Detection of Protein Content in Alfalfa Using Visible/ Near-Infrared Spectroscopy Technology

Jie Li,[a] Guifang Wu,[a,*] Fang Guo,[a,*] Lei Han,[a] Haowen Xiao,[b] Yang Cao,[b] Huihe Yang,[a] and Shubin Yan [a]

In this study, a quantitative model was developed using near-infrared spectroscopy to analyze protein content in dried purple alfalfa, employing preprocessing methods (SG, SNV, MSC, FD) and variable selection algorithms (CARS, IRIV) to optimize spectra. Models using ELM, PLSR, SVM, and LSTM were tested; the MSC-CARS-PLSR-SVM model achieved the highest accuracy, with a calibration determination coefficient (R^2) of 0.9982 and root mean square error (RMSE) of 0.1088, and a prediction R^2 of 0.9645 with RMSE of 0.5230, offering a precise and reliable method for protein content prediction.

DOI: 10.15376/biores.19.2.3808-3825

Keywords: Quantitative detection; Near-infrared spectroscopy; Machine learning; Protein content; Alfalfa hay

Contact information: a: College of Mechanical & Electrical Engineering, Inner Mongolia Agricultural University, Hohhot, 010018, P.R. China; b: Inner Mongolia Autonomous Region Agricultural and Pastoral Technology Extension Center, Hohhot, 010010, P.R. China;
**Corresponding authors: wgfsara@126.com and jennifer_guo@imau.edu.cn*

INTRODUCTION

Purple alfalfa, a perennial herbaceous plant belonging to the legume family, is widely considered to have originated in the Near East region, including Iran, Anatolia, the Turkmen Plateau, and the Transcaucasus（Bedaf *et al.* 2008）. It boasts high production potential and nutritional value, making it one of the most extensively cultivated forage crops worldwide and earning it the reputation of "the king of pastures." Beyond its nutritional benefits, purple alfalfa plays a crucial role in nitrogen fixation, enhancing soil fertility and aiding in the reduction of chemical fertilizers' use, which is significant for the sustainable development of agriculture (Ye *et al.* 2022). Purple alfalfa is vital in the development of grasslands and livestock industries, especially in arid and semi-arid regions（Cao *et al.* 2011）. Alfalfa hay is a critical feed source for dairy cows and other livestock, significantly impacting animal health and productivity due to its protein content（Fustini *et al.* 2017）. Improper drying methods can degrade the nutritional quality of alfalfa, diminishing its feed value and potentially leading to livestock poisoning, affecting the quality of dairy products. However, current methods for assessing the protein content in alfalfa hay have limitations, especially in terms of rapid and non-destructive testing. Therefore, developing a new technology for the quick and accurate assessment of protein content in purple alfalfa hay is particularly important.

Near-infrared spectroscopy (NIR) is a non-destructive analytical method capable of detecting different absorbance frequencies of specific molecules within substances. Its rapid and non-destructive nature makes it particularly well-suited for analyzing functional

groups in proteins, such as amide groups, whose infrared absorbance band characteristics can accurately reflect the content and quality of proteins (Huck *et al.* 2020). By measuring the interaction between the sample and near-infrared light within a specific wavelength range, materials with different components exhibit unique spectral features regarding light absorption, scattering, and reflection. The varying concentrations of the same component are indicated by different intensities of characteristic absorbance peaks (Hell *et al.* 2016). Analyzing these features provides chemical and physical information about agricultural products. Recently, NIR spectroscopy has been applied in various fields, including food, pharmaceutical, and chemical engineering (Lopes *et al.* 2015), tea (Shen *et al.* 2022), wood (Acuna-Gutierrez *et al.* 2021), and feed. There are also reports of using infrared spectroscopy to detect molds in food (Ma *et al.* 2023).

Near-infrared (NIR) spectroscopy has been employed in the grain and feed industries to determine the content of moisture, protein, fiber, and fat. By establishing a relationship model between the moisture content and the spectral characteristics of samples, the moisture content in grains and feeds can be determined, providing a method for the rapid and accurate assessment of their dryness (Phetpan 2019). Due to their chemical bond structures, proteins produce specific spectral features in light absorption and scattering, enabling the prediction of protein content in grains and feeds, which is crucial for feed production and grain quality control (Masithoh *et al.* 2020). NIR spectroscopy can determine fiber content through specific spectral responses generated by the fiber components in samples (Chen *et al.* 2020). Similarly, it can rapidly determine fat content by utilizing the optical properties of fats, offering key data support for feed formulation (Bilal *et al.* 2020). Discovering the links and patterns between NIR spectroscopy and agricultural products allows for the practical analysis and detection of agricultural product quality (Cortés *et al.* 2019).

With the rapid advancement of computer science and artificial intelligence, machine learning algorithms are increasingly applied to the processing of near-infrared (NIR) spectroscopy data. As a non-destructive analytical technique, NIR spectroscopy obtains chemical and physical information about samples by measuring their absorbance and scattering spectra (Mishra *et al.* 2019; Cortés *et al.* 2019; Zhang *et al.* 2020). The application of machine learning in the analysis and modeling of NIR spectroscopy data can achieve various objectives, such as prediction and classification (Ciza *et al.* 2019), feature extraction and dimensionality reduction, anomaly detection, and quality control (Gao *et al.* 2018), significantly enhancing the efficiency and accuracy of material analysis and testing.

In the construction of a quantitative detection model for purple alfalfa, the preprocessing of spectral data and the extraction of characteristic wavelengths are of paramount importance. Effective preprocessing of spectral data can eliminate or reduce the interference from non-target factors such as noise and baseline drift, thereby enhancing the accuracy and repeatability of the analysis (Saly *et al.* 2010). Feature extraction plays a crucial role in identifying the most representative and relevant information from complex spectral data, pinpointing the characteristic wavelengths closely related to protein content with greater precision (Zhang *et al.* 2023). These characteristic wavelengths are key to building a high-accuracy quantitative detection model. Compared to conventional models, models employing characteristic wavelengths significantly improve precision and efficiency by precisely identifying specific wavelengths closely associated with target attributes, such as protein content. This approach reduces the need to process redundant information, allowing the model to focus more on key data. Consequently, under similar conditions, it achieves higher predictive performance and stability with lower

computational costs (Li *et al*. 2023).

This study employed a variety of machine learning algorithms, including Extreme Learning Machine (ELM), Partial Least Squares Regression (PLSR), Support Vector Machine (SVM), and Long Short-Term Memory (LSTM) networks. The ELM algorithm demonstrated advantages in providing fast learning speed and high generalization capability, while PLSR is suited for handling high-dimensional data. SVM showed strong performance in small sample sizes, non-linearity, and high-dimensional pattern recognition, and the LSTM network excelled in processing time-series data. The combination of these methods enabled the study to predict the protein content in purple alfalfa hay more accurately and efficiently. This approach offers new perspectives and methods for related research. Compared to the authors' earlier article published in *BioResources* in 2023, Vol. 18, pages 5399-5416, the study displays several differences and innovations. Significantly, the spectrometer and detection range used in this study differ from previous research. Observations were made with the Quality Spec Pro visible/near-infrared spectrometer from ASD Inc., USA. It has a detection range of 350 to 1830 nm, which is better suited for capturing spectral information related to protein content. Moreover, in terms of research focus, this study concentrated on the quantitative analysis of protein content in alfalfa using visible/near-infrared spectroscopy, rather than merely classifying alfalfa's moldy state or drying method. A significant correlation was found between visible/near-infrared spectroscopy and the protein content in alfalfa, enabling rapid and accurate detection of protein in dried alfalfa. Additionally, in the establishment of machine learning models, this work introduced different algorithms and modeling methods from the aforementioned study and employed optimization algorithms to enhance prediction accuracy.

In summary, this study successfully enhanced the accuracy and efficiency of protein content detection in purple alfalfa hay by utilizing advanced spectrometric equipment and innovative data processing and modeling techniques. The objectives of this research were as follows: (1) Conditioning and conventional protein content analysis of alfalfa hay; (2) Preprocessing and average spectral analysis of dried alfalfa; (3) Determining the optimal characteristic wavelengths using Competitive Adaptive Reweighted Sampling and Iteratively Retains Informative Variables algorithms; (4) Constructing quantitative detection models for alfalfa using Extreme Learning Machine and Partial Least Squares Regression; (5) Improving model predictive capability by using Principal Components derived from PLSR as independent variables, with Support Vector Machine (SVM) and Long Short-Term Memory (LSTM) networks for regression prediction.

EXPERIMENTAL

Preparation of Experimental Samples

The samples utilized in this study were sourced from the experimental fields of Inner Mongolia Agricultural University. To ensure the accuracy and rigor of the experiment, while minimizing sampling errors, a strict sample selection and processing protocol was adopted. Specifically, prior to sampling, impurities such as weeds, nails, artificially damaged specimens, and decayed alfalfa were removed. Samples with similar plant heights and leaf areas were selected (average plant height ranged from 71.85 cm to 82.31 cm, leaf length from 1.25 cm to 2.25 cm, and leaf width from 1 cm to 2.5 cm). The initial moisture content of the harvested alfalfa was 80±3%. Drying treatments included

both natural sun drying and ventilated shade drying after mold formation. During the drying process, a DSH-50-10 electronic moisture meter was used to measure the moisture content every two hours, to monitor the drying process, and to ensure that the final moisture content of alfalfa stabilized between 15% and 20%. Naturally sun-dried alfalfa was placed under direct sunlight, while moldy alfalfa was sealed in Ziplock bags and stored in the laboratory until ventilated shade drying was performed after mold formation. After drying, alfalfa samples were ground into a fine powder (100 mesh) using a pulverizer, and 15g was weighed and stored in Ziplock bags. A total of 120 alfalfa hay samples were prepared, and the protein content of these samples was determined using a Kjeltec 8420 automatic Kjeldahl apparatus and recorded.

Infrared Spectral Acquisition

The spectrometer used in this study was the Quality Spec Pro from Analytical Spectral Devices, Inc. (ASD), USA. The wavelength range of the spectrometer was 350 to 1830 nm, with a spectral sampling interval of 1 nm. The visible-near infrared spectrometer was preheated for 30 minutes before collecting dark and reference spectra for calibration. Measurements were taken in a dark environment (dark box) to avoid stray light interference. The fiber optic probe was placed 12 cm above the sample's surface vertically. Each petri dish sample was measured three times, generating three sets of spectral data per sample. The average of these three sets was taken as the spectral reflectance test value for the alfalfa sample. After collection, the spectral data were imported into a computer for analysis using ViewSpecPro software, resulting in an average spectrum reflectance in the wavelength range of 350 to 1830 nm.

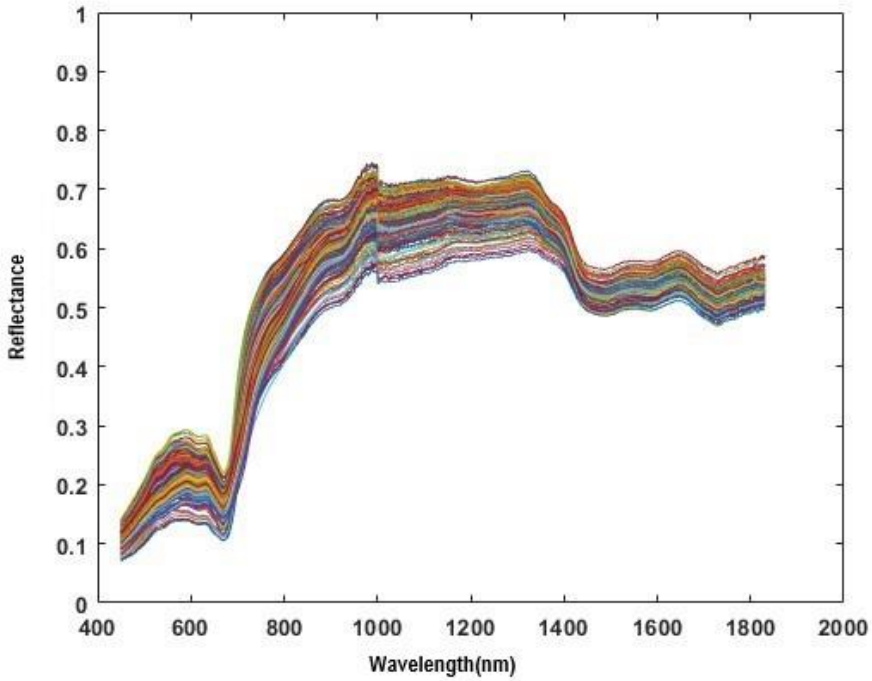

Fig. 1. Reflectivity curve of alfalfa hay samples

To improve the accuracy of visible/near-infrared spectroscopy measurements and enhance the signal-to-noise ratio of the spectra, noisy spectra in the 350 to 449 nm range were excluded. Thus, the effective wavelength range was 450 to 1830 nm, as shown in Fig.

1. In the graph, each curve corresponds to the spectral reflectance test values of a sample from 450 to 1830 nm, with a total of 120 curves. During the measurement process, dark and reference spectra were collected every 10 minutes for recalibration to ensure measurement accuracy.

Pretreatment of the Spectral Data

Due to the susceptibility of spectral data to instrument noise and surrounding environmental factors, the original spectral curves of alfalfa hay often contain numerous spikes, which can impact subsequent model building. Therefore, it is necessary to preprocess the average spectrum to eliminate machine noise and baseline drift. This study selected Savitzky-Golay (SG) convolution smoothing, Standard Normal Variate (SNV), Multiplicative Scatter Correction (MSC), and First Derivative (FD) algorithms for preprocessing. SG smoothing enhances the smoothness of the spectrum, reducing noise interference (Jiao *et al.* 2020). SNV is primarily used to address surface scattering effects and variations in light intensity on the spectrum (Oliveri *et al.* 2019). MSC is employed to eliminate the impacts of particle size and scattering caused by particle inhomogeneity (Makino *et al.* 2016). The first derivative operation (FD) eliminates baseline shifts (Yang *et al.* 2019). This study preprocessed spectral data using The Unscrambler X10.4 and MATLAB software. The processed spectral curve is shown in Fig. 2.

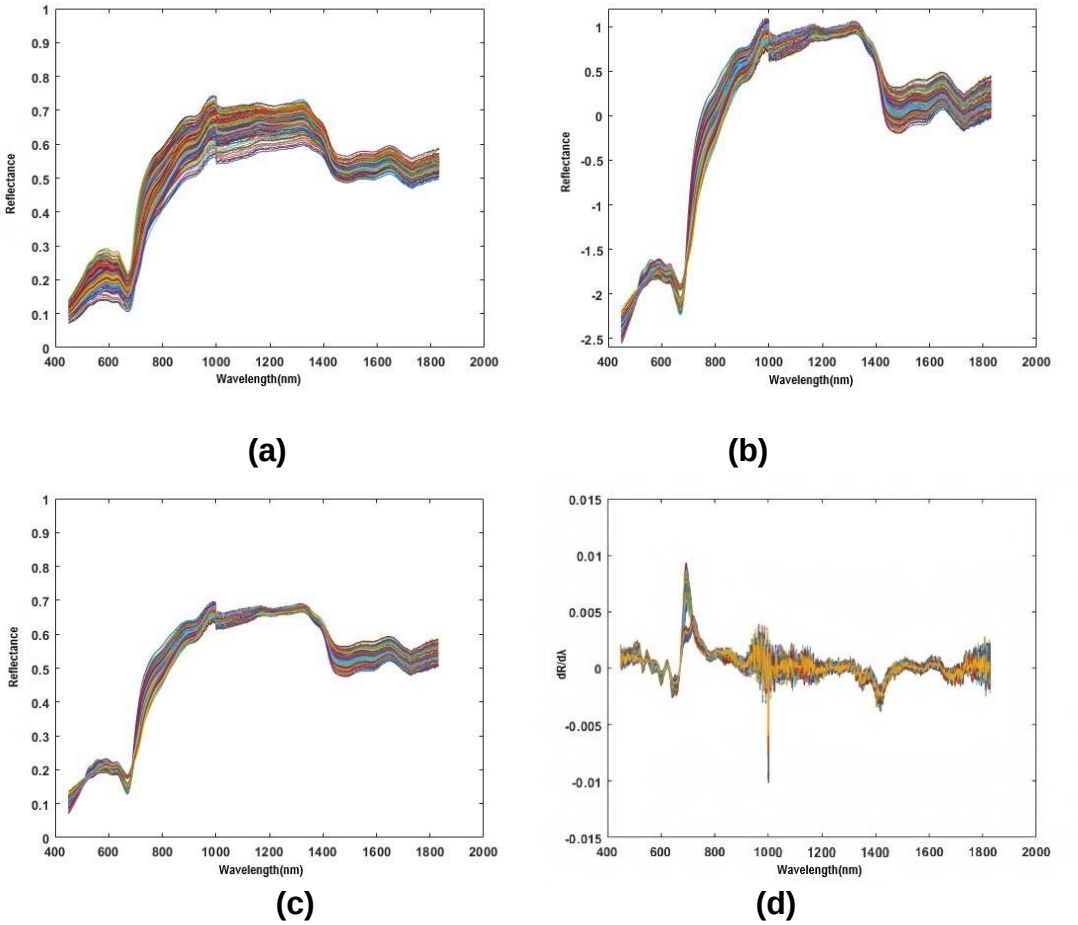

Fig. 2. Spectral curve after preprocessing：(a) SG, (b) SNV, (c) MSC, (d) FD.

The Savitzky-Golay filtering algorithm optimizes the moving average method and is extensively utilized for data denoising and smoothing. Its core advantage lies in its ability to maintain the original shape and width of the signal unaffected during filtering. This algorithm, through specific computational formulas, effectively balances the relationship between denoising and the preservation of signal characteristics in signal processing.

Standard Normal Variate (SNV) transformation is a preprocessing technique aimed at minimizing the impact of sample granularity, surface reflection properties, and path length differences on the reflectance of near-infrared spectroscopy data. This method adjusts the data to ensure the accuracy and consistency of the analysis, making it suitable for improving the quality of near-infrared spectroscopy data.

Multiplicative Scatter Correction (MSC) is a technique for enhancing spectroscopic data, primarily by reducing spectral variability caused by sample scattering, thus enhancing the correlation between spectroscopic data and analytical results. This method standardizes the data through necessary scaling and shifting corrections by comparing all sample spectra with a selected reference spectrum, ideally chosen based on the mean of all sample spectra.

In data acquisition, it is challenging to completely eliminate errors caused by background color or other factors. The application of the First-Order Derivative (FD) algorithm can effectively remove the influence of baseline drift or background noise, while enhancing the resolution and sensitivity by increasing the distinguishability of overlapping peaks. First-order derivative processing alters the shape of the spectrum, providing information through emphasizing the rate of change rather than the absolute intensity. This aids in the identification and quantitative analysis of specific components, although it may complicate data interpretation.

Characteristic Wavelength Selection

Feature extraction plays a pivotal role in near-infrared spectroscopy analysis (Jo *et al.* 2020), enabling the extraction of crucial information from complex spectral data related to the properties of the substances under investigation. This process reduces data dimensionality, simplifies the model-building process, and enhances the accuracy of predictions (Mei *et al.* 2019). In this study, aimed at facilitating rapid and non-destructive detection of nutritional substances in purple alfalfa, the collected spectral data in the 450 to 1830 nm range underwent preprocessing using the four different methods that were described earlier. This was followed by integrating Competitive Adaptive Reweighted Sampling (CARS) and Iteratively Retains Informative Variables (IRIV) algorithms to extract characteristic wavelengths.

Competitive adaptive reweighted sampling (CARS)

The Competitive Adaptive Reweighted Sampling (CARS) algorithm is a feature selection method based on competitive neural networks. It selects feature wavelengths highly relevant to the target variable through a competitive, adaptive approach. The algorithm iteratively adjusts weights based on the interaction and importance of feature wavelengths, thereby selecting the most representative wavelengths (Xie *et al.* 2022). The analysis process of the CARS algorithm is as follows:

(1). Monte Carlo model sampling: The dataset is randomly divided for model construction, with a split ratio of 80% to 90%, to establish a PLS (Partial Least Squares) model. This process yields the regression coefficient for the i^{th} wavelength .

(2). Exponential Decay Wavelength Selection: The method uses an exponentially decreasing function (EDF) to forcibly eliminate wavelengths having relatively small

absolute weight in regression coefficients. The retention rate of variables is . where '*j*' denotes the j^{th} Monte Carlo sampling, '*N*' represents the total number of Monte Carlo samplings, and the parameters '*a*' and '*b*' are constants.

(3). Adaptive Reweighting Sampling: Selection is conducted using the evaluation weights as in .

(4). Cyclic Iteration: The process involves iterative calculations based on a set number of cycle iterations. The optimal set of variables is based on the minimum cross-validation root mean square error, representing the desired characteristic variables.

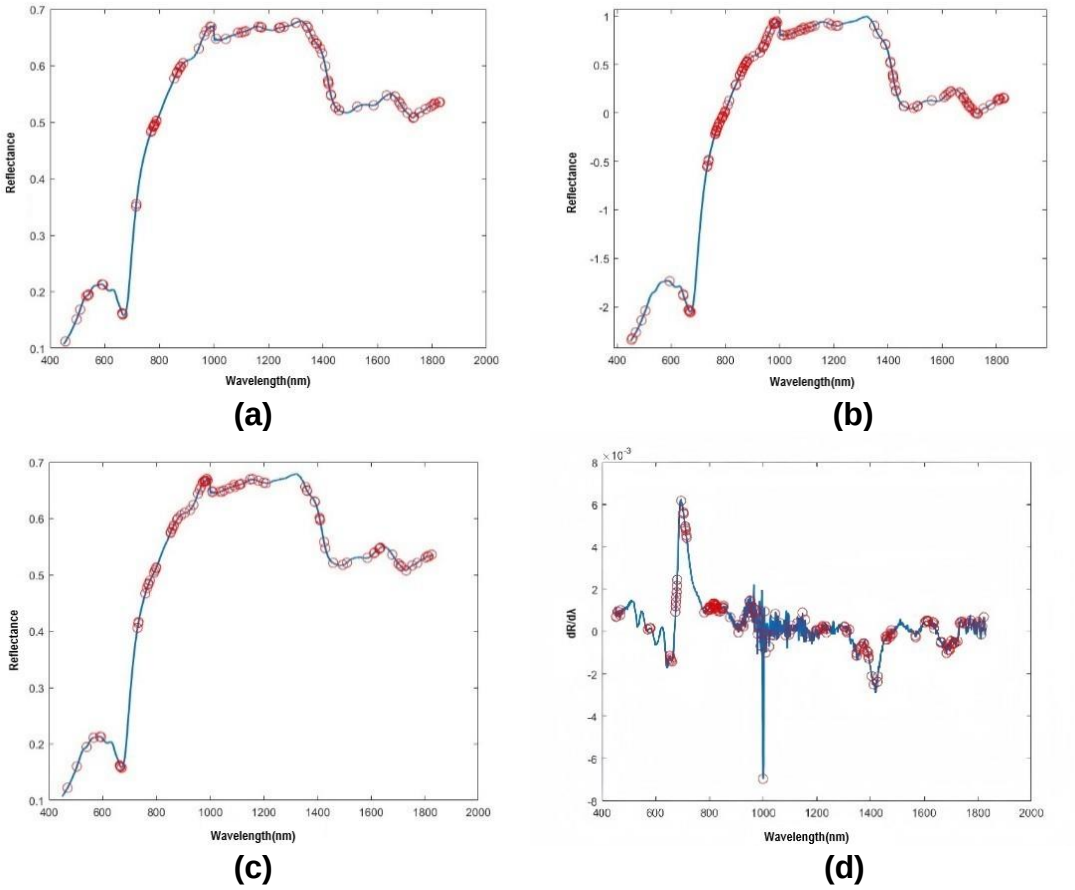

Fig. 3. Selection of Characteristic Wavelengths after CARS : (a) SG-CARS, (b) SNV-CARS, (c) MSC-CARS, (d) FD-CARS

When applying the CARS method for extracting feature wavelengths from preprocessed spectral data using four different methods, the Monte Carlo sampling was set to 50 times, employing a 10-fold cross-validation. The process of variable reduction exhibited an exponential decay, with a rapid decrease in the number of variables in the initial phase and a much slower decrease in the second phase, indicating "rough" and "fine" selection stages (Chen *et al.* 2020; Li *et al.* 2022). The change in the 10-fold cross-validation root mean square error initially decreases and then gradually increases, suggesting that less relevant wavelengths to protein content in alfalfa spectral data are discarded initially, and later, due to high selectivity, some critical parameters are excluded, leading to a gradual increase in error. The best iteration numbers for different preprocessing methods were as follows: SG convolution smoothing (18 iterations, 96 feature wavelengths, 6.95% of the full spectrum); Standard Normal Variate transformation (22

iterations, 143 feature wavelengths, 10.35% of the full spectrum); Multiplicative Scatter Correction (21 iterations, 96 feature wavelengths, 6.95% of the full spectrum); First Derivative operation (19 iterations, 163 feature wavelengths, 11.8% of the full spectrum). The feature wavelengths extracted by different preprocessing methods are illustrated in Fig. 3. The blue line represents the average spectral data after preprocessing, and the red circles denote the characteristic wavelengths extracted by the CARS algorithm.

Iterative Retention of Information Variables Method (IRIV)

The Iteratively Retains Informative Variables (IRIV) algorithm is a method used for feature selection to identify the most relevant subset of variables from a large pool about a target variable (Yu *et al.* 2018). It is particularly suitable for wavelength selection in spectral analysis. The basic process of the IRIV algorithm can be outlined as follows:

Iterative Retention of Informative Variables: A subset of variables is selected from the current variable set in each iteration. These variables are chosen because they maximally retain relevant information about the target variable.

Assessment of Variable Importance: The importance of each selected variable is evaluated. This is typically done by examining each variable's contribution to the model's predictive performance. The assessment is often based on statistical indicators such as the magnitude of regression coefficients, the impact of variables on model prediction error, and the consistency of model performance across different datasets. The assessment is often based on statistical indicators such as the magnitude of regression coefficients, the impact of variables on model prediction error, and the consistency of model performance across different datasets.

Cyclic Iteration and Optimization: The above process is repeated for a predetermined number of iterations or until specific stopping criteria are met (such as minimization of cross-validation error). After each iteration, the variable set is updated, removing those deemed unimportant or contributing less to the prediction of the target variable.

Determination of the Final Feature Set: The variable set obtained at the end of the iterative process represents the selected feature variables. These variables are considered the most important for predicting the target variable and can be used in subsequent data analysis or modeling processes.

The Iteratively Retains Informative Variables (IRIV) algorithm effectively selects the most crucial subset of variables from a large set through an iterative process, enhancing the model's explanatory power and predictive accuracy. This is particularly applicable to spectral data and other high-dimensional data analyses. The core of the IRIV method in feature wavelength selection involves iterative feature selection. In each iteration, it assesses the impact of remaining features on the model's performance and selects those that minimize the prediction error (mean square error). This process is repeated until the maximum number of iterations is reached, or no remaining features are left (Xu *et al.* 2019). In this study, the feature wavelengths obtained from SG convolution smoothing were 59 (4.27% of the full spectrum), from Standard Normal Variate transformation were 69 (5.00% of the full spectrum), from Multiplicative Scatter Correction were 57 (4.13% of the full spectrum), and from the First Derivative operation were 51 (3.69% of the full spectrum). The results of feature wavelength selection using the IRIV method after various preprocessing methods are shown in Fig. 4. The blue line represents the average spectral data after preprocessing, and the red circles denote the characteristic wavelengths extracted by the IRIV algorithm.

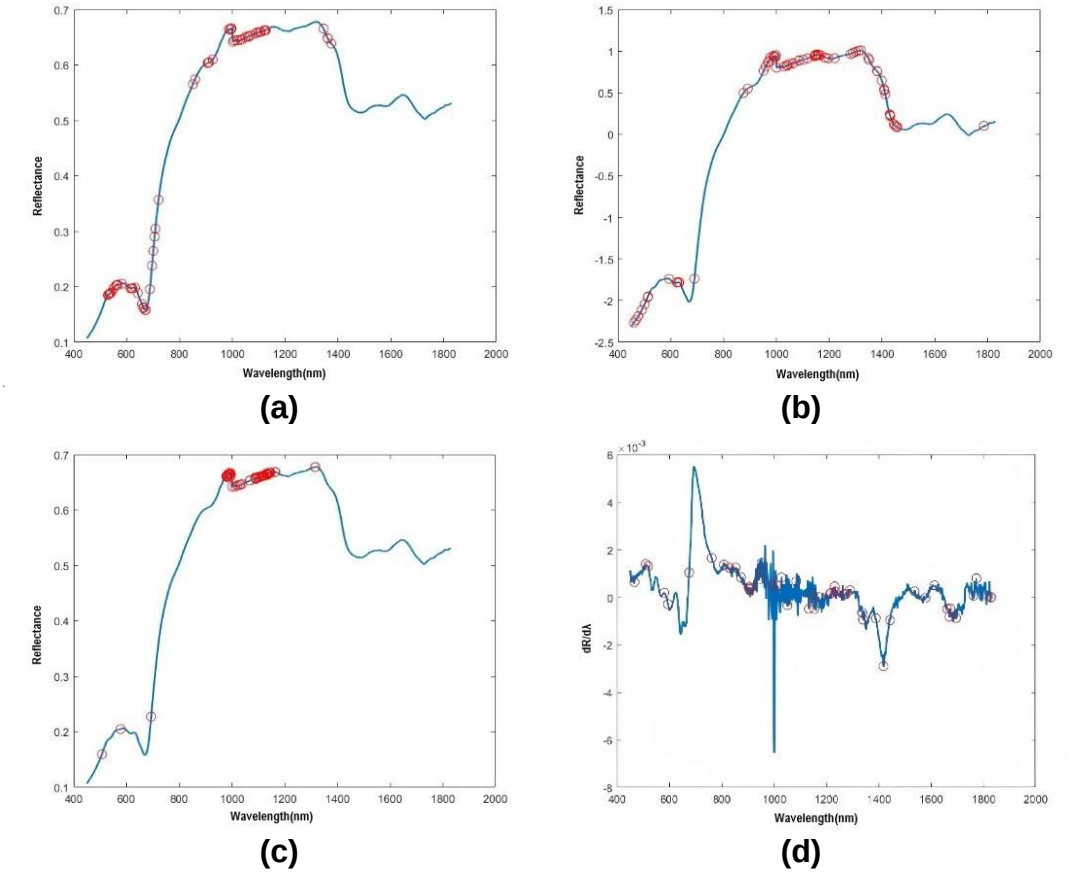

Fig. 4. Selection of characteristic wavelengths after iriv: (a) SG-IRIV, (b) SNV-IRIV, (c) MSC-IRIV, (d) FD-IRIV.

Based on the analysis results, the identified feature wavelengths predominantly correspond to functional groups such as C-H, O-H, N-H, C=O, and -CHO. The wavebands near 1100 to 1160 nm and 1428 to 1491 nm are associated with O-H groups (Rego *et al.* 2020), related to the moisture content in the feed; absorption peaks near 1470, 1500 to 1530, and 1640 to 1680 nm correspond to the stretching vibrations of N-H groups, which are related to crude protein in the feed (Rego *et al.* 2020). The selected feature wavelengths reflect the characteristic absorption bands of moisture, protein, and other substances in dried alfalfa. In subsequent modeling, these wavelengths can effectively reduce computational load, decrease the redundancy of spectral data, and improve model accuracy. The number of feature wavelengths extracted using different preprocessing methods and feature wavelength extraction techniques is summarized in Table 1.

Table 1. Wavelength Selection after CARS and IVIR

Pre-Processing Technique	Method	Feature Variables Number
SG	CARS	96
	IVIR	59
SNV	CARS	143
	IVIR	69
MSC	CARS	96
	IVIR	57
FD	CARS	163
	IVIR	51

Establishment and Evaluation of the Dried Alfalfa Protein Prediction Model

This study aimed to achieve rapid and non-destructive detection of nutritional substances in purple alfalfa using near-infrared spectroscopy technology. For this purpose, various machine learning algorithms, including Partial Least Squares Regression (PLSR), Extreme Learning Machine (ELM), Support Vector Machine (SVM), and Long Short-Term Memory networks (LSTM), were employed to establish and evaluate a prediction model for the protein content in purple alfalfa.

Initially, prediction models were developed using the PLSR method for purple alfalfa's full-spectrum and feature wavelength spectral data. PLSR is a classic regression method that establishes a linear regression model by maximizing the correlation between input and output variables (Niu *et al.* 2021). This study used full-spectrum or feature wavelength spectral data as input variables. Protein content was used as the output variable to build a predictive model for the nutritional substances in purple alfalfa.

Establishing the PLSR model involved two steps: model training and model validation. The collected purple alfalfa samples were divided into a calibration set and a prediction set, with the calibration set comprising 70% of the total samples and the prediction set comprising the remaining 30%. The calibration set was used for training and optimizing the model, while the prediction set was used to assess the model's generalizability and predictive accuracy. During training, the model's coefficients and intercept were determined by minimizing the sum of squared residuals. In the optimization process, the best number of principal components and regularization parameters were selected through cross-validation to enhance the model's stability and generalizability (Belini *et al.* 2011). After the training, the model was evaluated using the prediction set. The predictive accuracy and generalizability of the model were assessed by calculating the determination coefficient (R^2) and root mean square error (RMSE) for both the calibration and prediction sets. A determination coefficient closer to 1 indicates a better model fit to the data, and a smaller RMSE indicates a lower prediction error.

The study also employed the Extreme Learning Machine (ELM) method for model establishment. ELM is a nonlinear regression method based on artificial neural networks, which quickly trains the network to obtain good predictive results by randomly generating initial weights and biases.

In the ELM method, the model's performance is optimized by adjusting the number of neurons in the hidden layer and selecting the activation function. During training, the outputs of the hidden layer neurons are computed using randomly generated weights and biases. Then, the weights and biases of the output layer are calculated using the least squares method. The best-performing ELM model is obtained by continuously adjusting the number of neurons in the hidden layer and the activation function (Leuenberger and

Kanevski 2015; Pradhan *et al.* 2019; Jiang *et al.* 2020).

Like the PLSR method, establishing the ELM model also includes training and validation steps, with the same distribution of samples in the calibration and prediction sets. The optimal number of hidden layer neurons and activation function are selected through cross-validation to improve the model's stability and predictive accuracy.

After the model training, the model is evaluated using the prediction set. The predictive accuracy and generalizability of the model are assessed by calculating the determination coefficient (R^2) and root mean square error (RMSE) for both the calibration and prediction sets.

Through model establishment in this study, ELM and PLSR models based on full spectrum and feature wavelengths were developed to predict the protein content in purple alfalfa.

Experimental results are presented in Tables 2 and 3. Analyzing Table 1 from the perspective of spectral data preprocessing, it is observed that the various preprocessing methods improved the accuracy of both calibration and prediction sets of the models. From the perspective of feature wavelengths in Table 3, it was noted that the number of feature wavelengths extracted by the IVIR algorithm was significantly less than those extracted by the CARS method. Moreover, the models using the IVIR algorithm showed lower calibration and prediction set accuracies compared to other models, possibly due to the exclusion of wavelengths highly relevant to the protein content in dried alfalfa during the IVIR selection process, leading to poorer predictive accuracy in the calibration and test sets. From the perspective of model establishment, Table 3 indicates that the MSC-CARS-PLSR model had strong predictive capability, with a calibration set root mean square error (RMSE) of 0.1922, a determination coefficient (R^2) of 0.9972, a prediction set RMSE of 0.6581, and a determination coefficient of 0.9446. In establishing a prediction model for the protein value of dried alfalfa, it is evident that the PLSR model performed better than the ELM model. The prediction results show that the accuracy of the full-spectrum prediction model was lower than that of the feature wavelength prediction model, and the MSC-CARS-PLSR model improved the prediction accuracy by 15.8% and reduced the prediction set RMSE by 33.4% compared to the MSC-PLSR model. This demonstrates that extracting feature wavelengths significantly simplified the computational model and enhanced prediction accuracy.

Table 2. Prediction Results of Full-spectrum ELM and PLSR Models using Different Preprocessing Methods

Model	Pretreatment	Calibration Set		Prediction Set	
ELM	No Without	1.5110	0.8577	1.5303	0.6443
	SG	0.9731	0.9145	1.5084	0.6527
	SNV	1.0325	0.8963	1.9715	0.5823
	MSC	1.1485	0.9004	1.1923	0.7474
	FD	1.0041	0.8997	2.1570	0.5336
PLSR	No Without	1.4201	0.8443	1.3363	0.6850
	SG	0.4354	0.9744	1.1533	0.6997
	SNV	0.4451	0.9431	1.1126	0.7322
	MSC	0.4126	0.9752	0.9875	0.8155
	FD	0.5127	0.9321	1.3321	0.6954

Based on these results, to further improve model accuracy, two machine learning algorithms, Support Vector Machine (SVM) and Long Short-Term Memory network (LSTM), were introduced, forming two new model combinations: MSC-CARS-PLSR-SVM and MSC-CARS-PLSR-LSTM. These were aimed at utilizing the strengths of each algorithm to enhance further the accuracy and reliability of predicting the protein content in dried alfalfa.

Table 3. Prediction Results of ELM and PLSR Models after Feature Variable Selection

Model	Algorithm Combinations	Feature Variables Number	Calibration Set		Prediction Set	
ELM	SG-CARS-ELM	96	0.2600	0.9351	1.3271	0.7359
	SG-IRIV-ELM	59	1.1354	0.8021	1.7059	0.6059
	SNV-CARS-ELM	143	0.1464	0.9763	1.0822	0.7997
	SNV-IRIV-ELM	69	0.8813	0.9015	1.2114	0.7908
	MSC-CARS-ELM	96	0.1903	0.9655	1.2099	0.7204
	MSC-IRIV-ELM	57	1.3442	0.7501	2.1128	0.7454
	FD-CARS-ELM	163	0.2832	0.9456	1.5775	0.6446
	FD-IRIV-ELM	51	1.3402	0.7318	2.1308	0.7442
PLSR	SG-CARS-PLSR	96	0.2044	0.9930	0.7581	0.9362
	SG-IRIV-PLSR	59	1.4871	0.6725	2.0921	0.7237
	SNV-CARS-PLSR	143	0.0983	0.9953	0.7651	0.9212
	SNV-IRIV-PLSR	69	1.3304	0.7655	2.193	0.7315
	MSC-CARS-PLSR	96	0.1922	0.9972	0.6581	0.9446
	MSC-IRIV-PLSR	57	1.2184	0.7808	2.8432	0.6124
	FD-CARS-PLSR	163	0.1940	0.9939	1.3465	0.7723
	FD-IRIV-PLSR	51	1.5201	0.7133	2.6244	0.5982

The MSC-CARS-PLSR-SVM model combines the feature extraction and data preprocessing capabilities of the MSC-CARS-PLSR model with the robust regression function of SVM (Lee *et al.* 2023). SVM excels at complex, high-dimensional datasets and is especially adept at handling small-sample datasets and nonlinear problems (Vabalas *et al.* 2019). In this model, SVM served as a secondary prediction model, using the outputs of the PLSR model as its inputs to refine and optimize the prediction results further. The Radial Basis Function (RBF) was used as the kernel function in the model establishment process, with the penalty factor (c) and RBF parameter (g) set. The model was built with a 10-fold cross-validation method with the svmtrain function.

The MSC-CARS-PLSR-LSTM model combines the feature extraction ability of MSC-CARS-PLSR with the time series data processing advantage of LSTM. LSTM networks are suitable for processing data with strong time dependencies and can effectively capture long-term dependencies in time series (Fagerström *et al.* 2019). In this model, LSTM was used to analyze and predict time-varying alfalfa sample data further to enhance prediction accuracy. The model structure included a sequence input layer, an LSTM layer, a ReLU activation layer, a fully connected layer, and a regression layer. The Adam optimizer was used, with parameters set for mini-batch size, maximum number of iterations, initial learning rate, and learning rate drop strategy. Parameters were adjusted to enhance prediction accuracy.

These two new model combinations were used to process the same dataset, with the calibration and prediction sets comprising 70% and 30% of the total samples,

respectively, and were compared with the original MSC-CARS-PLSR model. Comparing the prediction results of the three models, the predictive capability of the MSC-CARS-PLSR-LSTM model was similar to that of the MSC-CARS-PLSR model. In contrast, the MSC-CARS-PLSR-SVM model provided more accurate and stable prediction results than the MSC-CARS-PLSR model, with a calibration set RMSE of 0.1088, a determination coefficient of 0.9982, a prediction set RMSE of 0.5230, and a determination coefficient of 0.9645. Specific experimental results are presented in Table 4. The scatter plot of predicted values versus actual values for the MSC-CARS-PLSR-SVM model is shown in Fig. 5.

Table 4. Model Prediction Results after Model Optimization

Algorithm Combinations	Calibration Set		Prediction Set	
MSC-CARS-PLSR	0.1922	0.9972	0.6581	0.9446
MSC-CARS-PLSR-SVM	0.1088	0.9982	0.5230	0.9645
MSC-CARS-PLSR-LSTM	0.2071	0.9943	0.6361	0.9449

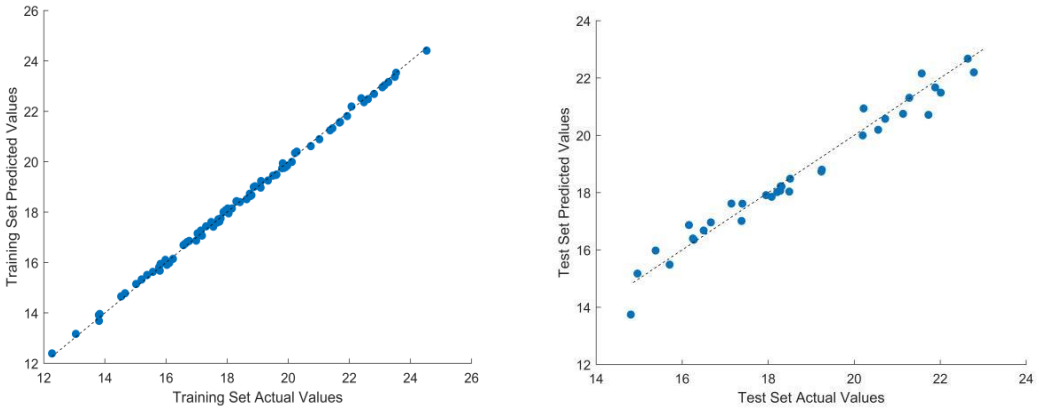

Fig. 5. Scatter Plot of Predicted vs. Actual Values for the MSC-CARS-PLSR-SVM Model

RESULTS AND DISCUSSION

This study successfully established predictive models for the nutritional substances in purple alfalfa using near-infrared spectroscopy combined with various preprocessing and feature extraction methods, and these models underwent detailed performance evaluations. Initially, in the preprocessing phase, four methods were applied to the average spectra of purple alfalfa samples in the 450 to 1830 nm range: Standard Normal Variate (SNV), Multiplicative Scatter Correction (MSC), Savitzky-Golay (SG) smoothing, and First Derivative (FD). The SNV method effectively reduced the impact of surface scattering and intensity changes, MSC addressed issues caused by particle inhomogeneity, and SG smoothing significantly enhanced spectral smoothness, effectively reducing noise interference.

In terms of feature extraction, both the Competitive Adaptive Reweighted Sampling (CARS) and Iteratively Retains Informative Variables (IRIV) algorithms were used, enabling more effective extraction of crucial information related to the protein content in purple alfalfa from complex spectral data. Additionally, the Partial Least Squares Regression (PLSR) and Extreme Learning Machine (ELM) methods were used to establish protein content prediction models for both the full spectrum and feature wavelengths of purple alfalfa. After evaluating the models on calibration and prediction sets, it was found that the feature wavelength models demonstrated superior predictive performance compared to full-spectrum models, as indicated by higher determination coefficients and lower root mean square errors (RMSE). Notably, the MSC-CARS-PLSR model showed a determination coefficient of 0.9972 and an RMSE of 0.1922 on the calibration set and a determination coefficient of 0.9446 and an RMSE of 0.6581 on the prediction set, indicating that this model can accurately and reliably predict the nutritional content of purple alfalfa.

Furthermore, model accuracy was enhanced by introducing a Support Vector Machine (SVM) and Long Short-Term Memory Network (LSTM) for regression prediction of the principal factors derived from PLSR. The MSC-CARS-PLSR-SVM model, in particular, exhibited a determination coefficient of 0.9982 and an RMSE of 0.1088 on the calibration set, and a determination coefficient of 0.9645 and an RMSE of 0.5230 on the prediction set, further improving the predictive accuracy of the model.

Overall, this study not only successfully established predictive models for the nutritional substances in purple alfalfa using near-infrared spectroscopy and advanced algorithms but also confirmed the effectiveness and reliability of these models in accurately predicting the nutritional content of purple alfalfa through comprehensive performance evaluations. These achievements provide new methods and technical support for purple alfalfa's quality assessment and nutritional substance monitoring.

However, there are some limitations to this study. First, the relatively small sample size may affect the generalizability of the models. Increasing the sample size and conducting more validation experiments could enhance model performance further. Second, this study focused solely on predicting the protein content of purple alfalfa and did not consider other essential nutrients. Future research could expand the scope of the models to predict more nutritional substances in purple alfalfa. Additionally, the reliability and stability of the models need further verification in practical applications to ensure their effective use in different scenarios.

CONCLUSIONS

1. This study successfully developed an accurate and non-destructive method to predict the protein content in purple alfalfa by integrating near-infrared spectroscopy with machine learning algorithms. The application of various preprocessing and feature extraction techniques led to the MSC-CARS-PLSR model exhibiting the best performance among all tested models, offering high precision and reliability.
2. The results demonstrate that the predictive accuracy of the model can be further enhanced by employing regression predictions with Support Vector Machines (SVM) and Long Short-Term Memory networks (LSTM). Specifically, the MSC-CARS-PLSR-SVM model showed superior predictive performance, reflected in higher determination coefficients and lower root mean square errors in both the calibration and prediction sets.
3. The methodologies and findings of this study provide new perspectives and technical support for quality assessment and nutritional substance monitoring of purple alfalfa, laying a foundation for future research and practical applications in related fields. Additionally, the study emphasizes the importance of increasing the sample size and further verifying the stability of the models to enhance the performance of the models.

ACKNOWLEDGMENTS

This project is supported by the National Natural Science Foundation of China (Grant No. 32060414); Natural Science Foundation of Inner Mongolia, China (2022MS06023,2023QN05034); Natural Science Foundation of The Autonomous Region Military-Civilian Integration Key Research & Soft Science Research Projects of Inner Mongolia, China (JMZD202201); Scientific Research Project of Universities in Inner Mongolia, China (NJZY21461). and the Inner Mongolia Engineering Research Center of Intelligent Equipment for the Entire Process of Forage and Feed Production.

REFERENCES CITED

Acuna-Gutierrez, C., Schock, S., Jimenez, V. M., and Mueller, J. (2021). "Detecting fumonisin B1 in black beans (*Phaseolus vulgaris* L.) by near-infrared spectroscopy (NIRS)," *Food Control* 2021(130), 130. DOI: 10.1016/j.foodcont.2021.108335

Bedaf, M. T., Masoud, B., Ghodratollah, S., Mengoni, A., and Bazzicalupo, M. (2008). "Diversity of *Sinorhizobium* strains nodulating *Medicago sativa* from different Iranian regions," *FEMS Microbiology Letters* 288(1), 40-46. DOI: 10.1111/j.1574-6968.2008.01329.x

Beć, K. B., Grabska, J., and Huck, C. W. (2020). "Near-infrared spectroscopy in bio-applications," *Molecules* 25(12), article 2948. DOI: 10.3390/molecules25122948

Belini, U. L., Hein, P. R. G., Tomazello Filho, M., Rodrigues, J. C., and Chaix, G. (2011). "Near infrared spectroscopy for estimating sugarcane bagasse content in medium density fiberboard," *BioResources* 6(2), 1816-1829. DOI: 10.15376/biores.6.2.1816-1829

Bilal, M., Zou, X., Arslan, M., Tahir, H. E., Azam, M., Junjun, Z., Basheer, S., and

Abdullah. (2020). "Rapid determination of the chemical compositions of peanut seed (*Arachis hypogaea*) Using portable near-infrared spectroscopy," *Vibrational Spectroscopy* 110, 103138. DOI: 10.1016/j.vibspec.2020.103138

Cao, H., Zhang, H. L., Gai, Q. H., Chen, H., and Zhao, M. L. (2011). "Introduction test and comprehensive evaluation of production performance of 22 alfalfa varieties," *Journal of Grass Industry* 20(6), 219-229.

Chen, H., Tan, C., and Lin, Z. (2020). "Quantitative determination of the fiber components in textiles by near-infrared spectroscopy and extreme learning machine," *Analytical Letters* 53(6). DOI: 10.1080/00032719.2019.1683742

Chen, X., Yang, Q., Han, J., Lin, L., and Shi, L. (2020). "Estimation of leaf water content in winter wheat *leaves* based on leaf-scale hyperspectral data," *Spectroscopy and Spectral Analysis* 40(3), 7. DOI: CNKI:SUN:GUAN.0.2020-03-047

Ciza, P. H., Sacre, P-Y, Waffo, C., Coic, L., Avohou, H., Mbinze, J. K., Ngono, R., Marini, R. D., Hubert, Ph., and Ziemons, E. (2019). "Comparing the qualitative performances of handheld NIR and Raman spectrophotometers for the detection of falsified pharmaceutical products," *Talanta* 202, 469-478. DOI: 10.1016/j.talanta.2019.04.049

Cortés, V., Blasco, J., Aleixos, N., Cubero, S., and Talens, P. (2019). "Monitoring strategies for quality control of agricultural products using visible and near-infrared spectroscopy: A review," *Trends in Food Science & Technology* 85, 138-148. DOI: 10.1016/j.tifs.2019.01.015

Fagerström, J., Bång, M., Wilhelms, D., and Chew, M. S. (2019). "LiSep LSTM: A machine learning algorithm for early detection of septic shock," *Scientific Reports* 2019(1), article 15132. DOI: 10.1038/S41598-019-51219-4

Fustini, M., Palmonari, A., Canestrari, G., Bonfante, E., Mammi, L., Pacchioli, M. T., Sniffen, G. C. J., Grant, R. J., Cotanch, K. W., and Formigoni, A. (2017). "Effect of undigested neutral detergent fiber content of alfalfa hay on lactating dairy cows: Feeding behavior, fiber digestibility, and lactation performance," *Journal of Dairy Science* 100(6), article 4475. DOI: 10.3168/jds.2016-12266

Gao, J., Nuyttens, D., Lootens, P., He, Y., and Pieters, J. G. (2018). "Recognising weeds in a maize crop using a random forest machine-learning algorithm and near-infrared snapshot mosaic hyperspectral imagery," *Biosystems Engineering* 170, 39-50. DOI: 10.1016/j.biosystemseng.2018.03.006.

Hell, J., Prückler, M., Danner, L., Henniges, U., Apprich, S., Rosenau, T., Kneifel, W., and Böhmdorfer, S. (2016). "A comparison between near-infrared (NIR) and mid-infrared (ATR-FTIR) spectroscopy for the multivariate determination of compositional properties in wheat bran samples," *Food Control* 365-369. DOI: 10.1016/j.foodcont.2015.08.003

Jiang, H., Liu, T., and Chen, Q. (2020). "Quantitative detection of fatty acid value during storage of wheat flour based on a portable near-infrared (NIR) spectroscopy system," *Infrared Physics & Technology* 109, article 103423. DOI: 10.1016/j.infrared.2020.103423

Jiao, Y., Li, Z., Chen, X., and Fei, S. (2020). "Preprocessing methods for near-infrared spectrum calibration," *Journal of Chemometrics* 29(3), article 682. DOI: 10.1002/cem.3306

Jo, S., Sohng, W., Lee, H., and Chung, H. (2020). "Evaluation of an autoencoder as a feature extraction tool for near-infrared spectroscopic discriminant analysis," *Food Chemistry* 331, article 127332. DOI: 10.1016/j.foodchem.2020.127332

Lee, Y.-J., Lee, T.-J., and Kim, H. J. (2023). "Classification analysis of copy papers using infrared spectroscopy and machine learning modeling," *BioResources* 19(1), 160-182. DOI: 10.15376/biores.19.1.160-182

Leuenberger, M., and Kanevski, M. (2015). "Extreme learning machines for spatial environmental data," *Computers & Geosciences* 85(DEC.PT.B), 64-73. DOI: 10.1016/j.cageo.2015.06.020

Li, L., Zhang, S., Zuo, Z., and Wang, Y. (2022). "Data fusion of multiple-information strategy based on Fourier transform near-infrared spectroscopy and Fourier-transform mid-infrared for geographical traceability of *Wolfiporia coco*s combined with chemometrics," *J. Chemometrics* 36(9), article e3436. DOI: 10.1002/cem.3436

Li, X., Wei, Z., Peng, F., Liu, J., and Han, G. (2023). "Non-destructive prediction and visualization of anthocyanin content in mulberry fruits using hyperspectral imaging," *Frontiers in Plant Science*, 14. DOI: 10.3389/fpls.2023.1137198

Lopes, J. A., Sousa, C., Ferreira, E. C., Mesquita, D. P., and Quintelas, C. (2015). "Near-infrared spectroscopy for the detection and quantification of bacterial contaminations in pharmaceutical products," *International Journal of Pharmaceutics* 492(1-2), 199-206. DOI: 10.1016/j.ijpharm.2015.07.005

Ma, T., Inagaki, T., and Tsuchikawa, S. (2023). "Demonstration of the applicability of visible and near-infrared spatially resolved spectroscopy for rapid and nondestructive wood classification," *Holzforschung* 12(2), 153-162. DOI: 10.1515/hf-2020-0074

Makino, Y., Oshita, S., and Kamruzzaman, M. (2016). "Hyperspectral imaging for real-time monitoring of water holding capacity in red meat," *LWT-Food Science & Technology* 66 (2016), 685-691. DOI:10.1016/j.lwt.2015.11.021

Masithoh, R. E., Amanah, H. Z., Yoon, W. S., *et al.* (2020). "Determination of protein and glucose of tuber and root flours using NIR and MIR spectroscopy," *Infrared Physics & Technology*. DOI: 10.1016/j.infrared.2020.103577

Mei, Q.-P., Li, T.-F., Yao, L.-Z., Liu, X.-H., Hu, Y.-L., and Hu, L. (2019). "Characterization of a wavelength selection method using near-infrared spectroscopy and partial least squares with false nearest neighbors and its application in the detection of the chemical oxygen demand of waste liquid," *Spectroscopy Letters* 52(9), 553-562. DOI: 10.1080/00387010.2019.1676261

Mishra, P., Nordon, A., Mohd Asaari, M. S., Lian, G., and Redfern, S. (2019). "Fusing spectral and textural information in near-infrared hyperspectral imaging to improve green tea classification modeling," *Journal of Food Engineering* 249, 40-47. DOI: 10.1016/j.jfoodeng.2019.01.009

Niu, C., Tan, K., Jia, X., and Wang, X. (2021). "Deep learning based regression for optically inactive inland water quality parameter estimation using airborne hyperspectral imagery," *Environmental Pollution* 286, article 117534. DOI: 10.1016/j.envpol.2021.117534

Oliveri, P., Malegori, C., Simonetti, R., and Casale, M. (2019). "The impact of signal pre-processing on the final interpretation of analytical outcomes – A tutorial," *Analytica Chimica Acta* 1058, 9-17. DOI: 10.1016/j.aca.2018.10.055

Phetpan, V. S. P. (2019). "In-line near infrared spectroscopy for the prediction of moisture content in the tapioca starch drying process," *Powder Technology* 345, 608-615. DOI: 10.1016/j.powtec.2019.01.050

Pradhan, K., Monoj, K., Minz, S., and Shrivastava, V. K. (2019). "Fast active learning for hyperspectral image classification using extreme learning machine," *IET Image Processing* 13(4), 549-555. DOI: 10.1049/iet-ipr.2018.5104

Rego, G., Ferrero, F., Valledor, M., Campo, J. Carlos, Forcada, S., Royo, L. J., and Soldado, A. (2020). "A portable IoT NIR spectroscopic system to analyze the quality of dairy farm forage," *Computers and Electronics in Agriculture* 175, article 105578. DOI: 10.1016/j.compag.2020.105578

Saly, R., Romero, R., Torres, T., Mojgan, M., Moshgbar, M., Jun, J., and Huang, H. (2010). "Practical considerations in data pre-treatment for NIR and Raman spectroscopy," *American Pharmaceutical Review* 13(6), 116-127.

Shen, S., Hua, J., Zhu, H., Yang, Y., Deng, Y., Li, J., Yuan, H., Wang, J., Zhu, J., and Jiang, Y. (2022). "Rapid and real-time detection of moisture in black tea during withering using micro-near-infrared spectroscopy," *LWT* 155(112970). DOI: 10.1016/j.lwt.2021.112970

Vabalas, A., Gowen, E., Poliakoff, E., and Casson, A. J. (2019). "Machine learning algorithm validation with a limited sample size," *PloS One* 14(11), article e0224365. DOI: 10.1371/journal.pone.0224365.

Xie, L., Hong, M., and Yu, Z. (2022). "A wavelength selection method combining direct orthogonal signal correction and Monte Carlo," *Spectroscopy and Spectral Analysis* 42(2), article 6. DOI: 10.3964/j.issn.1000-0593(2022)02-0440-06

Xu, Y., Zhang, H., Zhang, C., Wu, P., Li, J., Xia, Y., and Fan, S. (2019). "Rapid prediction and visualization of moisture content in single cucumber (*Cucumis sativus* L.) seed using hyperspectral imaging technology," *Infrared Physics Technol.* 102, 103034. DOI: 10.1016/j.infrared.2019.103034

Yang, J., Du, L., Gong, W., Shi, S., and Chen, B. (2019). "Analyzing the performance of the first-derivative fluorescence spectrum for estimating leaf nitrogen concentration," *Optics Express* 27(4), article 3978. DOI: 10.1364/OE.27.003978

Ye, Q., Zhu, F., Sun, F., Wang, T. C., Wu, J., Liu, P., Shen, C., Dong, J., and Wang. T. (2022). "Differentiation trajectories and biofunctions of symbiotic and un-symbiotic fate cells in root nodules of *Medicago truncatula*," *Molecular Plant*. 15(12), 1852-1867. DOI: 10.1016/j.molp.2022.10.019

Yu, L., Zhang, T., Zhu, Y., Zhou, Y., Xia, T., and Nie, Y. (2018). "Estimation of SPAD values in soybean leaves using IRIV algorithm for wavelength variable selection based on hyperspectral data," *Transactions of the Chinese Society of Agricultural Engineering* 34(16), 7. DOI: 10.11975/j.issn.1002-6819.2018.16.019

Rego, G., Ferrero, F., Valledor, M., Campo, J. C., Forcada, S., Royo, L. J., and Soldado, A. (2020). "A portable IoT NIR spectroscopic system to analyze the quality of dairy farm forage," *Computers and Electronics in Agriculture*, 175, 105578. DOI: 10.1016/j.compag.2020.105578

Zhang, J., Guo, Z., Ren, Z., Wang, S., Yue, M., Zhang, S., Yin, X., Du, J., and Ma, C. (2023). "Variable selection methods to determine protein content in paddy using near-infrared hyperspectral imaging," *Journal of Food Measurement and Characterization* 17, 4506-4519. DOI: 10.1007/s11694-023-01964-y

Zhang, M. S., Zhang, B., Li, H., Shen, M. S., Tian, S. J., Zhang, H. H., Ren, X. L., Xing, L. B., and Zhao, J. (2020). "Determination of bagged 'Fuji' apple maturity by visible and near-infrared spectroscopy combined with a machine learning algorithm," *Infrared Physics & Technology* 111. DOI: 10.1016/j.infrared.2020.103529

Article submitted: January 8, 2024; Peer review completed: March 9, 2024; Revised version received: March 24, 2024; Accepted: March 30, 2024; Published: April 26, 2024.
DOI: 10.15376/biores.19.2.3808-3825