bending vibration frequency of the NH2 group. The absorption peak at 1246 cm^{-1} is associated with the stretching vibration of the C-N group. Additionally, the absorption peak around 1246 cm^{-1} is related to the stretching vibration of the COOH group, and the absorption peak at 1417 cm^{-1} is associated with the in-plane bending vibration of the COOH group. The peaks at 600 and 650 cm^{-1} are related to the vibration frequency of the C-OH group.

Based on the characteristic absorption peak spectral frequencies of functional groups and moieties (FG&M) present in crude protein, absorption peak value data at corresponding frequencies can be extracted as characteristic values for crude protein content. This method can serve as a means of feature extraction for analyzing and evaluating crude protein content.

3.2. FTIS Preprocessing

The application of the airPLS method to the original spectral data and the correction preprocessing of the entire spectrum resulted in the transmittance spectrum shown in Figure 2a. The application of the SG algorithm and normalization to the correction preprocessing of the entire spectrum resulted in the transmittance spectrum shown in Figure 2b.



(**a**) (**b**)

Figure 2. The preprocessed results: (**a**) AirPLS; (**b**) SG.

The 90-sample spectra are divided into training, validation, and prediction sets according to a 7:1.5:1.5 ratio, resulting in 63 training samples, 13 validation samples, and 14 prediction samples. Based on Table 1, the use of the PLSR model and feature screening preprocessing methods resulted in a very small coefficient of determination in the final results, and the root mean square errors were quite large, indicating that the feature vector contained too many features, leading to significant interference and noise. Among the preprocessing methods, the SG algorithm for the entire spectrum performed the worst, and therefore this method was excluded from the preprocessing of the spectral data.

Table 1. Comparison of performance of PLSR models with different pretreatment methods.

Pretreatment	Calibration Set		Prediction Set	
	RMSEC	R_c^2	RMSEP	R_P^2
without	2.23	0.35	2.55	0.37
airPLS	2.22	0.35	2.14	0.55
SG	2.23	0.35	2.55	0.36

3.3. Effective Wavelength Selection

The selection of characteristic wavelengths in spectral data is based on the comparative results of pretreatment research and the selection of the most effective pretreatment procedure. However, even after preprocessing the actual spectrum, the prediction performance is still not good, due to the large amount of data. Therefore, it is necessary to perform dimensionality reduction on the full-band spectrum after airPLS.

The wavelength results obtained using the CARS method are shown in Figure 3. Through two methods of screening feature points, it was found that there are overlapping feature points around the absorption peaks at 2915, 1633, 1512, 1420, and 1242 cm^{-1} wavenumber. The absorption band in the 3000~2800 cm^{-1} region is most likely attributed to C-H and NH3, which can be referred to as free amino acids and phenols. The region of 1700~1550 cm^{-1} is the characteristic peak for proteins [37]. The 1750–940 cm^{-1} wavenumber range corresponds to the C-OH group and C-O stretching vibrations in the phenol structure. The CARS feature selection method, when dealing with large-scale data features, converges through each iteration and gradually eliminates irrelevant or weakly correlated features. However, more is needed to guarantee the complete elimination of weakly correlated features. In fact, certain spectral positions exhibit significant noise that cannot be sufficiently reduced through data preprocessing methods, resulting in the retention of some weakly correlated feature points.

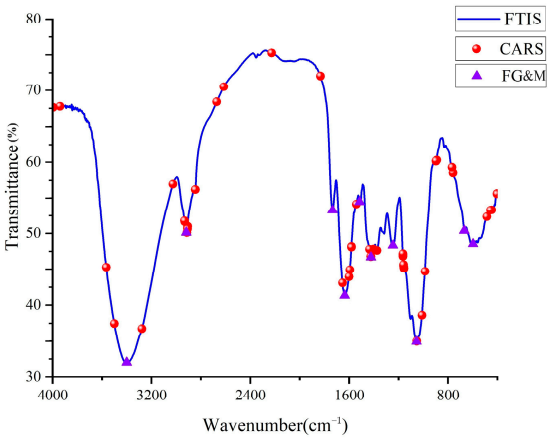


Figure 3. Using CARS and FG&M to select features related to the spectral characteristics of crude protein.

The results of the CARS method for extracting feature values are shown in Figure 4. As the number of iterations increases, the number of selected wavelengths gradually decreases and becomes stable when the number of iterations approaches 20. The RMSECV value gradually decreases in the early stage, and then reaches the minimum value when the number of iterations is close to 17, after which the RMSECV value begins to rise. This indicates that at the 17th iteration, the CARS method has achieved the best balance between the number of selected wavelengths and the predictive performance of the model. This suggests that the CARS method gradually screens out the most representative wavelengths during the iteration process, and ultimately retains fewer but more informative wavelengths.

The PLSR models developed using feature wavelengths selected in various ways are shown in Table 2. The PLSR model developed using feature wavelengths selected by CARS exhibited superior predictive performance for protein, with a prediction set R_P^2 of 0.89 and an RMSEP of 0.75. The PLSR model developed using feature wavelengths selected based on protein absorption peaks showed a moderate predictive ability for protein, with a relatively low coefficient of determination. The RFR models developed using different feature selection methods all exhibited excellent predictive performance for CP. Among them, the RFR model developed using feature wavelengths selected by CARS showed the

best predictive performance for CP, with a prediction set R_P^2 of 0.96 and an RMSEP of 0.41, indicating its high capability in predicting CP content.

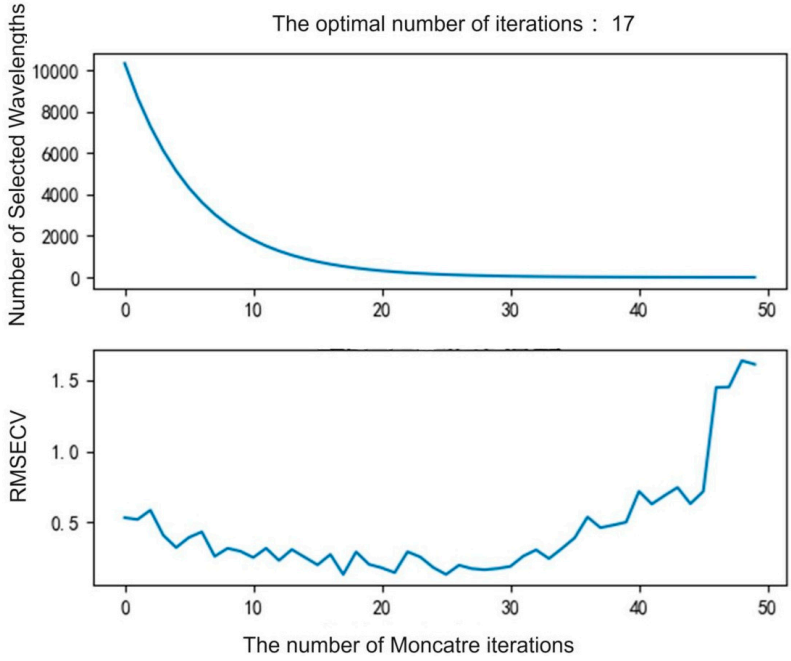


Figure 4. The results of the CARS method for extracting feature values.

Table 2. Comparison of performance of PLSR and RFR models with different feature selection methods.

Model	Pretreatment	Number	Calibration Set		Prediction Set	
			RMSEC	R_c^2	RMSEP	R_P^2
PLSR	airPLS-CARS	55	0.74	0.95	0.75	0.89
	airPLS-FG&M	20	1.32	0.85	1.51	0.54
RFR	airPLS-CARS	55	0.58	0.96	0.41	0.96
	airPLS-FG&M	20	0.95	0.92	0.79	0.94

The feature extraction using airPLS-CARS and the use of RFR for predicting the CP content of alfalfa demonstrated excellent predictive capabilities. The R_c^2 = 0.96 and R_p^2 = 0.96 values were higher than those of the PLSR model, and the RMSE was comparable, with RMSE values being smaller than those of the PLSR model. This indicates that the RFR model has more stable predictive performance.

Alfalfa CP content was predicted using FG&M and RFR, with an R_P^2 of 0.94 and an RMSEP of 0.79, which is slightly lower in accuracy than the CARS feature extraction method, but the number of feature values is 4/11 of the CARS extracted feature values. It is inevitable that water molecules cannot be completely removed from the samples and the manufacturing environment, and potassium bromide itself will absorb water molecules from the air, so the measured spectra show an absorption peak around 3400 cm^{-1} due to water, which interferes with the peak feature values of NH2 in this region [38]. The CARS method avoids the 3400 cm^{-1} feature peak and instead selects feature peaks at 3368, 3445, and 3477 cm^{-1} around it, which can avoid interference from water. The prediction results of airPLS-CARS-PLSR, airPLS-FG&M-PLSR, airPLS-CARS-RFR, and airPLS-FG&M-RFR are shown in Figure 5.

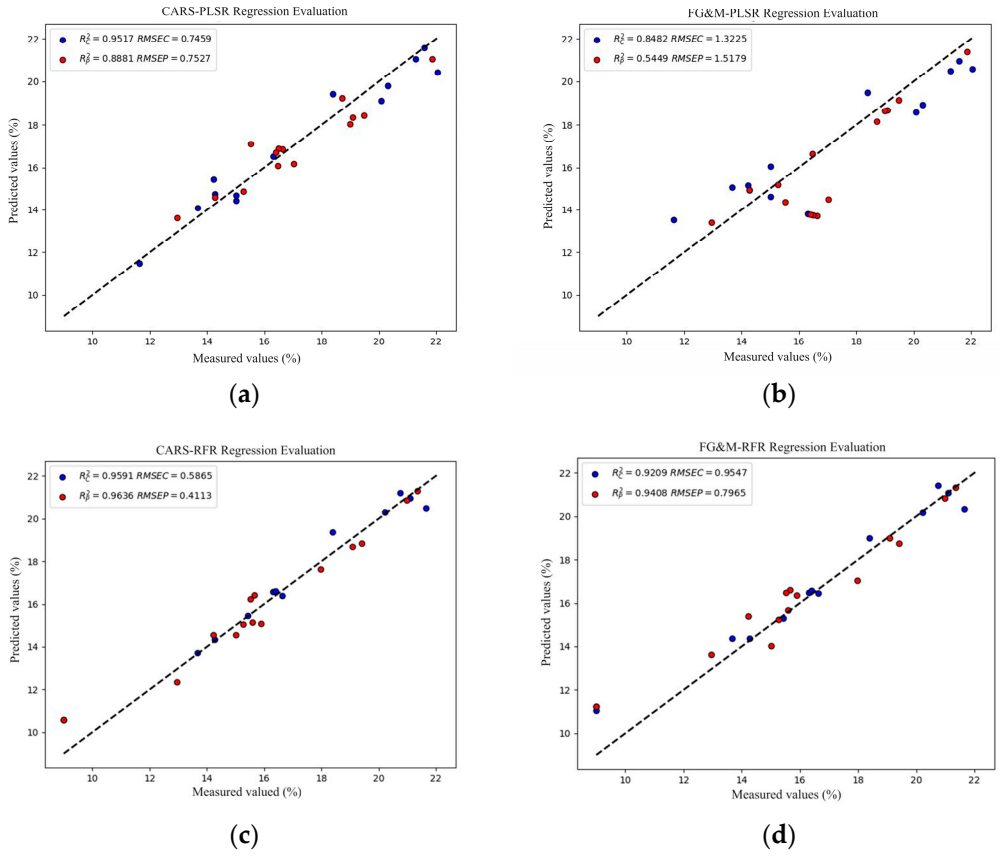


Figure 5. Predicted CP content: (**a**) airPLS-CARS-PLSR; (**b**) airPLS-FG&M-PLSR; (**c**) airPLS-CARS-RFR; and (**d**) airPLS-FG&M-RFR.

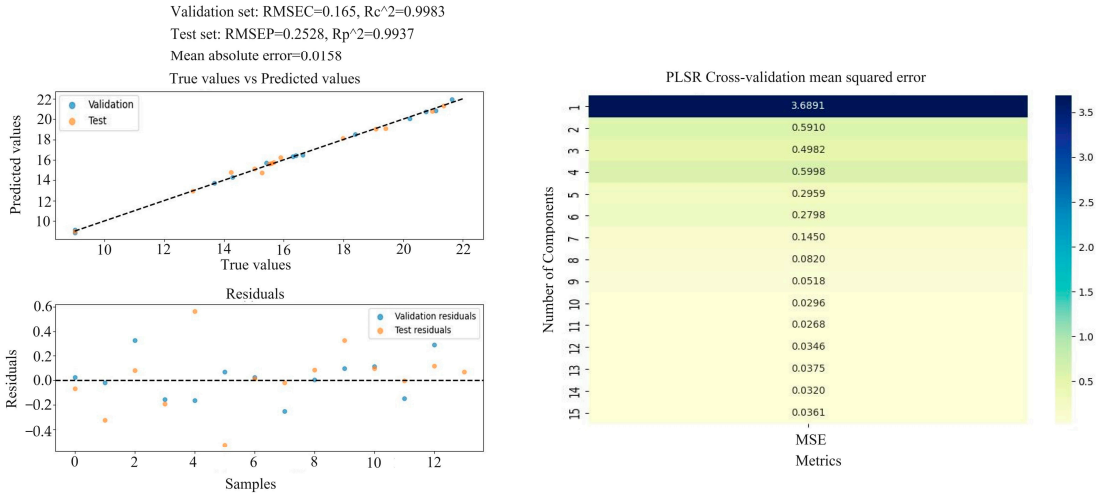
3.4. Discussion of Final Models

Through the feature vectors obtained using different feature screening methods, the protein is predicted by the PLSR and RFR models, both of which have good predictive performance. Among them, the RFR prediction performance is better and more stable than that of the PLSR model. In order to obtain a relatively stable and better predictive performance, the cross-validation (CV) of K-fold Cross-Validation is introduced to find the optimal number of principal components, thereby optimizing the predictive performance of the PLSR prediction model [39]; Figures 6–8 show the optimized PLSR and RFR optimization prediction results.

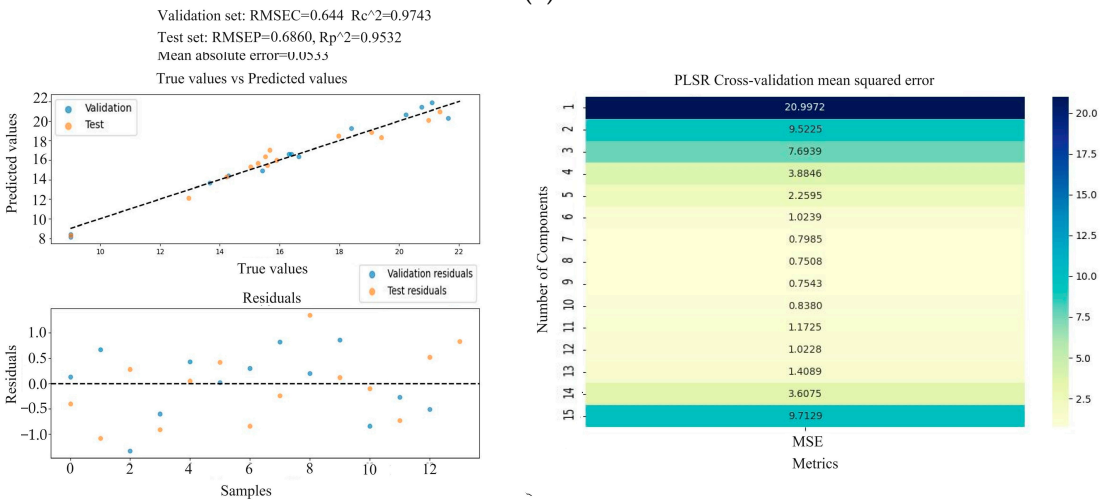
Among them, the PLSR model developed using the CARS-selected features has a predictive performance of R_P^2 = 0.99 and RMSEP = 0.25 (as shown in Figure 6a), and the average of the absolute error between the prediction and the target is 0.0158. The heatmap of the optimal number of principal components shows that when the number of principal components is 11, the total RMSE of the model is minimized. The PLSR model developed using FG&M-selected features has a predictive performance of R_P^2 = 0.95 and RMSEP = 0.69 (as shown in Figure 6b), and the average of the absolute error between the prediction and the target is 0.0533. The heatmap of the optimal number of principal components shows that when the number of principal components is eight, the total RMSE of the model is minimized.

The introduced CV and parameter optimization of the RFR model, as well as the automated hyperparameter tuning, can efficiently find the optimal parameter combination and thereby improve the overall performance of the model. Figure 6 shows the optimized RFR prediction results and optimization results. The RFR model developed using the FG&M-selected features has a prediction performance of R_p^2 = 0.96 and RMSEP = 0.19 (as shown in Figure 7a), with an average absolute error of 0.33 between the predictions and the targets. The random forest feature importance analysis reveals the key feature

peaks in the model, as shown in Figure 7b, with the critical peaks at 1426.07, 1054.13, 1051.69, and 1420.84 cm^{-1}. The optimized model parameters are as follows: max_depth: 10; max_features: 'sqrt'; min_samples_leaf: 1; min_samples_split: 2; n_estimators: 15.



(**a**)



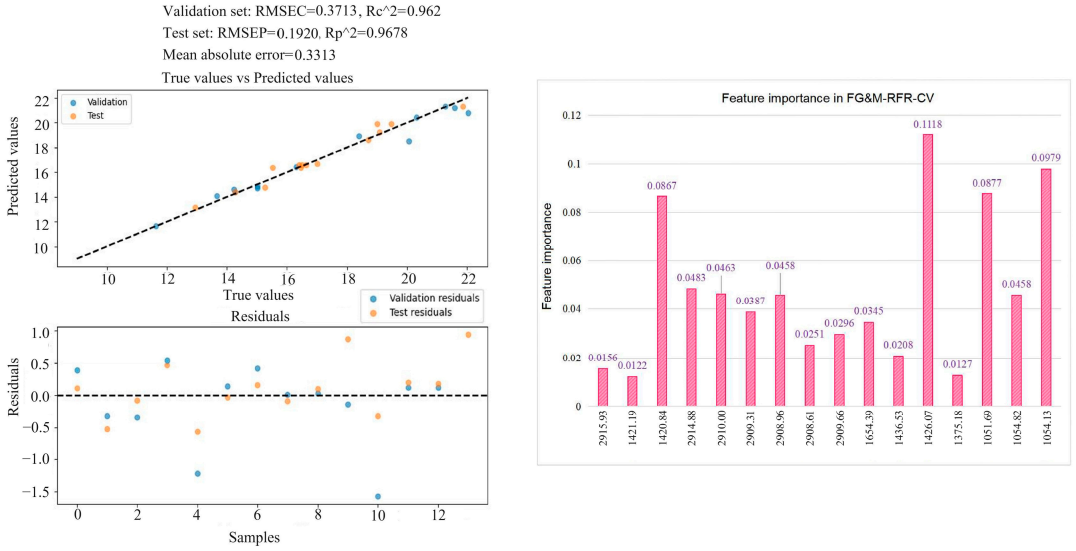
(**b**)

Figure 6. Optimization prediction results: (**a**) airPLS-CARS-PLSR-CV; (**b**) airPLS-FG&M-PLSR-CV.

The developed RFR model based on the features selected by CARS for protein prediction performance gives an R-squared value of 0.93 and an RMSEP of 0.54 (as shown in Figure 8a), with an average absolute error of 0.38 between the predicted and target values. Through the random forest feature importance analysis (as shown in Figure 8b), the key features in the model are identified as 897.61, 986.85, 898.66, 1157.66, 1165.53, and 2909.66 cm^{-1}. The optimized model parameters are as follows: max_depth: 10; max_features: 'sqrt'; min_samples_leaf: 1; min_samples_split: 2; n_estimators: 15. The overall performance of the CARS-RFR-CV model is lower than that of the FG&M-RFR-CV model. The RFR model is suitable for predicting a small number of features, as an excessive number of feature variables can hinder its computational process and interfere with the calculation results.

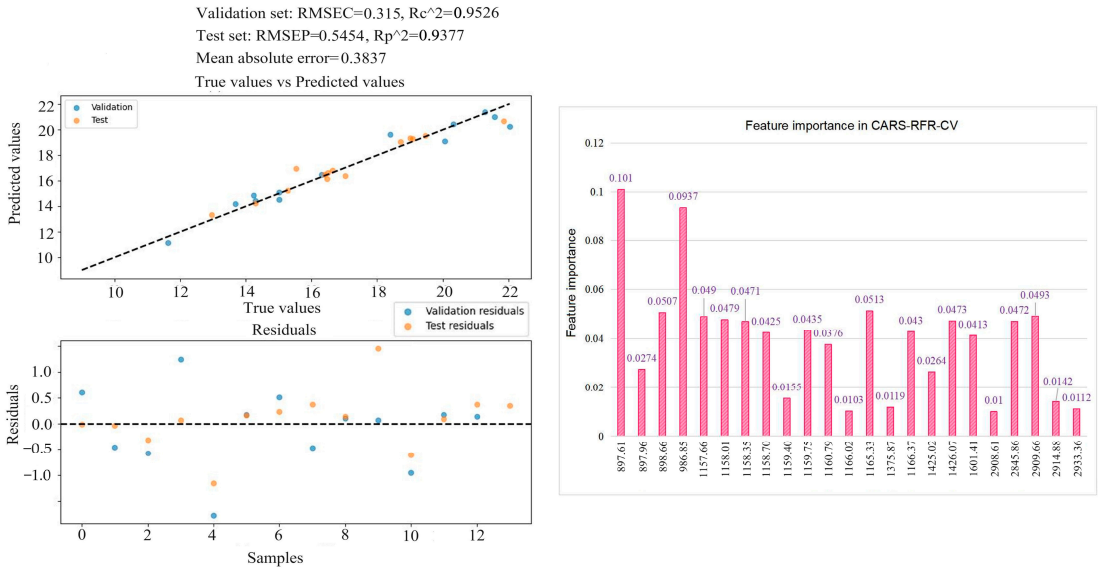
The final decision coefficient for the predictive performance of the crude protein content prediction model is as follows: airPLS-CARS-PLSR-CV > airPLS-FG&M-RFR-CV > airPLS-FG&M-PLSR-CV > airPLS-CARS-RFR-CV. The root mean square error of the prediction is as follows: airPLS-FG&M-RFR-CV < airPLS-CARS-PLSR-CV < airPLS-CARS-RFR-CV < airPLS-FG&M-PLSR-CV. The airPLS-CARS-PLSR-CV model has the best

predictive performance and relatively stable predictive results for predicting crude protein content, with the prediction accuracy as the primary evaluation metric and a smaller root mean square error. CARS can select wavelength variables with high weights and delete wavelength variables with low weights to improve model performance [40]. The feature wavelength variables screened out based on the spectral characteristics of FG&M of proteins have certain predictive ability, but they may ignore some key hidden feature wavelengths, causing a decline in predictive performance. RFR is a method that randomly selects the feature set size, evaluates the size of each decision tree, and then estimates the best size based on the decision tree prediction. It is very fast because it only considers a few candidates. The RFR model does not require too much feature processing and has strong robustness to outliers and interference values [41,42].



(**a**) (**b**)

Figure 7. The optimized results of the RFR model based on FG&M-selected features: (**a**) predictive analysis results; (**b**) feature importance in FG&-MRFR-CV.



(**a**) (**b**)

Figure 8. The optimized results of the RFR model based on CARS-selected features: (**a**) predictive analysis results; (**b**) feature importance in CARS-RFR-CV.

4. Conclusions

This study established a promising airPLS-CARS-PLS-CV model and compared its predictive performance and model robustness to the PLSR-based prediction model. Among the four models, the airPLS-CARS-PLS-CV model reduced the RMSEP for crude protein by 88.89%, 88.87%, and 58.33% compared to the full PLSR model, the airPLS model, and the airPLS-CARS-PLS model, respectively. These results indicate that the wavelength selection procedure can effectively improve the predictive performance of the model. Feature selection using the FG&M feature wavelengths can reduce the number of features, but the predictive performance is slightly lower than that of the airPLS-CARS-PLS-CV model. In the case of a small number of feature values, the model based on airPLS-FG&M-RFR-CV has better predictive capability and better stability. Overall, the results confirm the feasibility of using Fourier transform infrared spectroscopy for the rapid determination of crude protein in alfalfa.

Author Contributions: Conceptualization, H.D. and Y.M.; methodology, H.D.; software, H.D. and T.L.; validation, H.D., Y.Z. and Y.M.; formal analysis, H.D. and W.J.; investigation, H.D. and Y.Z.; resources, Y.M.; data curation, H.D.; writing—original draft preparation, H.D.; writing—review and editing, H.D., Y.Z. and Y.M.; supervision, Y.M. and H.S.; project administration, Y.M. and W.J.; funding acquisition, Y.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Interdisciplinary Research Fund of Inner Mongolia Agricultural University (No. BR231517), the National Natural Science Foundation of China (No. 32060771), the Fundamental Research Funds of Inner Mongolia Agricultural University (No. BR221314, BR221032), and the Science and Technology Planning Project of Inner Mongolia Autonomous Region (No. 2023YFDZ0037).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: The authors acknowledge the chemical testing technical support provided by Inner Mongolia Bocetancy Quality Inspection Technology Co., Ltd. (Hohhot, China).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Du, H.J.; Lei, T.; Ma, Y.; Li, Y.; Yang, X.T. Effect of process parameters on quality of alfalfa block. *BioResources* **2024**, *19*, 3234–3248. [CrossRef]
2. Tong, Y.; Dong, J.W.; Huang, L.; Li, Y.Z.; Yan, H.M.; Zhai, J.; Wang, J.; Jin, Z.; Zhang, G. A large forage gap in forage availability in traditional pastoral regions in China. *Fundam. Res.* **2023**, *3*, 188–200. [CrossRef]
3. İnal, F.; Coşkun, B.; Alataş, M.S.; Kahraman, O.; Özbilgin, A. Quality classification of alfalfa hays according to protein and fiber contents. *Eurasian J. Vet. Sci.* **2020**, *36*, 193–198. [CrossRef]
4. Liao, Z.; Su, K.; Jiang, X.; Wang, J.; You, Y.; Wang, L.; Chang, S.; Wei, C.; Zhang, Y.; Li, C. Spatiotemporal variation and coupling of grazing intensity and ecosystem based on four quadrant model on the Inner Mongolia. *Ecol. Indic.* **2023**, *152*, 110379. [CrossRef]
5. Ye, S.; Zhang, J.X.; Hu, W.; Jonathan, D.; Zhang, G.Z.; Mukamel, S.; Jiang, J. A machine learning protocol for predicting protein infrared spectra. *J. Am. Chem. Soc.* **2020**, *142*, 19071–19077. [CrossRef] [PubMed]
6. Feng, Y.P.; Shi, Y.; Zhao, M.Y.; Shen, H.H.; Xu, L.C.; Luo, Y.K.; Liu, Y.Z.; Xing, A.J.; Kang, J.; Jing, H.C.; et al. Yield and quality properties of alfalfa (*Medicago sativa* L.) and their influencing factors in China. *Eur. J. Agron.* **2022**, *141*, 126637. [CrossRef]
7. Yuan, N.; Sun, L.; Du, S.; Ge, G.T.; Wang, Z.J.; Li, Y.Y.; Bao, J.; Zhao, M.; Si, Q.; Hao, J.F.; et al. Effects of Harvesting Period and Storage Duration on Volatile Organic Compounds and Nutritive Qualities of Alfalfa. *Agriculture* **2022**, *12*, 1115. [CrossRef]
8. Díaza, F.J.; Grattanb, S.R.; Reyesc, J.A.; Roza-Delgadod, B.; Benese, S.E.; Jiméneza, C.; Dortaa, M.; Tejedor, M. Using saline soil and marginal quality water to produce alfalfa in arid climates. *Agric. Water Manag.* **2018**, *199*, 11–21. [CrossRef]
9. He, Y.Q.; Ren, Y.; Liu, J.H.; Liu, L.; Yang, H.; Li, Z.P.; Zhan, Q.W. Study on rapid determination of qualities of alfalfa hay based on NIRS. *Spectrosc. Spectr. Anal.* **2023**, *43*, 3753–3757.
10. Jia, W.; Georgouli, K.; Rincon, J.M.; Koidis, A. Challenges in the Use of AI-Driven Non-Destructive Spectroscopic Tools for Rapid Food Analysis. *Foods* **2024**, *13*, 846. [CrossRef]
11. Chu, X.L.; Chen, P.; Li, J.Y.; Liu, D.; Xu, Y.P. Progresses and Perspectives of Near Infrared Spectroscopy Analytical Technology. *J. Instrum. Anal.* **2020**, *39*, 1181–1188.

12. Amanah, H.Z.; Tunny, S.S.; Masithoh, R.E.; Choung, M.; Kim, K.; Kim, M.S.; Baek, I.; Lee, W.; Cho, B. Nondestructive Prediction of Isoflavones and Oligosaccharides in Intact Soybean Seed Using Fourier Transform Near-Infrared (FT-NIR) and Fourier Transform Infrared (FT-IR) Spectroscopic Techniques. *Foods* **2022**, *11*, 232. [CrossRef] [PubMed]
13. Liu, Q.; Zhang, W.; Zhang, B.; Du, C.W.; Wei, N.N.; Liang, D.; Sun, K.; Tu, K.; Peng, J.; Pan, L.Q. Determination of total protein and wet gluten in wheat flour by Fourier transform infrared photoacoustic spectroscopy with multivariate analysis. *J. Food Compos. Anal.* **2022**, *106*, 104349. [CrossRef]
14. Díaz, E.O.; Iino, H.; Koyama, K.; Kawamura, S.; Koseki, S.; Lyu, S. Non-destructive quality classification of rice taste properties based on near-infrared spectroscopy and machine learning algorithms. *Food Chem.* **2023**, *429*, 136907. [CrossRef] [PubMed]
15. Lu, Z.; Lu, R.; Chen, Y.; Fu, K.; Song, J.; Xie, L.; Zhai, R.; Wang, Z.; Yang, C.; Xu, L. Nondestructive Testing of Pear Based on Fourier Near-Infrared Spectroscopy. *Foods* **2022**, *11*, 1076. [CrossRef] [PubMed]
16. Sahachairungrueng, W.; Meechan, C.; Veerachat, N.; Thompson, A.K.; Teerachaichayut, S. Assessing the Levels of Robusta and Arabica in Roasted Ground Coffee Using NIR Hyperspectral Imaging and FTIR Spectroscopy. *Foods* **2022**, *11*, 3122. [CrossRef] [PubMed]
17. Perezmarin, D.; Garridovaro, A.; Guerrero, J. Non-linear regression methods in NIRS quantitative analysis. *Talanta* **2007**, *72*, 28–42. [CrossRef] [PubMed]
18. Mishra, P.; Angileri, M.; Woltering, E. Identifying the best rice physical form for non-destructive prediction of protein content utilising near-infrared spectroscopy to support digital phenotyping. *Infrared Phys. Technol.* **2021**, *116*, 103757. [CrossRef]
19. Baath, G.S.; Baath, H.K.; Gowda, P.H.; Thomas, J.P.; Northup, B.K.; Rao, S.C.; Singh, H. Predicting Forage Quality of Warm-Season Legumes by Near Infrared Spectroscopy Coupled with Machine Learning Techniques. *Sensors* **2020**, *20*, 867. [CrossRef] [PubMed]
20. Wang, Y.; Dai, Y.J.; Wang, Y.Y.; Yang, J.L.; Xiang, D.H.; Yang, Y.Q.; Zeng, S.W. Research on Non-destructive Detection of Protein and Fat Content in Torreya Based on Near-Infrared Spectroscopy Technology. *Sci. Technol. Food Ind.* **2024**, *45*, 18. (In Chinese with English Abstract) [CrossRef]
21. López-Calderón, M.J.; Estrada-Ávalos, J.; Rodríguez-Moreno, V.M.; Mauricio-Ruvalcaba, J.E.; Martínez-Sifuentes, A.R.; Delgado-Ramírez, G. Estimation of Total Nitrogen Content in Forage Maize (*Zea mays* L.) Using Spectral Indices: Analysis by Random Forest. *Agriculture* **2020**, *10*, 451. [CrossRef]
22. Tian, L.M.; Chen, F.; Lin, Y.L.; Ni, K.K.; Yang, F.Y. Research on detecting nutritional quality of forage oat silage by infrared spectroscopy. *Acta Agrestia Sin.* **2023**, *31*, 3841–3850. [CrossRef]
23. *GB/T5009.5-2016*; Determination of Protein Content in Food. Standardization Administration of China: Beijing, China, 2016.
24. Rinnan, Å.; Berg, F.V.D.; Engelsen, S.B. Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends Anal. Chem.* **2009**, *28*, 1201–1222. [CrossRef]
25. Li, Y.; Pan, T.; Li, H.; Chen, S. Baseline correction using local smoothing optimization penalized least squares. In Proceedings of the 2022 IEEE International Symposium on Advanced Control of Industrial Processes (AdCONIP), Vancouver, BC, Canada, 7–9 August 2022. [CrossRef]
26. Zhang, F.; Tang, X.; Tong, A.; Wang, B.; Wang, J.; Lv, Y.; Tang, C.; Wang, J. Baseline correction for infrared spectra using adaptive smoothness parameter penalized least squares method. *Spectrosc. Lett.* **2020**, *53*, 222–233. [CrossRef]
27. LI, Y.P.; Fang, T.; Zhu, S.Q.; Huang, F.; Chen, Z.Q.; Wang, Y. Detection of olive oil adulteration with waste cooking oil via Raman spectroscopy combined with iPLS and SiPLS. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2017**, *189*, 37–43. [CrossRef] [PubMed]
28. Tsagkaris, A.S.; Bechynska, K.; Ntakoulas, D.D.; Pasias, I.N.; Weller, P.; Proestos, C.; Hajslova, J. Investigating the impact of spectral data pre-processing to assess honey botanical origin through Fourier transform infrared spectroscopy (FTIR). *J. Food Compos. Anal.* **2023**, *119*, 105276. [CrossRef]
29. Díaz, E.O.; Kawamura, S.; Ishizu, H.; Nagata, T.; Koseki, S. Non-destructive assessment of amylose content in rice using a quality inspection system at grain elevators. *Food Chem.* **2022**, *379*, 132144. [CrossRef] [PubMed]
30. Singh, D.; Singh, B. Feature wise normalization: An effective way of normalizing data. *Pattern Recognit.* **2022**, *122*, 108307. [CrossRef]
31. Yang, Y.T.; Jiang, M.; Zhan, R.L.; Xia, H.; Chao, Y.N.; Wang, Y.; Ren, Y.H.; Peng, L.X. Development and Optimization of a Rapid Detection Model for Characteristic Nutrient Content in Highland Barley. *Sci. Technol. Food Ind.* **2024**, 1–17. (In Chinese with English Abstract)
32. Jia, B.; Chen, Q.W.; Lan, J.M.; Zhang, K.L.; Xiang, C.Q.; Guo, T.; Chen, M.; Xiao, X. Determination of Polygoni Multiflori Polysaccharide Contents during Concoction by Near Infrared Spectroscopy. *J. Instrum. Anal.* **2024**, *43*, 481–488. [CrossRef]
33. Hu, W.F.; Tang, W.H.; Li, C.; Wu, J.J.; Ma, Q.F.; Luo, X.C.; Wang, C.; Tang, R.N. Estimating nitrogen concentration of rubber leaves based on a hybrid learning framework and Near-Infrared spectroscopy. *Spectrosc. Spectr. Anal.* **2023**, *43*, 2050–2058.
34. Zhu, Y.; Fan, S.; Zuo, M.; Zhang, B.; Zhu, Q.; Kong, J. Discrimination of New and Aged Seeds Based on On-Line Near-Infrared Spectroscopy Technology Combined with Machine Learning. *Foods* **2024**, *13*, 1570. [CrossRef]
35. Sun, J.; Yang, W.; Zhang, M.; Feng, M.; Xiao, L.; Ding, G. Estimation of water content in corn leaves using hyperspectral data based on fractional order Savitzky-Golay derivation coupled with wavelength selection. *Comput. Electron. Agric.* **2021**, *182*, 105989. [CrossRef]
36. She, X.T.; Huang, J.; Cao, X.Q.; Wu, M.J.; Yang, Y. Rapid Measurement of Total Saponins, Mannitol, and Naringenin in Dendrobium officinale by Near-Infrared Spectroscopy and Chemometrics. *Foods* **2024**, *13*, 1199. [CrossRef]

37. Hssaini, L.; Razouk, R.; Bouslihim, Y. Rapid Prediction of Fig Phenolic Acids and Flavonoids Using Mid-Infrared Spectroscopy Combined with Partial Least Square Regression. *Front. Plant Sci.* **2022**, *13*, 782159. [CrossRef] [PubMed]
38. Wang, L.M.; Chen, S.; Ding, J.; Qi, B.K.; Wang, Z.J.; Jiang, L.Z.; Sui, X.N.; Babich, O.O.; Li, Y. Structural Characteristics of Protein Emulsion Formed during Enzyme-Assisted Aqueous Extraction of Soybean Oil. *Food Sci.* **2018**, *39*, 9–15. (In Chinese with English Abstract) [CrossRef]
39. Zhang, X.; Liu, C. Model averaging prediction by K-fold cross-validation. *J. Econom.* **2023**, *235*, 280–301. [CrossRef]
40. Wu, X.; Zeng, S.; Fu, H.; Wu, B.; Zhou, H.; Dai, C. Determination of corn protein content using near-infrared spectroscopy combined with A-CARS-PLS. *Food Chem. X* **2023**, *18*, 100666. [CrossRef]
41. Han, S.; Kim, H. Optimal Feature Set Size in Random Forest Regression. *Appl. Sci.* **2021**, *11*, 3428. [CrossRef]
42. Broeiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.