

# Near-infrared Spectroscopy Detection of Rice Protein Content Based on Stacking Multi-model Fusion

Shengye WANG<sup>1</sup>, Siting WU<sup>1</sup>, Jinming LIU<sup>1\*</sup>, Chunqi WANG<sup>2</sup>, Zhijiang LI<sup>2</sup>

1. College of Information and Electrical Engineering, Heilongjiang Bayi Agricultural University, Daqing 163319, China; 2. College of Food Science, Heilongjiang Bayi Agricultural University, Daqing163319, China

**Abstract** [ **Objectives** ] This study was conducted to achieve rapid and accurate detection of protein content in rice with a particle size of 1.0 mm. [ **Methods** ] A multi-model fusion strategy was proposed on the basis of Stacking ensemble learning. A base learner pool was constructed, containing Partial Least Squares (PLS), Support Vector Machine (SVM), Deep Extreme Learning Machine (DELM), Random Forest (RF), Gradient Boosting Decision Tree (GBDT), and Multilayer Perceptron (MLP). PLS, DELM, and Linear Regression (LR) were used as meta-learner candidates. Employing integer coding technology, systematic dynamic combinations of base learners and meta-learners were generated, resulting in a total of 40 non-repetitive fusion models. The optimal combination was selected through a comprehensive evaluation based on multiple assessment indicators. [ **Results** ] The combination "PLS-DELM-MLP-LR" (code 1367) achieved coefficients of determination of 0.973 2 and 0.978 0 on the validation set and independent test set, respectively, with relative root mean square errors of 2.35% and 2.36%, and residual predictive deviations of 6.107 5 and 6.747 9, respectively. [ **Conclusions** ] The Stacking fusion model significantly enhances the predictive accuracy and robustness of spectral quantitative analysis, providing an efficient and feasible solution for modeling complex agricultural product spectral data.

**Key words** Rice protein; Near-infrared spectroscopy; Stacking ensemble learning; Multi-model fusion; Integer encoding

DOI:10.19759/j.cnki.2164-4993.2026.01.010

Rice is a staple food for over half of the world's population. Its nutritional quality, particularly protein content, is one of the key indicators for assessing its edible and market value<sup>[1]</sup>. Although traditional chemical detection methods yield accurate results, they involve complex procedures, are time- and labor-intensive, destroy samples, and pose risks of chemical reagent pollution, making it difficult to meet the demands of modern agricultural products for rapid, non-destructive, and online quality detection<sup>[2]</sup>. Near-infrared spectroscopy (NIRS) analysis technology has been widely applied in the field of agricultural product quality detection due to its advantages of high speed, non-destructiveness, and environmental friendliness<sup>[3]</sup>.

However, spectral data typically exhibit characteristics such as high dimensionality, nonlinearity, and multicollinearity, while single models often have limitations and struggle to balance accuracy with generalization capability. Currently, modeling approaches for spectral data are primarily divided into linear and nonlinear categories. Linear models such as PLS can handle multicollinearity, but are inadequate for capturing complex relationships<sup>[4]</sup>. Nonlinear models such as SVM, RF, GBDT, and neural networks can characterize nonlinear features, but are prone to overfitting<sup>[5]</sup>. To overcome the limitations of single models, ensemble learning enhances overall performance by integrating predictions from multiple models, aiming to achieve superior generalization capability compared with single-component learners. Among

these methods, Stacking is an advanced ensemble strategy that employs a meta-learner to learn and combine the primary predictions of multiple base learners, thereby forming a hierarchical model fusion framework<sup>[6]</sup>.

Based on rice samples with a 1.0 mm grinding particle size, this study was conducted to transcend traditional single-model approaches by innovatively integrating multiple representative linear and nonlinear algorithms into a unified Stacking framework. To systematically explore the optimal model structure, integer encoding technology was introduced for standardized identification and dynamic combination screening of the learner pool. Through comprehensive evaluation, a Stacking fusion model that combines high predictive accuracy, strong generalization capability and practical efficiency was constructed. This study provides reliable technical support for the rapid and precise detection of rice protein content, while offering a novel solution for multi-model fusion modeling in handling complex spectral data.

## Materials and Methods

### Samples and spectral collection

In This study, 150 representative rice varieties were selected from high-quality rice cultivation areas in Northeast China. Sample pretreatment involved using a THU35C hulling device for three rounds of dehulling, followed by two rounds of polishing with a TM05C rice polishing machine. The resulting milled rice samples were packaged in No.8 ziplock bags. They were pulverized using an FZ-102 miniature plant sample grinder and sieved through a standard sieve to accurately obtain samples with a particle size of 1.0 mm. The crushed samples were packaged in No.3 ziplock bags and stored in a dry and ventilated environment. Spectral data

Received: September 3, 2025 Accepted: November 6, 2025

Shengye WANG (2001 –), male, P. R. China, devoted to research about intelligent analysis and processing of multi-source data.

\* Corresponding author.

collection was performed using a TANGO Fourier transform near-infrared spectrometer equipped with an integrating sphere diffuse reflection detection module. The wavenumber range covered 3 946–11 542  $\text{cm}^{-1}$ , and the spectral resolution was set at 8.0  $\text{cm}^{-1}$ . The data recording mode was selected as absorbance. To minimize the impact of random noise, each sample was scanned three times, and arithmetic means were calculated to generate a standard spectral dataset. Simultaneously, the Kjeldahl method was applied to accurately determine the protein content of each sample, serving as the reference value for model establishment and validation.

### Spectral preprocessing and dataset division

To eliminate interference from baseline drift, light scattering, and noise, necessary preprocessing was applied to the raw spectral data. In this study, we compared the effects of multiple preprocessing methods and selected the optimal combination: Savitzky-Golay filters (SG) and baseline correction (BC)<sup>[7]</sup>. Under the Monte Carlo Cross-Validation (MCCV) framework, an analysis of residual statistical distribution characteristics was employed for outlier detection. Samples with significantly high residuals deviating from the main data cluster were identified as outliers and removed. Subsequently, the random selection method was used to extract samples from the original dataset to form an initial modeling subset, while a small number of unselected samples served as an external test set to evaluate the generalization capability of the model. Following this, based on the spatial distance maximization principle of the Kennard – Stone method, highly representative samples from the initial modeling subset were sequentially selected into the calibration set, while a validation subset was simultaneously constructed for model performance assessment<sup>[8]</sup>.

### Optimal wavelength selection

The raw near-infrared spectral data of rice samples often contain environmental noise and non-target signals. The Uninformative Variable Elimination (UVE) method, based on the principles of Partial Least Squares (PLS) Regression, introduces virtual noise channels to establish variable stability criteria, enabling dimensionality reduction and feature selection for high-dimensional spectral data<sup>[9]</sup>. First, PLS regression is performed on the original spectral matrix  $X_{(m \times n)}$  and the target parameter matrix  $Y_{(n \times 1)}$  to determine the optimal number of latent variables. Subsequently, a random noise matrix  $N_{(m \times n)}$  equivalent to the instrument error level is generated and horizontally concatenated with the original spectral matrix to form an extended matrix  $X_{N(n \times 2m)}$ . The first  $m$  columns represent the real spectral variables, while the latter  $m$  columns serve as virtual noise channels. Full-spectrum regression modeling is performed on  $X_N$  and  $Y$ , and the regression coefficient matrix  $B_{(1 \times 2m)}$  is obtained using leave-one-out cross-validation. The mean and standard deviation of the regression coefficients for each column are calculated, and a stability indicator is constructed.

$$MEAN_i = \frac{1}{k} \sum_{j=1}^k b_{ij}, S_i = \sqrt{\frac{1}{k-1} \sum_{j=1}^k (b_{ij} - MEAN_i)^2} \quad (1)$$

$$H_i = \frac{|MEAN_i|}{S_i} \quad (2)$$

The maximum value of  $H_i$  from the virtual noise channels (columns  $m+1$  to  $2m$ ) is taken as the screening threshold  $H_{\max}$ . When  $H_i < H_{\max}$ , the corresponding spectral variables are eliminated, retaining effective characteristic wavelengths to construct the optimized spectral matrix  $X_{\text{uve}}$ .

### Encoding strategy and stacking framework

This study selected six algorithms with excellent performance in spectral modeling to construct a base learner pool, each assigned an integer code: PLS(1), SVM(2), DELM(3), RF(4), GBDT(5), MLP(6). The meta-learner is responsible for the final integration of the outputs from the base learners: PLS(1), DELM(3), LR(7). To systematically explore the performance of different combinations, integer encoding technology was employed to uniquely identify each model combination. Three out of the six base learners are selected for combination, and one out of the three meta-learners is chosen. The base learner codes are arranged in the order of the selected model numbers, while the meta-learner code is listed separately. For example, "1237" indicates that the base learners are PLS(1), SVM(2), and DELM(3), and the meta-learner is LR(7). According to the principle of combinatorial mathematics, selecting 3 out of 6 base learners [ $C(6,3) = 20$  combinations] and pairing them with one of the 3 meta-learners generates 60 potential combinations. However, under the constraint that "base and meta-learners cannot be duplicated", 40 valid combinations are retained after eliminating redundancies.

This study employs a two-layer Stacking framework. First layer (base learner layer): Multiple base learners are trained on the calibration set using 5-fold cross-validation, generating their predictions (*i.e.*, meta-features) for the calibration set, validation set, and independent test set. Second layer (meta-learner layer): The meta-features of the calibration set generated in the first layer are used as new input features, with the corresponding reference protein content as the output, to train the meta-learner. Finally, the trained meta-learner is applied to predict the meta-features of the validation set, yielding the final results of the Stacking model<sup>[10]</sup>.

### Model evaluation and optimal combination selection

To comprehensively and fairly evaluate the performance of each combination, multiple statistical indicators were used for assessment: the coefficient of determination ( $R^2$ ), root mean square error (RMSE), relative root mean square error (rRMSE), and residual predictive deviation (RPD).  $R^2$  is used to measure the model's explanatory power for the calibration set, validation set, and independent test set data. RMSE is used to measure the deviation between predicted and actual values, reflecting the predictive accuracy of the model. rRMSE is the ratio of RMSE to the mean, effectively eliminating the influence of mean or dimensional

differences on RMSE. Combinations with high  $R^2$ , comparable RMSE, rRMSE below 5%, RPD greater than 3, and acceptable operational efficiency are prioritized as the optimal Stacking fusion model<sup>[11]</sup>.

## Results and Analysis

### Performance comparison of single models

First, the performance of the six base learners as single models under UVE-optimized wavelengths was evaluated, as shown in Table 1.

The results indicated significant differences in the predictive capabilities of various algorithms for rice protein content. Nonlinear models demonstrated excellent fitting performance on the calibration set, but their  $R^2$  values showed considerable fluctuation on the validation set. The linear model PLS exhibited stable performance, though its predictive accuracy ceiling may be lower than some high-performing nonlinear models. These findings confirm the limitations of single models and underscore the necessity of leveraging ensemble methods to combine their strengths.

**Table 1 Performance comparison of linear and nonlinear single models under UVE optimal wavelength**

1.0 mm UVE Single model	$R_c^2$	$R_v^2$	$R_t^2$	RMSEC	RMSEV	RMSET	rRMSEC	rRMSEV	rRMSET	RPDV	RPDT
RF	0.908 6	0.698 2	0.751 1	0.269 9	0.536 8	0.552 5	0.040 0	0.078 9	0.079 3	1.851 4	2.065 6
PLS	0.966 2	0.965 9	0.963 4	0.164 1	0.180 4	0.211 7	0.024 3	0.026 5	0.030 4	5.416 4	5.229 1
SVM	0.972 7	0.967 5	0.970 9	0.147 4	0.176 3	0.188 9	0.021 8	0.025 9	0.027 1	5.543 4	5.862 2
DELM	0.966 4	0.969 7	0.964 1	0.163 5	0.170 2	0.209 9	0.024 2	0.025 0	0.030 1	5.838 3	5.437 0
GBDT	0.993 2	0.829 5	0.837 5	0.073 4	0.403 5	0.446 5	0.010 9	0.059 3	0.064 1	2.462 9	2.555 7
MLP	0.956 5	0.917 8	0.931 1	0.186 0	0.280 1	0.290 6	0.027 6	0.041 2	0.041 7	3.548 0	3.927 0

### Performance screening of multi-model stacking fusion combinations

A systematic evaluation was conducted on 40 valid Stacking combinations. Most Stacking fusion models outperformed the best single model, demonstrating the effectiveness of the ensemble

strategy. A gradient distribution in performance was observed among different combinations, indicating that the selection of base learners and their compatibility with the meta-learner are critical to the final performance.

**Table 2 Comparison of Stacking model performance when choosing PLS as a meta learner**

1.0 mm UVE Stacking model	Length (best)	$R_c^2$	$R_v^2$	$R_t^2$	RMSEC	RMSEV	RMSET	rRMSEC	rRMSEV	rRMSET	RPDV	RPDT
2341	699	0.962 6	0.968 5	0.979 0	0.172 5	0.173 5	0.160 6	0.025 6	0.025 5	0.023 0	5.630 7	6.895 4
2351	699	0.962 6	0.969 3	0.978 2	0.172 7	0.171 1	0.163 4	0.025 6	0.025 2	0.023 5	5.711 0	6.774 4
2361	699	0.962 5	0.969 3	0.978 6	0.172 8	0.171 1	0.161 9	0.025 6	0.025 2	0.023 2	5.710 7	6.839 3
2451	699	0.930 3	0.947 6	0.943 2	0.235 7	0.223 7	0.263 9	0.034 9	0.032 9	0.037 9	4.367 1	4.195 6
2461	699	0.937 4	0.946 6	0.963 4	0.223 3	0.225 8	0.211 9	0.033 1	0.033 2	0.030 4	4.326 7	5.225 8
2561	699	0.935 2	0.952 2	0.952 6	0.227 3	0.213 5	0.240 9	0.033 7	0.031 4	0.034 6	4.576 1	4.595 1
3451	699	0.962 2	0.970 7	0.964 3	0.173 6	0.167 3	0.209 3	0.025 7	0.024 6	0.030 0	5.841 5	5.290 1
3461	699	0.962 0	0.970 8	0.963 0	0.174 1	0.166 9	0.212 9	0.025 8	0.024 5	0.030 6	5.854 1	5.201 1
3561	699	0.961 9	0.970 5	0.964 5	0.174 2	0.167 8	0.208 6	0.025 8	0.024 7	0.029 9	5.822 6	5.307 0
4561	699	0.900 8	0.943 4	0.870 0	0.281 1	0.232 5	0.399 2	0.041 6	0.034 2	0.057 3	4.203 0	2.773 6

**Table 3 Comparison of Stacking model performance when choosing DELM as a meta learner**

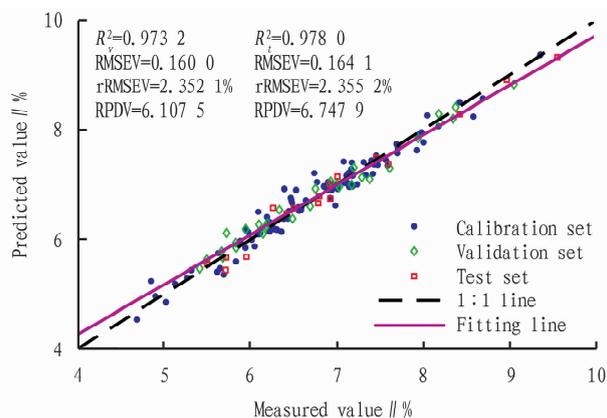
1.0 mm UVE Stacking model	Length (best)	$R_c^2$	$R_v^2$	$R_t^2$	RMSEC	RMSEV	RMSET	rRMSEC	rRMSEV	rRMSET	RPDV	RPDT
1243	699	0.861 8	0.822 4	0.862 1	0.331 8	0.411 8	0.411 1	0.049 2	0.060 5	0.059 0	2.372 8	2.692 8
1253	699	0.893 0	0.863 1	0.901 8	0.292 0	0.361 5	0.347 0	0.043 3	0.053 2	0.049 8	2.702 7	3.190 4
1263	699	0.939 3	0.957 1	0.946 5	0.219 9	0.202 5	0.256 1	0.032 6	0.029 8	0.036 8	4.825 8	4.323 6
1453	699	0.955 0	0.967 0	0.967 1	0.189 3	0.177 6	0.200 9	0.028 0	0.026 1	0.028 8	5.500 9	5.509 3
1463	699	0.854 9	0.865 7	0.847 3	0.339 9	0.358 0	0.432 6	0.050 4	0.052 6	0.062 1	2.729 1	2.559 4
1563	699	0.955 1	0.968 1	0.967 2	0.189 1	0.174 5	0.200 6	0.028 0	0.025 7	0.028 8	5.600 4	5.520 2
2453	699	0.924 0	0.939 5	0.938 9	0.246 0	0.240 4	0.273 7	0.036 5	0.035 3	0.039 3	4.065 3	4.045 3
2463	699	0.938 1	0.947 0	0.958 3	0.222 1	0.225 0	0.226 1	0.032 9	0.033 1	0.032 5	4.341 9	4.896 6
2563	699	0.926 6	0.939 4	0.945 7	0.241 8	0.240 5	0.257 9	0.035 8	0.035 4	0.037 0	4.062 8	4.292 5
4563	699	0.848 7	0.832 6	0.832 9	0.347 2	0.399 7	0.452 6	0.051 4	0.058 8	0.065 0	2.444 4	2.446 3

**Table 4** Comparison of Stacking model performance when choosing LR as a meta learner

1.0 mmUVE Stacking model	Length (best)	$R_c^2$	$R_v^2$	$R_t^2$	RMSEC	RMSEV	RMSET	rRMSEC	rRMSEV	rRMSET	RPDV	RPDT
1237	699	0.960 3	0.970 9	0.970 3	0.177 8	0.166 8	0.190 8	0.026 3	0.024 5	0.027 4	5.858 5	5.803 5
1247	699	0.958 1	0.968 2	0.964 8	0.182 7	0.174 3	0.207 6	0.027 1	0.025 6	0.029 8	5.604 4	5.333 1
1257	699	0.959 8	0.966 3	0.962 9	0.179 0	0.179 5	0.213 2	0.026 5	0.026 4	0.030 6	5.444 7	5.1917
1267	699	0.959 4	0.967 3	0.963 9	0.179 8	0.176 7	0.210 4	0.026 6	0.026 0	0.030 2	5.530 0	5.261 3
1347	699	0.957 7	0.970 0	0.968 0	0.183 6	0.169 3	0.197 9	0.027 2	0.024 9	0.028 4	5.769 9	5.592 9
1357	699	0.958 4	0.966 9	0.968 5	0.182 1	0.177 9	0.196 6	0.027 0	0.026 2	0.028 2	5.492 7	5.631 2
1367	699	0.959 0	0.973 2	0.978 0	0.180 8	0.160 0	0.164 1	0.026 8	0.023 5	0.023 6	6.107 5	6.747 9
1457	699	0.955 8	0.969 3	0.967 6	0.187 7	0.171 2	0.199 1	0.027 8	0.025 2	0.028 6	5.706 8	5.559 2
1467	699	0.955 8	0.969 5	0.967 4	0.187 7	0.170 5	0.199 8	0.027 8	0.025 1	0.028 7	5.729 5	5.540 0
1567	699	0.957 3	0.969 0	0.966 6	0.184 5	0.172 1	0.202 4	0.027 3	0.025 3	0.029 1	5.677 0	5.469 3
2347	699	0.962 8	0.968 3	0.979 4	0.172 2	0.174 0	0.158 9	0.025 5	0.025 6	0.022 8	5.613 9	6.966 2
2357	699	0.963 7	0.970 7	0.976 3	0.169 9	0.167 3	0.170 5	0.025 2	0.024 6	0.024 5	5.839 7	6.494 6
2367	699	0.963 6	0.966 5	0.978 9	0.170 2	0.179 0	0.160 8	0.025 2	0.026 3	0.023 1	5.460 1	6.885 5
2457	699	0.931 3	0.945 9	0.943 7	0.234 0	0.227 2	0.262 7	0.034 7	0.033 4	0.037 7	4.300 1	4.213 8
2467	699	0.939 7	0.950 1	0.957 7	0.219 1	0.218 3	0.227 6	0.032 5	0.032 1	0.032 7	4.475 6	4.863 7
2567	699	0.935 7	0.951 8	0.949 9	0.226 2	0.214 6	0.247 9	0.033 5	0.031 5	0.035 6	4.553 3	4.466 5
3457	699	0.962 4	0.968 8	0.963 6	0.173 0	0.172 5	0.211 1	0.025 6	0.025 4	0.030 3	5.665 4	5.244 4
3467	699	0.963 6	0.970 7	0.956 0	0.170 4	0.167 2	0.232 3	0.025 2	0.024 6	0.033 3	5.844 4	4.766 2
3567	699	0.962 8	0.971 3	0.958 1	0.172 1	0.165 6	0.226 6	0.025 5	0.024 3	0.032 5	5.901 9	4.886 6
4567	699	0.903 5	0.937 6	0.854 6	0.277 2	0.244 0	0.422 2	0.041 1	0.035 9	0.060 6	4.003 7	2.622 2

Through performance ranking, the combination coded "1367" was found to exhibit the best overall performance in multiple indicators. This combination achieved the highest  $R^2$  and the lowest RMSE on both the validation set and the independent test set, and its RPD value was significantly greater than 3, indicating exceptionally strong model generalization without overfitting. These results further confirm the model's outstanding predictive capability.

weighting on the outputs of the three highly diverse base learners. Compared with single models, this Stacking model significantly reduces RMSE, with predictions closer to the measured values. Its structure is concise and allows for parallel training, making it suitable for rapid detection systems. Fig. 1 shows that both the validation set and independent test set data points are distributed along the 1 : 1 line, with  $R^2$  exceeding 0.97, RMSE around 0.16, and significantly smaller prediction errors, and the number of outliers is reduced, demonstrating excellent prediction consistency, which meets practical detection requirements<sup>[12]</sup>.

**Fig. 1** Scatter plot of measured and predicted values in the optimal stacking combination

### Analysis of the optimal Stacking model

The optimal combination "PLS – DELM – MLP – LR" skillfully integrates the linear stability and feature extraction capability of PLS, the deep nonlinear feature learning ability of DELM, and the powerful function approximation capacity of MLP. The meta-learner employs a simple LR, which, though not introducing additional complex nonlinearity, effectively performs optimal linear

## Conclusions and Discussion

The multi-model fusion method based on Stacking proposed in this study successfully achieved near-infrared spectral detection of rice protein content with a 1.0 mm grinding particle size. By constructing a diverse learner pool comprising both linear and non-linear algorithms and innovatively adopting integer encoding technology to systematically guide model combination and screening, the optimal Stacking fusion model (coded 1367) was ultimately established, with PLS, DELM, and MLP as base learners and LR as the meta-learner.

The model fully integrates the advantages of different algorithms, demonstrating high accuracy and stability on both the validation set and the independent test set. It significantly enhances the predictive accuracy, robustness, and generalization capability for rice protein content, overcoming the potential limitations of single models. This study not only provides a practical and effective solution for the rapid, non-destructive, and high-precision detection of rice protein content, but also offers an innovative "multi-model fusion modeling" approach and reference for rapid quality detection of agricultural products and complex spectral

