Analytical Methods

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d4ay01202j

Exploring soil multi-parameter stacking measurement through Raman and NIR dual-spectroscopy†

Qiong Sang,[a] Xiaoyu Zhao,[*a] Yue Zhao,[a] Lijing Cai,[a] Jinming Liu,[a] Liang Tong[b] and Zhe Zhai[*c]

The excessive use of fertilizers can lead to increased production costs, degraded soil quality, diminished product excellence, and environmental contamination. To address this issue, a solution involving soil testing and customizing fertilizer application has been proposed. The current standard methodology for soil parameter assessment relies on chemical analysis performed by trained laboratory technicians, which only allows for the measurement of one indicator at a time. Hence, a novel approach utilizing the fusion of near-infrared (NIR) and Raman dual-spectral features has been suggested to simultaneously determine five crucial indicators (hydrolyzed N, available P, quick-release K, OM, and pH) in soil with a single scan. In this research, seven preprocessing techniques and four feature extraction methods were initially explored to optimize the composite NIR and Raman feature variables. Subsequently, a regressor with a two-layer network structure (RF, LR, SVR; ELM, and PLS) was developed using the stacking algorithm. This methodology synergizes the strengths of the five base learners, minimizes the risk of overfitting, and demonstrates high computational efficiency for linear data correlations and robust fitting capabilities for nonlinear data correlations. Additionally, it showcases strong generalization capabilities, noise resilience, and robustness. The model produced relevant results for hydrolyzed N, available P, quick release K, OM, and pH measurements, with R_p^2 values of 0.9966, 0.9722, 0.9855, 0.9557, and 0.9951, RMSEP values of 2.9547, 2.9972, 7.6550, 0.0765, and 0.0313, and RPD values of 6.0855, 2.4655, 3.0511, 8.3084, and 10.6977. This work delivers a twofold contribution by presenting a swift method for simultaneous measurement of multiple soil parameters, enabling concurrent ploughing, soil surveying, and fertilizer application. Furthermore, it introduces a stacking measurement model based on dual fusion features, showcasing detailed model parameters. The stacking model outperformed mono-spectral models (NIR and Raman) and the dual PLS model in terms of R_p^2, RPD, and RMSEP values, and fluctuation ranges, demonstrating enhanced stability, predictive prowess, and reliable observations. Overall, the stacking model offers a cost-effective, rapid, and precise solution for online evaluation of soil physicochemical conditions, catering to the requirements of modern agricultural production well. This innovative approach signifies a significant leap forward and provides a solid theoretical foundation for the enhancement of associated online monitoring systems and tools.

Received 27th June 2024
Accepted 2nd September 2024

DOI: 10.1039/d4ay01202j

rsc.li/methods

1 Introduction

In modern agricultural production, two prominent challenges involve the disparity between soil demand and fertilizer supply, alongside the improper application of fertilizers.[1] Misguided use of fertilizers can result in a range of adverse effects including stunted growth, disease outbreaks, pest invasions, soil compaction, and contamination of soil, water sources, and air. It has been observed that tailoring fertilizer application to soil conditions proves highly effective. Common soil assessments typically encompass tests for hydrolyzed N, available P, quick-release K, organic matter content, and pH levels. Historically, chemical techniques have been utilized to gauge these five soil parameters. A standard protocol in the agricultural sector of the People's Republic of China[2] (https://www.docin.com/p-60500571.html) involves methodologies such as boiling samples with sulphuric acid and hydrogen peroxide to quantify hydrolyzed N, digesting samples with sulfuric acid and hydrogen peroxide for available P, subjecting samples to a similar boiling process for quick-release K, and employing a potassium dichromate solution to estimate organic matter content. pH levels are typically determined using various

[a]Heilongjiang Bayi Agricultural University, China. E-mail: xy_zhao77@163.com
[b]Qiqihar University, China
[c]Chinese Academy of Forestry Sciences, China

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d4ay01202j

extraction agents and potentiometric analysis.[3] Chemical analyses, while precise, come with drawbacks such as the need for specific chemicals, high operational costs, environmental impact, expert handling requirements, time-intensive procedures, and limited capacity for simultaneous measurements. While suitable for detailed research, chemical methods fall short in meeting the demands of swift on-site assessments where measurements must align with fertilizer applications. This has spurred ongoing efforts towards developing quicker, cost-efficient, and user-friendly alternatives like NIR and Raman spectroscopy. ESI, Table 1† provides a summary of measurement statistics for the five soil indicators utilizing NIR and Raman techniques from both domestic and international research studies.

In ESI, Table 1,† the results highlighted with a dark gray background represent the superior research outcomes based on NIR and Raman technologies, while those highlighted with a light gray background indicate the superior Raman or NIR measurement models. These reviews of national and international studies highlight three points. Firstly, among the reported explorations of soil indicator measurements based on NIR and Raman spectroscopy techniques, most of them used NIR spectroscopic techniques, and there are more studies on the measurement of nitrogen and organic matter content of soils, and fewer studies on the measurements of other indicators,[4] such as effective phosphorus, quick release potassium, and pH, and simultaneous measurements of all five indicators of soils have not been found.[5–7] Secondly, Raman spectroscopy technology has not yet been used to measure potassium content and pH values in soil.[8–10] Finally, the accuracy of models for measuring hydrolyzed N, available P, quick-release K, OM and pH was generally poor, except for OM measurements based on NIR techniques.

Therefore, the study aims to combine Raman and NIR dual spectral information to first accurately measure essential soil parameters for fertilization (hydrolyzed N, available P, quick-release K, OM, and pH) and secondly to improve the precision of measuring the five parameters.

Raman spectroscopy and NIR spectroscopy fall under the same category of molecular vibrational spectroscopy. Raman spectroscopy[11] is the scattering effect of molecules on incident laser light, which carries information about the vibration and rotation of molecules, while NIR spectroscopy[12] is the absorption effect of molecules on excitation light, which carries information about the hydrogen-containing groups of organic molecules. Typically, asymmetric vibrations of molecules and vibrations of polar groups trigger changes in the dipole moment of the molecule, and such molecular vibrations are manifested as NIR activity. Molecular symmetry vibrations and vibrations of non-polar groups deform the molecule, leading to changes in the polarisation rate, and such molecular vibrations are manifested as Raman activity. It can be seen that the Raman spectra are suitable for the non-polar bonding vibrations of the same atoms, such as C–C, S–S, and N–N bonds and other symmetric backbone vibrations, can be obtained from the Raman spectra with abundant information, while the polar bonding of different atoms, such as C═O, C–H, N–H, O–H, *etc.*, is presented in the NIR spectra, and the symmetric backbone vibrations of the molecules are almost invisible in the NIR spectra.[13–16] In addition, with the rapid development of the level of manufacturing, the price of portable NIR spectrometers and portable Raman spectrometers has been reduced year by year. Therefore, the method of combining the use of NIR and Raman spectroscopy to improve the accuracy of soil parameter measurements has a scientific basis and practical application feasibility.

2 Material and spectrum acquisition

2.1 Introduction to spectroscopic instruments

Spectroscopic instruments were utilized in the research, with one notable device being the Advantage 532 Raman spectrometer manufactured by DeltaNu in the United States. This particular instrument boasts a Raman wavelength range spanning from 200 to 3400 cm^{-1}, operating with an excitation wavelength of 532 nm. It can deliver a maximum excitation power of 100 mW and conduct scans lasting up to 60 seconds. A visual depiction of the spectrometer can be observed in ESI, Fig. 1(a).† Another pivotal instrument employed was the TANGO NIR spectrometer produced by Bruker in Germany. It covers a spectral range of 11 520–4000 cm^{-1} with a resolution of 8 cm^{-1} and conducts 32 scans. This spectrometer is displayed in ESI, Fig. 1(b).†

2.2 Soil sample acquisition

The investigation involved analyzing 95 soil plots situated in Mingshui County, Suihua City, Heilongjiang Province, China. The soil samples were collected from Mingshui County positioned at 125.90° E longitude and 47.18° N latitude. Descriptions of the soil include it being black, featuring medium levels of organic matter and total nitrogen, high in quick-release nitrogen, low in quick-release phosphorus, and rich in total potassium. The soil samples were gathered in March 2023, and subsequently dried, milled, sieved, and packed as illustrated in ESI, Fig. 2.† The Heilongjiang Beifeng Agricultural Capital Group employed the national standard method to measure hydrolyzed N, available P, quick-release K, organic matter (OM), and pH in the samples, with statistical values for these parameters shown in ESI, Table 2.† The following link provides a detailed description of the operating procedure for the national standard method (**https://www.docin.com/p-60500571.html**).

2.3 Spectrum acquisition

In preparation for spectroscopic analysis, soil samples were enclosed in cylindrical quartz cups with a 4 cm diameter at the cup base and a 5 mm height allocated for the samples. These cups were then positioned on a Raman spectroscopy carrier stage in a dimly lit room, with the laser port located 50 mm away from the stage. By averaging five spectra to portray the Raman spectra of each sample, a total of 95 Raman spectra were recorded, as indicated in Fig. 1(a). Similarly, the soil was packed in a 4.6 cm diameter cylindrical quartz cup at the base with

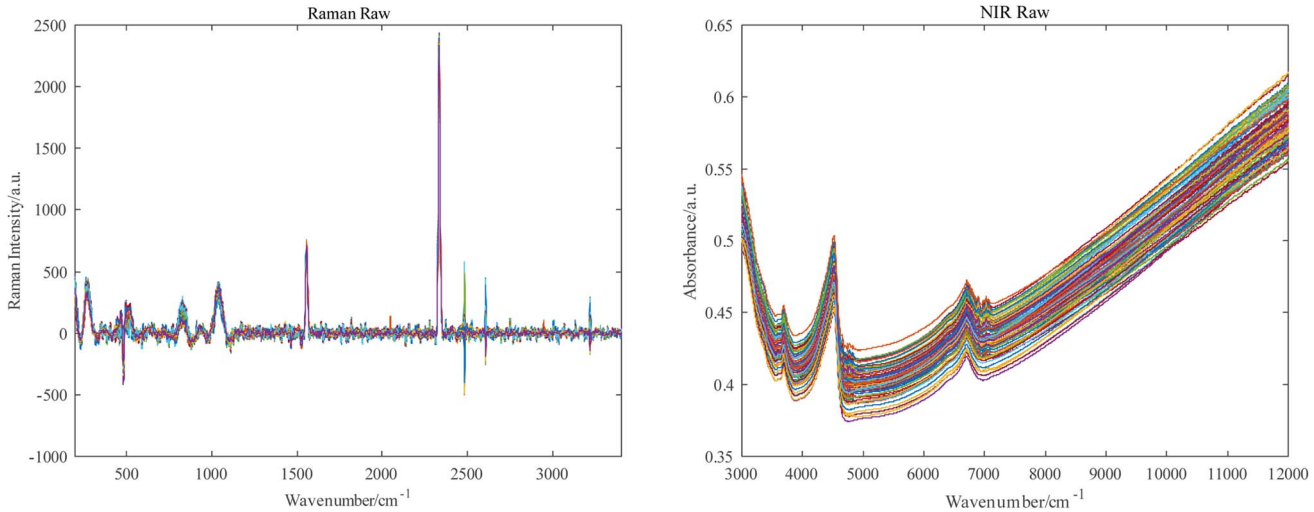

Fig. 1 Raw Raman spectra (a) and NIR spectra (b).

a 5 mm sample height for NIR spectroscopy. This cup was placed at the discharge outlet of the NIR spectrometer, and an average of five spectra was used to depict the NIR spectra, resulting in 95 NIR spectra showcased in Fig. 1(b). Among these 95 combined Raman and NIR spectra, 66 were designated for calibrating the soil measurement model, while the remaining 29 were set aside for the model's validation.

3 Methods and evaluation indicators

3.1 Theory and methodology

Due to various external factors such as human intervention, differences in environmental conditions, and instrumental limitations, as well as internal factors like sample characteristics, the original spectral data often contain significant amounts of irrelevant information such as noise, baseline drift, and stray light.[17] As a result, several preprocessing steps are typically applied to the spectra to enhance the accuracy of measurements. Initially, the spectra undergo Savitzky–Golay (SG) smoothing denoising to reduce electrical, optical, and mechanical noise. Subsequently, a 1st derivative operation is performed to remove baseline drift and accentuate characteristic spectral features. Moreover, Standard Normal Variation (SNV) correction and Multiple Scattering Correction (MSC) are employed to mitigate the impact of soil particle size, surface scattering, and light-range transformations on the spectra. Combinations of these techniques, including SG smoothing with Multivariate Scattering Correction (SG + MSC), SG smoothing with Standard Normal Variable transformation (SG + SNV), and SG smoothing with first-order derivative (SG + 1st derivative), are applied, resulting in seven preprocessing variations.[18]

The preprocessed spectra exhibit substantial data redundancy and covariance, necessitating effective feature extraction techniques to reduce dimensionality, minimize modeling requirements, and enhance operational efficiency. In this study, Competitive Adaptive Re-weighted Sampling Algorithm (CARS), Successive Projection Algorithm (SPA), Uninformative Variable Exclusion Algorithm (UVE), and Principal Component Analysis (PCA) methods are employed for spectral data feature extraction. CARS identifies variables closely associated with the measured component by evaluating the absolute value of the regression coefficient.[19] SPA iteratively selects wavelengths to minimize redundancy and covariance in spectral information.[20] UVE identifies and removes variables that do not contribute to prediction accuracy, thus enhancing model simplicity and stability.[21] PCA transforms data into linearly independent combinations, facilitating dimensionality reduction.[22]

Partial Least Squares Regression (PLSR) is selected as the modeling technique based on optimized preprocessing and feature-extracted data. PLSR projects independent and dependent variables into a new space to maximize covariance and perform regression analysis, making it widely utilized in contemporary research.[23]

The Extreme Learning Machine (ELM), proposed by Huang *et al.*, is a fast-learning algorithm for single hidden layer feed forward neural networks. ELM's random initialization of input weights and hidden layer biases, along with the use of the generalized inverse of output weights, ensures rapid learning and strong generalization. While ELM overcomes the limitations of traditional nonlinear techniques, its single model may exhibit instability. Therefore, integrated modeling, such as stacking, is employed to improve stability and performance by combining predictions from multiple models.[24]

Stacking, a classical integrated learning algorithm, enhances overall performance by aggregating predictions from multiple base models. By optimizing the ELM model combined with PLS using the stacking method, the study maximizes the advantages of both techniques, improving model accuracy, stability, and generalization across diverse data structures and problem domains.[25]

3.2 Evaluation indicators

Indices such as the coefficient of determination (R^2), root mean square error (RMSE), and relative analytical error (RPD) were used to compare the evaluation models.

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

In eqn (1), R^2 represents the coefficient of determination, $\hat{y}_i$ denotes the predicted value of the i-th sample, y_i stands for the reference value of the i-th sample, $\bar{y}$ is the mean value of the reference values of the samples, and n signifies the sample size.

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

In eqn (2) RMSE is the root mean square error, where $\hat{y}_i$ represents the predicted value of the i-th sample, y_i is the reference value of the i-th sample, $\bar{y}$ stands for the mean value of the reference values of the samples, and n denotes the sample size.

$$\mathrm{RPD} = \frac{\mathrm{SD}}{\mathrm{RMSE}} \quad (3)$$

In eqn (3), RPD stands for the relative analytical error, while SD denotes the standard deviation of the parameter values of the modeling or prediction set, and RMSE represents the root mean square error.

Within the regression model, R^2 serves to reflect the explanatory power of the model on the sample spectrum. A value approaching 1 indicates a stronger fit between the predicted and true values. RMSE, on the other hand, reveals the prediction bias of the model. A smaller value suggests a more stable and robust predictive ability. RPD provides insight into the overall predictive capability of the model, with larger values indicating superior performance. Specifically, an RPD value of $\geq$2.0 suggests that the model is suitable for practical measurement of the five indicators; a range between 1.0 and 2.0 implies poor predictive performance; and an RPD value of $\leq$1.0 signifies very limited predictive capability.[26]

4 Measurement models based on bispectral features

To achieve the optimal data format for constructing dual spectra of Raman and NIR, the subsequent step involves establishing measurement models based on individual spectra of Raman and NIR.

4.1 Screening of preprocessing methods for individual Raman and NIR data

The Raman and NIR raw data underwent seven kinds of preprocessing treatments, including Savitzky–Golay, MSC, SNV, 1st derivative, Savitzky–Golay + MSC, Savitzky–Golay + SNV, and Savitzky–Golay + 1st derivative. ESI, Fig. 3† illustrates the impact of these preprocessing methods.

The preprocessing spectral data mentioned above were utilized to establish a PLSR model for five key indicators: hydrolyzed N, available P, quick-release K, OM, and pH. ESI, Table 3† provides the statistical performance of all models based on various preprocessing data. ESI, Table 4† presents the optimal single spectral model performance and corresponding Raman and NIR spectral preprocessing methods.

From ESI, Table 4,† three conclusions can be drawn. Firstly, the training set R^2 of the five indicators is generally higher than that of the test set, while the training set RMSE is generally lower than that of the test set, indicating overfitting of the training data and the need for feature extraction. Secondly, the accuracy of the measurements of all five indicators was insufficient, necessitating measures to improve them, such as the utilization of dual data. Thirdly, implementing a tailored preprocessing approach for each indicator model significantly enhances the regression accuracy of the model. As indicated in ESI, Table 4,† the 1st derivative Raman data were effective for hydrolyzed N, quick-release K, OM, and pH, while the MSC Raman data were optimal for available P. Additionally, the 1st derivative NIR data proved effective for hydrolyzed N, available P, quick-release K, and pH, with SG + 1st derivative NIR data being ideal for OM. The preprocessing data are subsequently integrated to establish measurement models. Furthermore, to reduce data correlation and improve model measurement accuracy, feature extraction will also be carried out next. ESI, Fig. 4† describes the process of feature extraction and ESI, Table 5† presents the statistical performance of all models based on Raman and NIR feature extraction data. The analysis has demonstrated that the 1st derivative and MSC methods stand out for their superior performance as preprocessing techniques for Raman models, whereas the SG + 1st derivative and 1st derivative emerge as the optimal preprocessing approaches for NIR models.

4.2 Dual spectral data fusion and feature extraction

Consequently, the Raman 1st derivative data, Raman MSC data, NIR 1st derivative data, and NIR SG + 1st derivative data were individually subjected to maximum–minimum normalization, compressing the data within the range of [0, 1]. Subsequently, the fusion data of these three types of dual data were obtained through sequential concatenation of Raman and NIR data. Specifically, the resulting data include Raman 1st derivative + NIR 1st derivative data, Raman MSC + NIR 1st derivative data, and Raman 1st derivative + NIR SG + 1st derivative data, as depicted in Fig. 2.

The characteristic variables of the fused spectra were then extracted using four methods, CARS, SPA, UVE, and PCA, for each of the five measurement models.

4.2.1 CARS-based feature extraction of dual spectral data. For the extraction of Raman and NIR dual spectral feature variables based on CARS, the Monte Carlo (MC) value was set to 50. Fig. 3(a) reveals that the minimum RMSECV value of 16.8673 and 57 preferred data points were attained at the 29th sampling. In Fig. 3(b), the minimum RMSECV value of 8.3932 and 93 preferred data points were achieved at the 26th sampling. Fig. 3(c) demonstrates that the minimum RMSECV value of 30.2301 and 42 preferred data points were obtained at the 31st sampling. Fig. 3(d) displays the minimum RMSECV value of 0.4428 and 49 preferred data points recorded during the 30th

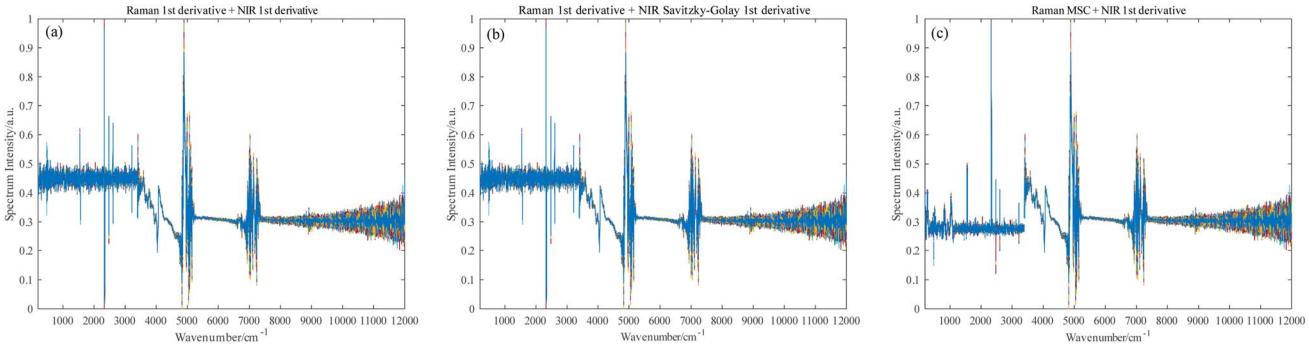

Fig. 2 Raman and NIR dual spectra, Raman 1st derivative + NIR 1st derivative (a), Raman 1st derivative + NIR Savitzky–Golay 1st derivative (b), and Raman MSC + NIR 1st derivative (c).

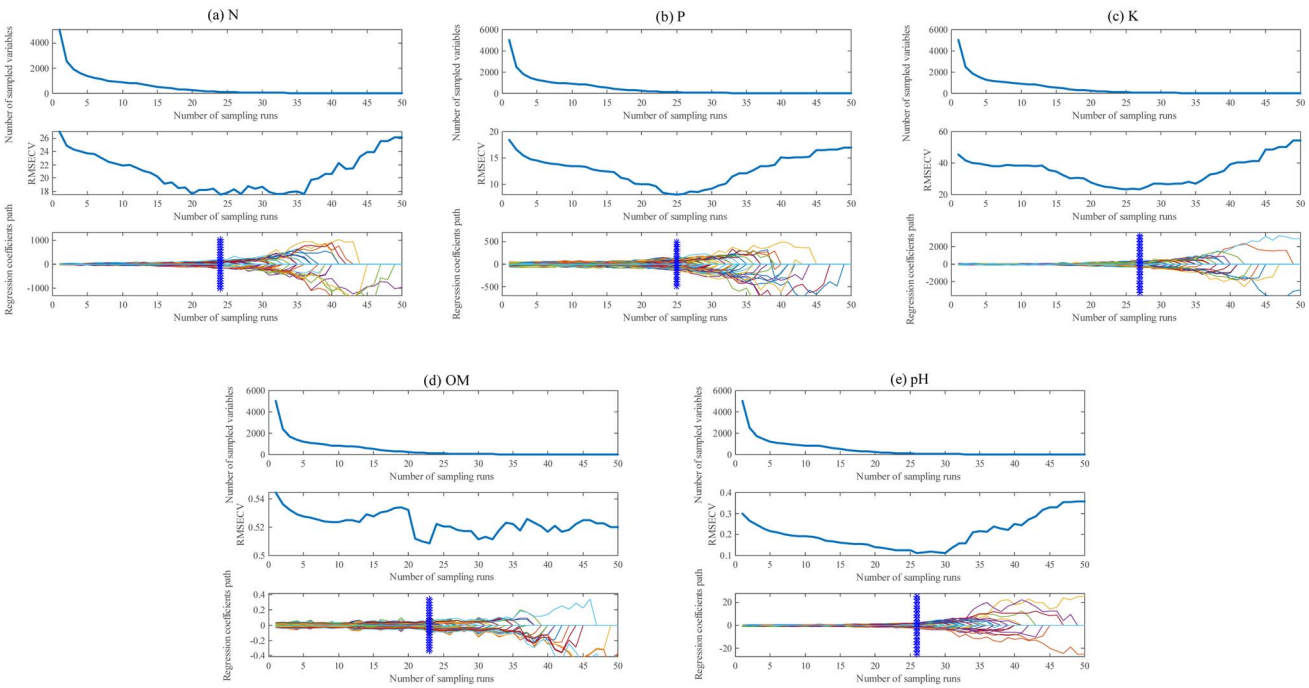

Fig. 3 CARS feature extraction for hydrolyzed N (a), available P (b), quick-release K (c), OM (d) and pH (e) models using fusion data.

sampling. Finally, Fig. 3(e) illustrates the minimum RMSECV value of 0.1213 and 79 preferred data points identified at the 27th sampling.

4.2.2 SPA-based feature extraction of dual spectral data. Raman and NIR dual spectral features were extracted using SPA. The study defined the number of characteristic variables to range between 1 and 50. For instance, when the minimum RMSE was 12.5234, four characteristic variables were identified for measuring hydrolyzed N content as shown in ESI, Fig. (5a and b);† with a minimum RMSE of 12.0154, one characteristic variable was determined for measuring available P content as shown in ESI, Fig. (5c and d)† shown; and when the minimum RMSE reached 44.6045, 17 characteristic variables were necessary for measuring quick-release K content as shown in ESI, Fig. (5e and f).† Similarly, for OM measurement, a minimum RMSE of 0.33929 led to the identification of four characteristic variables as shown in ESI, Fig. (5g and h),† while an RMSE of 0.19035 resulted in the discovery of 11 characteristic variables for pH value measurement as shown in ESI, Fig. (5i and j).†

4.2.3 UVE-based feature extraction of dual spectral data. The extraction of Raman and NIR dual spectral feature variables based on UVE revealed specific thresholds and selected wavelengths for each model. For example, the hydrolyzed N model had a threshold of 20.25, with 97 wavelengths selected; the available P model's threshold was 21.29, with 254 wavelengths selected; the quick-release K model's threshold stood at 21.55, with 115 wavelengths selected; the OM model's threshold was 21.08, with 67 wavelengths selected; and the pH model's threshold was 21.36, with 231 wavelengths selected as shown in ESI, Fig. 6.†

4.2.4 PCA-based feature extraction of dual spectral data. Utilizing PCA for feature extraction, the study analyzed the contribution rates of principal components (PCs) in each model. The cumulative contribution rates of the first three PCs

Table 1 Statistical performance of models based on bispectral eigenvariables

Spectroscopy	Category	Feature extraction	Principal components	R_c^2	R_p^2	RMSEC	RMSEP	RPD
Raman-NIR	Hydrolyzed N (mg kg^{-1})	**CARS**	**6**	**0.9992**	**0.9964**	**1.4097**	**2.9766**	**5.1750**
		SPA	4	4.5251	0.5945	32.1348	12.5233	7.4350
		UVE	5	0.9276	0.2275	9.0945	17.3659	5.9008
		PCA	8	0.9974	0.3871	1.7685	15.3965	6.3224
	Available P (mg kg^{-1})	**CARS**	**6**	**0.9975**	**0.9375**	**1.1328**	**3.5907**	**2.1407**
		SPA	1	4.9985	0.2682	21.3183	12.0153	2.9526
		UVE	8	0.9973	0.1014	1.1911	13.3147	3.3055
		PCA	7	0.9975	0.1000	1.1401	13.3032	2.3788
	Quick-release K (mg kg^{-1})	**CARS**	**6**	**0.9982**	**0.9845**	**2.7446**	**8.0202**	**2.7829**
		SPA	17	0.6011	0.5275	35.3339	44.6044	2.9614
		UVE	4	0.9053	0.7916	19.3090	30.3886	2.7871
		PCA	7	0.9942	0.7202	5.0204	34.3046	2.9834
	OM (mg kg^{-1})	**CARS**	**6**	**0.9950**	**0.9355**	**0.0327**	**0.1027**	**7.5122**
		SPA	4	3.3896	0.3580	0.4108	0.3392	8.4762
		UVE	8	0.9963	0.1562	0.0275	0.3889	10.3430
		PCA	7	0.9911	0.1483	0.0426	0.3907	8.7996
	pH	**CARS**	**6**	**0.9988**	**0.9939**	**0.0152**	**0.0342**	**9.7403**
		SPA	11	0.2679	0.7845	0.3041	0.1903	11.2676
		UVE	9	0.9995	0.7827	0.0103	0.1911	31.6612
		PCA	8	0.9985	0.7867	0.0179	0.1893	10.6567

Table 2 Standard deviation detection of CARS-PLS models using dual features

Category	R_c^2	R_p^2	RMSEC	RMSEP	RPD
Hydrolyzed N	0.9988 ± 0.0003	0.9944 ± 0.0015	1.7252 ± 0.2203	3.9570 ± 0.5399	4.9808 ± 0.5347
Available P	0.9966 ± 0.0010	0.9466 ± 0.0243	1.3310 ± 0.1671	3.1751 ± 0.6096	2.1849 ± 0.1857
Quick-release K	0.9975 ± 0.0010	0.9840 ± 0.0051	3.2228 ± 0.5671	7.7988 ± 1.2125	2.8770 ± 0.1739
OM	0.9912 ± 0.0022	0.9376 ± 0.0175	0.0433 ± 0.0052	0.0947 ± 0.0142	7.6629 ± 0.5154
pH	0.9980 ± 0.0006	0.9886 ± 0.0039	0.0194 ± 0.0030	0.0452 ± 0.0083	9.6847 ± 1.4358

were calculated, indicating their significance in explaining variance within the models. ESI, Fig. 7a–e† depict the distribution of training and testing samples in the PC space, suggesting a balanced allocation for rational training and testing.

Subsequently, eigenvariables were employed to construct PLSR models for the five indicators, with Table 1 presenting the statistical performance of these models.

Table 1 highlights an interesting observation: the red-highlighted SPA method may only capture the most crucial components, resulting in the extraction of just one feature variable, but the model's evaluation metrics are disappointingly weak. This suggests that in high-dimensional hyperspectral data, there exists a plethora of correlated dimensions and redundant information, making appropriate and precise feature extraction an essential factor.

The bold-highlighted CARS feature model in Table 1 exhibited superior accuracy when compared to other models. To validate this conclusion, feature variables were extracted from individual spectra (NIR and Raman) as detailed in ESI, Table 5.† The statistical performance of these variables indicated that CARS remains the optimal feature extraction method for measuring the five key soil indicators. Up to this point, the best data preprocessing and feature extraction methods have been identified through statistical screening. Subsequently, the PLS measurement models were developed based on these feature data, and their stability was analyzed using a 5-fold cross-validation (cv = 5) method. The standard deviation performance after running the model 50 times is documented in Table 2.

The performance of the model on different data subsets can be more comprehensively understood through 5-fold cross-validation to ensure the stability of the model performance. Comparing the results from Table 2 and the reported studies in ESI, Table 1,† it was observed that the PLS model based on dual spectra achieved the highest R_c^2 and R_p^2 values with relatively small fluctuations, indicating good fitting ability and prediction accuracy. The small values of RMSEC and RMSEP, along with RPD values greater than 2, suggested minimal measurement bias for all parameters, indicating practical usability. Additionally, comparing the R_p^2 values of the available P and OM models (0.9466 and 0.9376) with those of the other models (0.9944, 0.9840, and 0.9886), as well as the standard deviations labeled in red, suggested the potential for further optimization of the PLS models to enhance measurement accuracy and stability.

4.3 Stacking models using dual features

4.3.1 Construction process of stacking models. The research has developed a two-layer network structure for

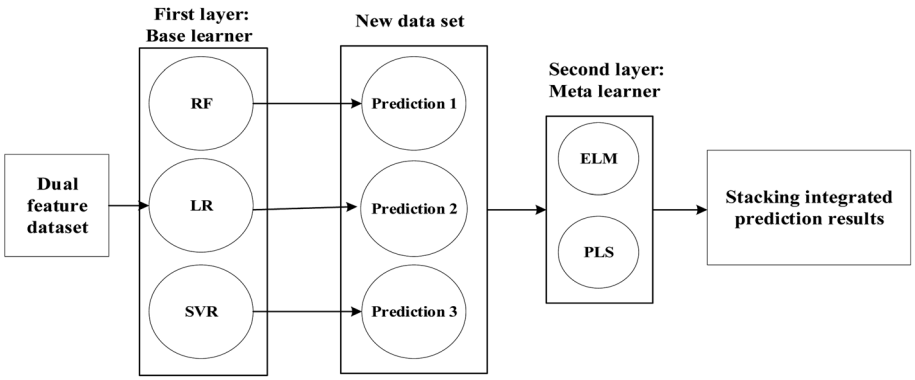

Fig. 4 Architecture of the stacking learning model.

creating a stacking measurement model. The initial layer comprises three fundamental learners: Random Forest (RF), Linear Regression (LR), and Support Vector Regression (SVR). The second layer includes two base learners, namely Partial Least Squares (PLS) and Extreme Learning Machine (ELM), as depicted in Fig. 4. The RF method integrates multiple decision trees to mitigate the risk of overfitting. On the other hand, the LR method is well-suited for datasets with strong linear correlations, offering high computational efficiency and strong model interpretability. SVR methods are adept at learning complex data patterns through the use of kernel functions for nonlinear transformations, making them advantageous in handling nonlinear relationships and high-dimensional data.

The first layer network amalgamates the benefits of these three methods, thereby enhancing the measurement accuracy of multi-parameter models. The second layer aims to select models with strong generalization ability and complementary advantages to rectify biases in the first layer base learner for the training model. This further improves prediction performance by making a second prediction on the data generated by the first layer base model.

The PLS method excels at handling high-dimensional data with a larger number of features than the number of samples. However, this may be impacted by the data's distribution, potentially diminishing the model's predictive ability in cases of weak correlation or poor fitting for nonlinear relationships. On the other hand, the ELM method efficiently handles complex nonlinear patterns through nonlinear mapping in the implicit layer. Nonetheless, it is sensitive to noise and may overfit or reduce measurement accuracy in noisy datasets. However, the PLS method leverages the variance interpretation ability of the regression vector to select effective features, thereby demonstrating resistance and robustness to noise. The complementary advantages of PLS and ELM underscore their strong generalization abilities, leading to their selection as second layer meta-learners.

To ensure optimal model fitting, the study utilizes the grid search method to optimize the parameters of the base learners. This optimization, within an acceptable range of computational complexity, further improves the measurement accuracy of the available P and OM models and enhances the reproducibility of the hydrolyzed N, available P, quick release K, and OM models.

The stacking training process diagram is shown in Fig. 5, and the specific process is as follows:

Step 1: first, the characteristic data set is divided into the training set D and the test set V.

Step 2: next, a 5-fold cross-validation method is applied to train each base learner. The dataset D is split into five distinct subsets, D_1, D_2, D_3, D_4, and D_5, evenly. Subsequently, the training set is created by combining four subsets while denoting one subset as the testing set to establish both training and testing sets for the primary learner. This process allows each primary learner to have five sets of training and testing sets, resulting in five test results S_1, S_2, S_3, S_4, and S_5.

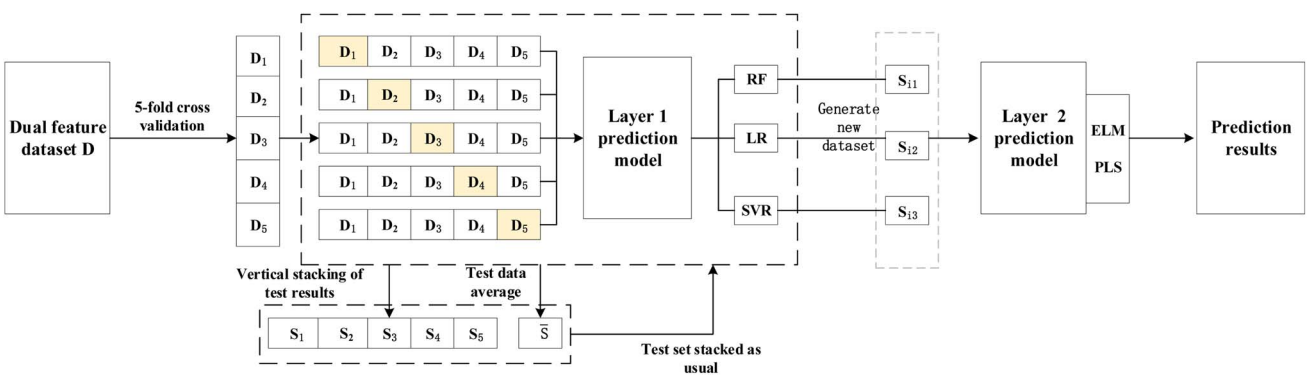

Fig. 5 Stacking training flowchart.

Step 3: following this, a new training dataset is generated. The first layer of the stacking network comprises three base learners, all of whom are trained and tested using the aforementioned five sets of training and testing sets. After completing 5-fold cross-validation for the n base learner, the resulting five prediction sets are vertically stacked in rows to generate the prediction set S_{i1}, S_{i2}, and S_{i3}, $(i = 1, 2, 3, 4, 5)$ for the sample data under this base learner. Simultaneously, these five prediction results are averaged to obtain S_n $(i = 1, 2, 3)$. Once the training of the first layer's three base learners is finalized, the predicted set from each base learner is concatenated with the predicted mean in columns to create a new training set S_i, n and a new test set S_n for the second layer. The input dataset for constructing the second layer meta-learner is the dataset (S_n, S_{in}) $(i = 1, 2, 3, 4, 5; n = 1, 2, 3)$.

Step 4: the new training set and test set obtained from the first-level primary learner are fed into the second-level meta-learner for additional training, resulting in the final prediction of soil data.

5 Results and analysis

5.1 Verification of stability assessment

The 5-fold cross-validation performance of the measurement models, which combine two-layer learners (RF, LR, SVR; ELM, and PLS) using the stacking method, is presented in Table 3 after 50 runs.

The comparative analysis between Tables 2 and 3 highlights the effectiveness of the stacking model in maintaining the exceptional performance of the original PLS model while addressing its previous shortcomings. Notably, the stacking approach resulted in improvements across various metrics. For instance, the R_p^2 for the available P model increased from 0.9466 to 0.9491, and for the OM model, it improved from

Table 3 Standard deviation detection of stacking models

Category	R_c^2	R_p^2	RMSEC	RMSEP	RPD
Hydrolyzed N	0.9997 ± 0.0001	0.9952 ± 0.0009	0.8220 ± 0.1290	3.7063 ± 0.3633	5.5100 ± 0.4267
Available P	0.9994 ± 0.0002	0.9491 ± 0.0166	0.5519 ± 0.1044	3.1447 ± 0.5454	2.2145 ± 0.1797
Quick-release K	0.9994 ± 0.0002	0.9842 ± 0.0032	1.5914 ± 0.2642	7.8042 ± 0.7719	2.8898 ± 0.1695
OM	0.9963 ± 0.0008	0.9520 ± 0.0105	0.0290 ± 0.0038	0.0833 ± 0.0099	7.8319 ± 0.4808
pH	0.9994 ± 0.0002	0.9901 ± 0.0028	0.0114 ± 0.0016	0.0428 ± 0.0062	9.9606 ± 0.3505

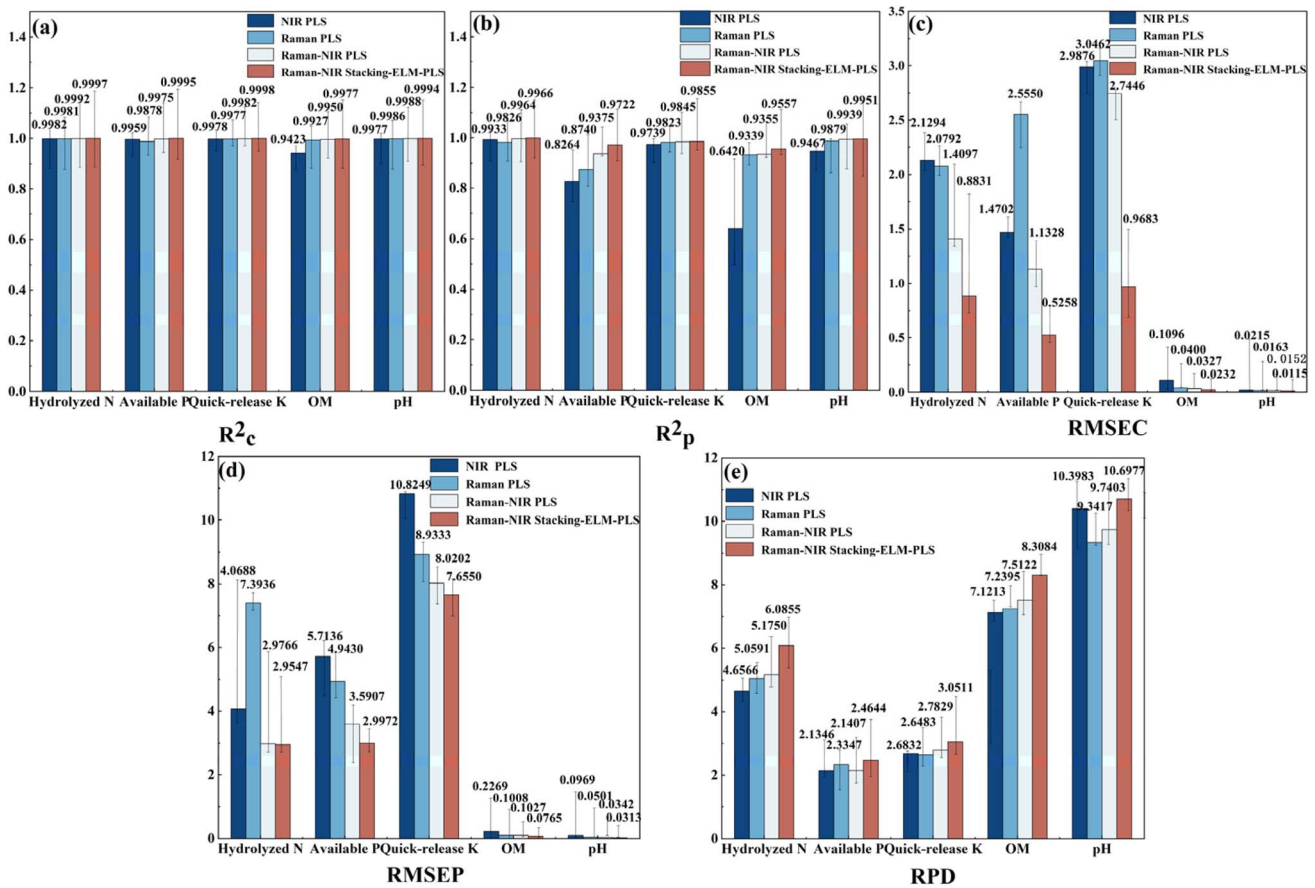

Fig. 6 Comparison of stacking models based on bar charts.

Table 4 Model performance comparison under noise and non-noise conditions

Category	Models	Testing set without noise		Testing set with noise		Relative change rate (%)		Average relative change rate (%)
		R_p^2	RPD	R_p^2	RPD	R_p^2	RPD	
Hydrolyzed N	PLS	0.9964	5.1750	0.8149	7.4792	18.22	44.53	31.38
	Stacking-ELM-PLS	**0.9966**	**6.0855**	**0.9723**	**7.5048**	**2.44**	**23.32**	**12.88**
Available P	PLS	0.9375	2.1407	0.7207	2.2080	23.13	3.14	13.14
	Stacking-ELM-PLS	**0.9722**	**2.4644**	**0.9216**	**2.2090**	**5.20**	**10.36**	**7.78**
Quick-release K	PLS	0.9845	2.7829	0.8339	2.8824	15.30	3.58	9.44
	Stacking-ELM-PLS	**0.9855**	**3.0511**	**0.9257**	**2.8858**	**6.07**	**5.42**	**5.75**
OM	PLS	0.9355	7.5122	0.8574	6.4937	8.35	13.56	10.96
	Stacking-ELM-PLS	**0.9557**	**8.3084**	**0.8793**	**7.9636**	**7.99**	**4.15**	**6.07**
pH	PLS	0.9939	9.7403	0.8911	9.4571	10.34	2.91	6.63
	Stacking-ELM-PLS	**0.9951**	**10.6977**	**0.9072**	**10.2883**	**8.83**	**3.82**	**6.33**

0.9376 to 0.9520. Moreover, the standard deviation fluctuations for all five models were reduced from ±0.0015 to ±0.0009, from ±0.0243 to ±0.0166, from ±0.0051 to ±0.0032, from ±0.0175 to ±0.0105, and from ±0.0039 to ±0.0028, respectively. Similarly, the stacking model significantly reduced the RMSEC value and its fluctuation range, while slightly decreasing the RMSEP value. It also led to an increase in the RPD value and a reduction in its fluctuation range. Specifically, the maximum standard deviation of RPD decreased by 1.48%, RMSEP by 4.44%, RMSEC by 1.25%, R_p^2 by 0.31%, and R_c^2 by 0.06%. The consistency of the stacking model under different data segmentations is confirmed by the 5-fold cross-validation, which further proves its reliability in practical applications. As shown in Tables 2 and 3, the stacking model not only improves the performance indicators, but also significantly reduces the volatility of these indicators. These results validate the effectiveness and advantages of stacked models in spectral analysis, especially in improving model stability.

5.2 Verification of accuracy assessment

To comprehensively assess the stacking model, bar charts illustrating evaluation indicators were generated for the PLS model based on NIR data, Raman data, fusion of Raman and NIR data, and the stacking model based on fusion data. Fig. 6 shows that among the four model types with five parameters, the stacking model demonstrated superior performance in terms of R_c^2 and R_p^2, along with the lowest values for RMSEC and RMSEP. Additionally, it achieved the highest RPD value. The ablation experiments demonstrate that stacked models outperform the near-infrared single spectral model, the Raman single spectral model, and the dual-spectral PLS model in terms of the five indicators: R_c^2, R_p^2, RMSEC, RMSEP and RPD, indicating the excellent accuracy of the stacked models.

5.3 Validation of robustness analysis

To evaluate the robustness of the model to various interference sources, Gaussian noise, Poisson noise, uniform noise and power law noise are introduced into the characteristic variables of CARS. The mean of Gaussian noise is 0, the standard deviation is 0.001, the variance of Poisson noise is 0.001, the variance of uniform noise is 0.004, and the variance of power law noise is 0.009. Table 4 lists the performance of Raman-NIR PLS and stacking-ELM-PLS models under the combination of linear and nonlinear noise.

As indicated in Table 4, the impact of noise stress on the measurement accuracy of the two models is substantial. The stacking-ELM-PLS model shows the least variation in indicators when subjected to noise stress, with average relative change rates of 12.88%, 7.78%, 5.75%, 6.07%, and 6.33%. In contrast, the average relative change rates of the PLS model are 31.38%, 13.14%, 9.44%, 10.96%, and 6.63%. This suggests that the stacking-ELM-PLS model exhibits robustness. Additionally, it is noteworthy that the stacking-ELM-PLS model maintains the highest R_p^2 and RPD values under noise stress, further underscoring its superior resistance to noise.

6 Conclusion

The study introduces an efficient soil multi-parameter rapid measurement technique that involves the fusion of Raman and NIR data, utilizing a two-layer network structure (incorporating RF, LR, SVR; ELM, and PLS) regressor grounded on the stacking algorithm. The regressor amalgamates the strengths of five distinct base learners, presenting a diminished risk of overfitting, heightened computational efficiency concerning linear correlation data, enhanced fitness regarding nonlinear correlation data, robust generalization potential, noise resistance, and robustness. By employing Raman and NIR spectroscopy fusion data, a soil multi-parameter measurement model is conceived using the stacking algorithm. The model yields remarkable R_p^2 values for hydrolyzed N, available P, quick-release K, OM, and pH, of 0.9966, 0.9722, 0.9855, 0.9557, and 0.9951, respectively, with corresponding RMSEP values of 2.9547, 2.9972, 7.6550, 0.0765, and 0.0313, and RPD values of 6.0855, 2.4644, 3.0511, 8.3084, and 10.6977. The outcomes of model tests affirm that the stacking models, leveraging dual data sources, showcase commendable fitting precision, robust predictive capabilities, exceptional stability, and reliable observational values. Ultimately, the model stands poised to swiftly assess the physical and chemical conditions of soil in an online setting.

Data availability

The data supporting this article have been included as part of the ESI.†

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This research was supported by the Natural Science Foundation of Heilongjiang Province in China (No. LH2022C061) and the Heilongjiang Bayi Agricultural University Youth Innovative Talent Project (ZRCQC202205).

References

1 F. Li, Problems and countermeasures of soil fertilizer in agricultural production, *Agriculture and Technology*, 2020, **40**(11), 64–65, DOI: **10.19754/j.nyyjs.20200615021**.
2 Ministry of Agriculture and Rural Affairs, *People's Republic of China Ministry of Agriculture and Rural Affairs Announcement No. 424*, Bulletin of Ministry of Agriculture and Rural Affairs of the People's Republic of China, 2021, (08), p. 102.
3 J. Zhang, *et al.*, Comparative analysis and trend study on relevant standards of organic fertilizers in Chinese and Russian agricultural standards, *Standardization in China*, 2022, (11), 189–192, DOI: **10.3969/j.issn.1002-5944.2022.11.029**.
4 X. Liu, Near infrared diffuse reflectance spectroscopy for the detection of soil organic matter and available N, *Chinese Journal of Agricultural Mechanization*, 2013, **34**(2), 202–206, DOI: **10.3969/j.issn.2095-5553.2013.02.052**.
5 T. Wang, *Research on Detection Method of Soil Physicochemical Information Based on Spectral Technology*, Zhejiang University, 2019.
6 S. Jia, *et al.*, Determination of soil available phosphorus and potassium by near infrared spectroscopy combined with recursive partial least squares algorithm, *Spectrosc. Spectr. Anal.*, 2015, (9), 2516–2520, DOI: **10.3964/j.issn.1000-0593(2015)09-2516-05**.
7 Y.-De Liu, *et al.*, Near-infrared spectroscopy of total phosphorus and total potassium in the soil of navel orange orchard in southern Jiangxi Province, *Transactions of Agricultural Engineering*, 2013, (18), 156–162, DOI: **10.3969/j.issn.1002-6819.2013.18.019**.
8 T. Dong, *et al.*, Rapid and Quantitative Determination of Soil Water-Soluble Nitrogen Based on Surface-Enhanced Raman Spectroscopy Analysis, *Appl. Sci.*, 2018, **8**(5), 701, DOI: **10.3390/app8050701**.
9 L. Zheng, *et al.*, Analysis of soil phosphorus concentration based on Raman spectroscopy, in *Multispectral, Hyperspectral, and Ultraspectral Remote Sensing Technology, Techniques and Applications IV*, Kyoto (JP), 2012, vol. 8527, p. 852718.
10 Z. Xing, *et al.*, Characterizing typical farmland soils in China using Raman spectroscopy, *Geoderma*, 2016, **268**, 147–155, DOI: **10.1016/j.geoderma.2016.01.029**.
11 Z. Xing, *et al.*, Application of FTIR-PAS and Raman spectroscopies for the determination of organic matter in farmland soils, *Talanta*, 2016, **158**, 262–269, DOI: **10.1016/j.talanta.2016.05.076**.
12 An Xiaofei, *et al.*, Effect of soil moisture on real-time detection of soil total nitrogen by near-infrared spectroscopy, *Spectrosc. Spectral Anal.*, 2013, **33**(3), 677–681, DOI: **10.3964/j.issn.1000-0593(2013)03-0677-05**.
13 Y. Zhang, *et al.*, Soil nitrogen content forecasting based on real-time NIR spectroscopy, *Comput. Electron. Agric.*, 2016, **124**, 29–36, DOI: **10.1016/j.compag.2016.03.016**.
14 Y. He, *Classification and identification of blast resistant varieties based on near infrared spectroscopy*, Tongfang Knowledge Network (Beijing) Technology Co., Ltd, Beijing, 2021, DOI: **10.27122/d.cnki.ghlnu.2021.000275**.
15 P. Zhou, *et al.*, Development and performance tests of an on-the-go detector of soil total nitrogen concentration based on near-infrared spectroscopy, *Precis. Agric.*, 2021, **22**(5), 1479–1500, DOI: **10.1007/s11119-021-09792-0**.
16 Y. Li, J.-Y. Zhang and Y.-Z. Wang, FT-MIR and NIR spectral data fusion: a synergetic strategy for the geographical traceability of Panax notoginseng, *Anal. Bioanal. Chem.*, 2018, **410**(1), 91–103, DOI: **10.1007/s00216-017-0692-0**.
17 M. Li, *Rapid Analysis of Methanol Content in Methanol Gasoline Based on Raman NIR Spectral Data Fusion*. Xi 'an Shiyou University, 2020.
18 S. Li, *et al.*, Nondestructive identification of soybean milk powder based on near infrared spectroscopy and optimized pretreatment method, *Food Res. Dev.*, 2020, **41**(17), 144–150, DOI: **10.12161/j.issn.1005-6521.2020.17.023**.
19 L. Wu, *et al.*, Characterization of Tobacco with Near-Infrared Spectroscopy with Competitive Adaptive Reweighted Sampling and Partial Least Squares Discrimination, *Anal. Lett.*, 2016, **49**(13/15), 2290–2300, DOI: **10.1080/00032719.2016.1144763**.
20 J.-H. Cheng and D.-W. Sun, Combining the genetic algorithm and successive projection algorithm for the selection of feature wavelengths to evaluate exudative characteristics in frozen–thawed fish muscle, *Food Chem.*, 2016, **197**(PA), 855–863, DOI: **10.1016/j.foodchem.2015.11.019**.
21 Q. Li, *et al.*, Optimization of quantitative models of total nitrogen and total sugar in tobacco by variable screening without information variable elimination, *Anal. Chem.*, 2013, **41**(6), 917–921, DOI: **10.3724/sp.J.1096.2013.21017**.
22 S. Chen, *et al.*, Prediction of earthquake death toll based on PCA-PSO-ELM model, *J. Geodesy Geodyn.*, 2024, **44**(1), 105–110, DOI: **10.14075/j.jgg.2023.03.107**.
23 C. Zhang, *Research on Mechanism and Method of Rape Disease Detection Based on Spectral and Spectral Imaging Technology*, Zhejiang University, 2016.
24 G.-B. Huang, Q.-Y. Zhu and C.-K. Siew, Extreme learning machine: theory and applications, *Neurocomputing*, 2006, **70**(1–3), 489–501, DOI: **10.1016/j.neucom.2005.12.126**.

25 X. Li, *et al.*, Multi-model fusion stacking ensemble learning method for the prediction of berberine by FT-NIR spectroscopy, *Infrared Phys. Technol.*, 2024, **137**, 105169, DOI: **10.1016/j.infrared.2024.105169**.

26 P. D. T. Nie and Y. He, The Effects of Drying Temperature on Nitrogen Concentration Detection in Calcium Soil Studied by NIR Spectroscopy, *Appl. Sci.*, 2018, **8**(2), 269, DOI: **10.3390/app8020269**.