ROYAL SOCIETY
OF CHEMISTRY

PAPER

Check for updates

Enhancing multi-type coal quality prediction accuracy with fusion spectra and classification models using NIRS and XRF techniques

Zhedong Zhang, [ab] Jiaxuan Li, [ab] Rui Gao, [ab] Yang Zhao, [c] Yan Zhang, [d] Lei Zhang, [iD] *[ab] Zefu Ye, [e] Zhujun Zhu, [e] Peihua Zhang, [a] Wangbao Yin*[ab] and Suotang Jia[ab]

The various analytical indices of coal are important criteria for evaluating the quality of commercial coal. Coals of different qualities exhibit different physical and chemical characteristics in their utilization. In the case of multiple coal types, the spectral characteristics of different coals may overlap within certain wavelength ranges, or be affected by interference or noise from other coal types, leading to low accuracy in coal quality prediction. Rapid and accurate coal quality testing is of great significance for improving industrial production efficiency and enhancing corporate profitability. This study employs near-infrared spectroscopy (NIRS) and X-ray fluorescence spectroscopy (XRF) combined techniques to explore the accuracy and feasibility of predicting coal quality based on coal type classification models. In terms of classification algorithms, coal samples are identified and classified using Support Vector Machine (SVM) based on fusion spectra. Regarding the modeling approach, Partial Least Squares (PLS) is utilized to establish both an overall model for all coal samples and individual classification models corresponding to each coal type. The results show that the precision, accuracy, recall, and F_1 score of this classification algorithm reached 96.49%, 97.50%, 95.83%, and 96.41%, respectively. The determination coefficients (R^2) for the classification model's predictions of ash, volatile matter, and sulfur in coal quality indicators reached 0.992, which represents improvements of 1.85%, 5.31%, and 10.10% over the overall model. The root mean square errors of prediction ($RMSE_P$) for these indicators were 0.062, 0.080, and 0.008, showing reductions of 0.24%, 0.68%, and 0.05% compared to the overall model. It indicates that the method of first identifying the coal type and then predicting coal quality indicators using the corresponding classification model can significantly improve the accuracy of coal quality detection in complex coal type scenarios.

1 Introduction

Coal is one of the most important energy sources in the world, widely used in power generation, steel production, cement manufacturing, and other fields. It is closely related to national economic development.[1,2] At the same time, as an energy reserve coal plays an indispensable role in ensuring the stable development of the national economy and strategic security.[3] However, the quality of coal varies widely.[4] Coal chemical enterprises often face the challenge of multiple coal types with significant differences in their physical and chemical

properties. Therefore, it is particularly important to find methods that can quickly and reasonably classify coal and perform multi-parameter analysis.[5,6]. Currently, coal quality testing still relies on national standard methods based on traditional physical and chemical principles. However, there are issues such as complex procedures, lengthy processes, and susceptibility to human interference.[7,8] Therefore, developing rapid and accurate coal quality testing technologies capable of simultaneously detecting multiple analysis indicators is of great value for improving industrial production efficiency and enterprise benefits.

With the increasing maturity of analytical techniques, various rapid coal quality analysis technologies and instruments have emerged in the market. These include Prompt Gamma Neutron Activation Analysis (PGNAA), Laser-Induced Breakdown Spectroscopy (LIBS), Near-Infrared Spectroscopy (NIRS), X-ray Fluorescence Spectroscopy (XRF), and others. Among them, PGNAA utilizes neutron beam radiation to activate coal, causing different isotopes in coal to interact with neutrons and produce high-energy prompt characteristic

aState Key Laboratory of Quantum Optics and Quantum Optics Devices, Institute of Laser Spectroscopy, Shanxi University, Taiyuan, China. E-mail: k1226@sxu.edu.cn

bCollaborative Innovation Center of Extreme Optics, Shanxi University, Taiyuan, China

cSchool of Semiconductor and Physics, North University of China, Taiyuan 030051, China

dSchool of Optoelectronic Engineering, Xi'an Technological University, Xi'an, China

eShanxi Gemeng US-China Clean Energy R&D Center Co., Ltd., Taiyuan, China

This journal is © The Royal Society of Chemistry 2024

J. Anal. At. Spectrom., 2024, 39, 2433–2442 | 2433

gamma rays. Coal quality multi-index analysis is achieved by analyzing the energy spectrum.[9] However, the use of such equipment is limited due to its radioactive hazards, which restrict its widespread application. LIBS utilizes high-energy pulsed lasers to generate high-temperature plasma. By detecting the specific wavelength radiation emitted when atoms and ions in the plasma transition from excited states to the ground state, qualitative or quantitative analysis of coal samples can be performed.[10] However, the measurement repeatability of this technology is constrained by Rayleigh–Taylor instability and self-absorption effects.[11,12]

NIRS is a non-destructive analytical method based on the absorption characteristics of near-infrared light in samples. It utilizes the absorption of specific wavelength ranges caused by the vibration, rotation, and interatomic vibration of molecules to perform quantitative analysis.[13] Due to its rapid, non-destructive nature, and the absence of sample pre-treatment requirements, NIRS has been successfully applied in the analysis and assessment of coal quality. For example, Wang[14] *et al.* proposed a near-infrared spectroscopy method for measuring coal industrial indicators based on the synergistic adaptive moving window PLS genetic algorithm. The root mean square error (RMSE) for calorific value and volatile matter were 0.964 MJ kg^{-1} and 0.889%; Begum[15] *et al.* utilized five absorption bands within the visible to near-infrared range to establish a multiple linear regression model for predicting the calorific value of coal. The coefficients of determination (R^2) and mean absolute percentage error were 0.92 and 4.84%; Gao[16] *et al.* proposed a novel technique for the ultra-reproducible measurement of coal calorific value using a combination of near-infrared spectroscopy (NIRS) and X-ray fluorescence spectroscopy (XRF). They established a comprehensive-segmented prediction model based on partial least squares (PLS). The root mean square error of prediction (RMSE$_P$), average relative error (ARE), and standard deviation (SD) for predicting coal calorific value were reported as 0.71 MJ kg^{-1}, 1.18%, and 0.07 MJ kg^{-1}. NIRS demonstrates high repeatability in measurements, but its capability is limited to detecting coal calorific value and volatile matter indicators that are positively correlated with organic functional groups such as C and H.

The working principle of XRF involves using high-energy X-rays to excite the atoms in the sample, causing the atoms to release energy in the form of fluorescence. By detecting and analyzing the fluorescence radiation, the composition elements and content of the sample can be determined.[17] For instance, Hicks[18] *et al.* employed XRF for measuring the ash content of coal, achieving an SD value of 1.7%. Similarly, Ma[19] *et al.* utilized XRF for measuring the repeatability of ash-forming elements such as Si, Al, Fe, Ca, Mg, Na, K, and Ti in coal, which was significantly lower than the requirements of national standard analysis. XRF demonstrates excellent measurement repeatability, but it also has some limitations. For example, it cannot effectively excite low atomic number elements such as C and H in coal, thus primarily used for measuring coal ash content or high atomic number element concentrations. We previously developed a coal quality analysis technique by NIRS-XRF,[20] achieving comprehensive detection of organic and inorganic components in coal. The detection accuracy for coal calorific value reached an R^2 of 0.98, RMSE$_P$ of 0.19 MJ kg^{-1}, and MARD$_P$ of 0.95%. However, in practical applications, the wide variety and diversity of coal sources in coking and other industries have significantly reduced the measurement accuracy and applicability of this technique due to the matrix effect.

In response to the above challenges, this study proposes a classification-based coal quality analysis method and conducts experimental verification, aiming to enhance the capability of NIRS-XRF combined spectroscopic analysis technique to handle complex coal types. This method first utilizes SVM algorithm to identify the coal types of samples, and then predicts their ash content, volatile matter, sulfur content, and other indicators using the corresponding PLS regression models for each coal type.

2 Experiment

2.1 Samples

The experiment gathered 187 coal samples from the coal preparation plant of Yangguang Coking Group in Shanxi Province. According to the Classification of coal in China (Chinese national standard GB/T 5751-2009), these samples are classified into four types of coal: gas coal, fat coal, coking coal, and lean coal. To ensure consistency, all samples underwent the standard general analysis coal preparation process, involving crushing, blending, sizing, dewatering, and grinding to a particle size of 0.2 mm. The coal preparation plant tests the coal quality to obtain standard values for ash, volatile matter, and sulfur according to Proximate analysis of coal (Chinese national standard GB/T 212-2008) and Determination of total sulfur in coal (Chinese national standard GB/T 214-2007).

To build and validate the model, the 187 samples were randomly divided into two subsets using a 7 : 3 ratio: one is the training set, consisting of 130 samples used to collect spectral data and build the coal quality prediction model; the other is the test set, consisting of 57 samples used to verify the accuracy and stability of the model. The coal quality indicators for the samples in the training and test set are shown in Table 1.

2.2 Experimental setup

The experimental NIRS-XRF setup for coal analysis is shown in Fig. 1. The experimental setup mainly consisted of the NIRS module, XRF module, delivery module and control module.

In the NIRS module, a halogen light source (AvaLight-HAL-S-Mini, Avantes, Netherlands) with a wavelength range of 360 to 2500 nm is used. The near-infrared spectrometer is a Fourier-transform type (C15511-01, HAMAMATSU, Japan), covering the wavelength range of 1100 to 2500 nm in the near-infrared region, with a spectral resolution of 5.7 nm. A reflective fiber optical probe (FCR-7UVIR200, Avantes, Netherlands) is employed, containing 7 cores of 200 μm each: 6 cores connect to the light source and 1 core connects to the spectrometer. The probe is positioned perpendicular to the sample surface, and the spectrometer outputs the spectral data to a computer for further processing. The reference tile (WS-2, Avantes) is placed

2434 | *J. Anal. At. Spectrom.*, 2024, **39**, 2433–2442

This journal is © The Royal Society of Chemistry 2024

Table 1 Statistics of coal quality indices for 187 samples categorized by coal types

Coal type	Ash (%)	Volatile matter (%)	Sulfur (%)	Number of samples in training set	Number of samples in test set
Gas coal	4.20–11.64	35.15–42.28	0.30–0.86	18	12
Fat coal	4.47–13.92	26.44–34.92	0.30–0.90	47	18
Coking coal	8.07–11.18	18.04–23.98	0.38–0.89	38	17
Lean coal	9.00–11.15	12.56–17.54	0.38–0.62	27	10

in front of the sample to facilitate absorbance calculations during each measurement. During the experiment, each coal sample is scanned 4000 times, and the average of these 4000 spectra is taken as its near-infrared spectral data.

The XRF module is based on energy-dispersive X-ray fluorescence (ED-XRF). In the X-ray tube (VF-50J-45°, VAREX, USA), a high voltage power supply (MNX50W, SPELLMAN, USA) generates a high voltage applied between the filament and the target. Electrons emitted from the filament are accelerated by the electric field and collide with the target material. The collision of accelerated electrons with the anode target produces X-rays, which serve as the radiation source to excite the sample. The X-rays pass through a beryllium window at the bottom and irradiate the coal sample, causing the emission of secondary X-rays, which are detected by the silicon drift detector (SDD, VICO-DV H20, KETEK, Germany). The measured energy spectrum is then subjected to quantitative analysis by a computer. During the experiment, the voltage and current of the X-ray tube are set to 16 kV and 0.6 mA, respectively. The filament voltage and current are set to 1.5 V and 2.5 A, respectively. The time constant for the SDD is 2 μs. A hydrogen generator (RH-300, Tengzhou Rapp Analytical Instrument, China) is used to fill the chamber with hydrogen gas, eliminating the absorption of low-energy X-ray fluorescence by the air.

The delivery module features a linear guide rail with a translation stage, on which the coal sample is placed in a sample holder. During measurements, a stepper motor driver controls the movement of the translation stage, allowing the sample to pass sequentially beneath the XRF detection module and the NIRS detection module. The computer collects the XRF and NIRS spectra respectively. After measurements are completed, the translation stage returns to its initial position.

The control module consists primarily of a PLC and a computer. The PLC is used to control the operating sequence of other modules, including the translation stage and high voltage power supply, while the computer is used for spectral analysis.

2.3 Spectral pretreatment

Spectral pretreatment is a crucial component of chemometric measurements, as it directly impacts the final robustness and reliability of the model. To enhance the features of the target analytical components in the spectra, spectral preprocessing is necessary. Here, we first employ the Savitzky–Golay convolution smoothing method to eliminate random noise interference. Then, we use the Standard Normal Variate (SNV) transformation method to eliminate the impact of scattering effects on spectral information caused by the uneven particle size of coal samples. Subsequently, the NIRS spectra and XRF spectra data are concatenated and standardized to a range of $[-1, 1]$,
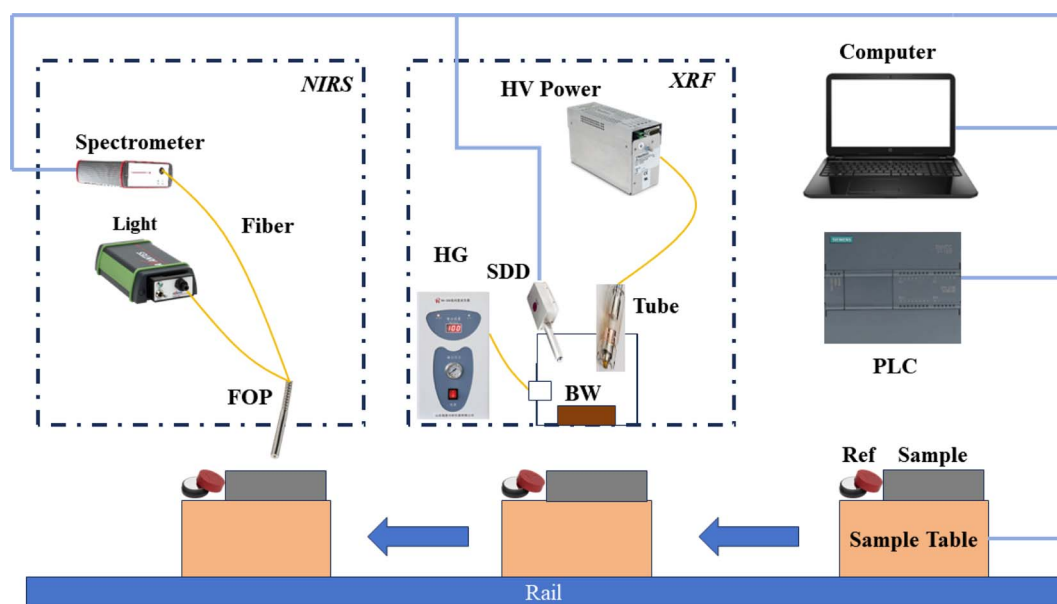

Fig. 1 Experimental NIRS–XRF setup for coal analysis (FOP: fiber optical probe; HG: hydrogen generator; SDD: silicon drift detector; HV power: high-voltage power supply; BW: beryllium window; PLC: programmable logic controller).

JAAS

Paper

ensuring data consistency and comparability, thus obtaining the NIRS-XRF fusion spectra for subsequent modeling. These spectral preprocessing steps help improve the quality and stability of our analysis model, ensuring accurate analysis and prediction.

2.4 Recognition algorithm

To achieve precise identification of the coal types to which the samples belong, this study employed the Support Vector Machine (SVM) algorithm. The training process of SVM involves finding the optimal parameters for the hyperplane, which can be achieved by solving an optimization problem. Lagrange multiplier is commonly used to solve this problem.[21] After training, SVM can be used to classify new data points.

In this study, the RBF kernel was selected, which has the advantage of nonlinear mapping. We optimized the penalty coefficient C and the kernel parameter γ. Parameter tuning is an important step to ensure that SVM achieves optimal performance and generalization ability. To optimize the hyperparameters of the SVM, we used grid search combined with K-fold cross-validation. First, we defined the value ranges for parameters C and γ, covering the parameter space using logarithmic values based on 2. The value ranges for $\log_2 C$ and $\log_2 \gamma$ were both set to $[-10, 0.2, 10]$. Then, a matrix was initialized to store the accuracies for different parameter combinations. During the grid search process, 5-fold cross-validation was used to evaluate the performance of each parameter set and calculate its average accuracy. Next, the parameter combination with the best performance on the validation set and its corresponding accuracy were identified. This method effectively selects the optimal SVM parameters for a given dataset and specific task, thereby maximizing the classifier's performance. Ultimately, the optimal values of C and γ were determined to be 73.5167 and 0.1436, respectively.

2.5 Classification prediction model

NIRS data and XRF data are both high-dimensional data. To reduce the dimensionality of the spectral data matrix and eliminate the adverse effects caused by multicollinearity between spectra, partial least squares (PLS) regression modeling method is employed here. PLS, as a multivariate statistical method, combines multiple linear regression, canonical correlation analysis, and principal component analysis, and can be used for regression analysis of problems with multiple dependent and independent variables.[22] For example, Leng *et al.*[23] utilized the PLS regression method to establish a quantitative model for certain chemical components based on near-infrared diffuse reflectance spectroscopy. They achieved a correlation coefficient R^2 of 0.9366 and reduced the root mean square error of prediction ($\mathrm{RMSE_P}$) to as low as 3.15%. In PLS modeling, the NIRS-XRF fused spectra were used as the independent variables for PLSR, with the certified values of A_{ad}, V_{ad}, and S as the dependent variables. 130 samples were used to build the model, while 57 samples were reserved for predictions using the newly established model. The optimal number of components for PLS regression was selected through cross-

validation. A 10-fold cross-validation was set up, with a random seed to ensure the reproducibility of the results. Data was randomly assigned to different cross-validation folds. For each possible number of components, cross-validation was performed and the mean absolute error (MAE) was calculated. The number of components that resulted in the smallest MAE was chosen as the optimal number of components.

To validate the robustness of the current PLS model, different calibration, segmentation, and prediction samples are used. The basic principle is to establish a linear regression model $Y = A \times X + B$ between the independent variables and the dependent variables, where Y is the response matrix containing m variables and n sample points, X is the input matrix containing p variables and n sample points, A is the regression coefficient matrix, and B is the correction coefficient.

In this study, two distinct prediction models were developed using PLS. The first model, termed the 'overall model', utilizes all coal samples for regression analysis. The second model, referred to as the 'classification model', performs regression analysis based on the determined coal sample classification strategy, serving as a basis for comparison.

2.6 Evaluation indexes

2.6.1 Evaluation of prediction model. The evaluation metrics for the segmented coal prediction model in this study include the coefficient of determination (R^2) and the root mean squared error ($\mathrm{RMSE_P}$), calculated as follows:

$$R^2 = 1 - \frac{\sum\limits_{i=1}^{m} (y_i - \hat{y}_i)^2}{\sum\limits_{i=1}^{m} (y_i - \bar{y}_i)^2}, \tag{1}$$

$$\mathrm{RMSE_P} = \sqrt{\frac{\sum\limits_{i}^{n} (y_i - \hat{y}_i)^2}{n}}, \tag{2}$$

where y_i is the standard value of the sample assay indexes, $\hat{y}_i$ is the predicted value of the sample assay indexes, $\bar{y}_i$ is the average value of the sample assay indexes, m is the number of samples in the training set, and n is the number of samples in the test set. The closer R^2 is to 1, the better the linearity of the model on the training set; the closer $\mathrm{RMSE_P}$ is to 0, the closer the predicted values are to the true values, indicating a higher accuracy of the prediction model.

2.6.2 Evaluation of classification algorithm. The experiment validates the classification performance through a confusion matrix and calculates R^2 and $\mathrm{RMSE_P}$ for both the overall model before classification and the segmented model after classification, thus evaluating the models. The confusion matrix

Table 2 Schematic illustration of confusion matrix concepts

	Predicted positive	Predicted negative
Actual positive	True positive (TP)	False negative (FN)
Actual negative	False positive (FP)	True negative (TN)

2436 | *J. Anal. At. Spectrom.*, 2024, **39**, 2433–2442

This journal is © The Royal Society of Chemistry 2024

is a metric used to measure and assess the performance of machine learning classification models. In this study, the confusion matrix is used to measure the predictive results of SVM.

In a typical binary classification problem, the confusion matrix consists of the following four important elements: (1) True Positives (TP), the model correctly classifies positive samples as positive; (2) False Positives (FP), the model incorrectly classifies negative samples as positive; (3) True Negatives (TN), the model correctly classifies negative samples as negative; and (4) False Negatives (FN), the model incorrectly classifies positive samples as negative. These elements are typically arranged as shown in Table 2.

The confusion matrix can be used to calculate various performance metrics for classification algorithms, including:

(1) Accuracy: accuracy is the proportion of samples correctly classified by the model to the total number of samples, which was defined as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100\%, \qquad (3)$$

(2) Precision: precision represents the proportion of samples correctly classified as positive by the model among those predicted as positive, which was defined as:
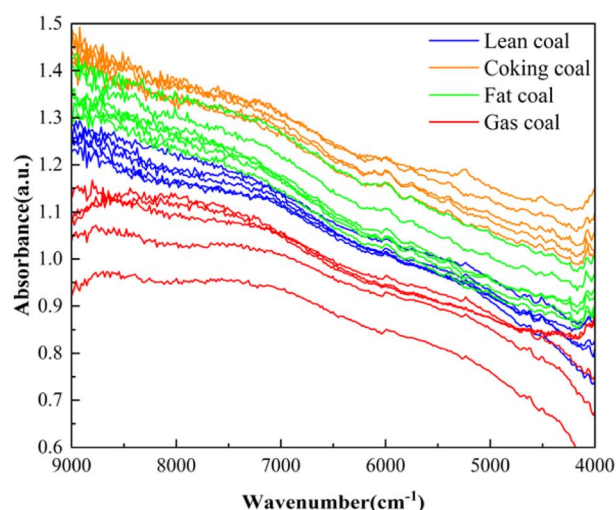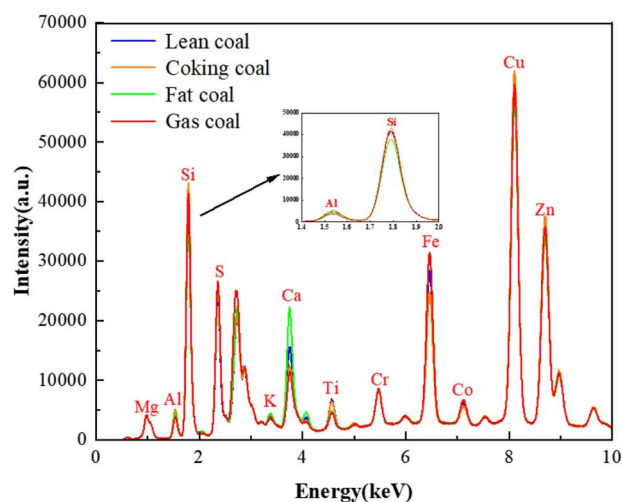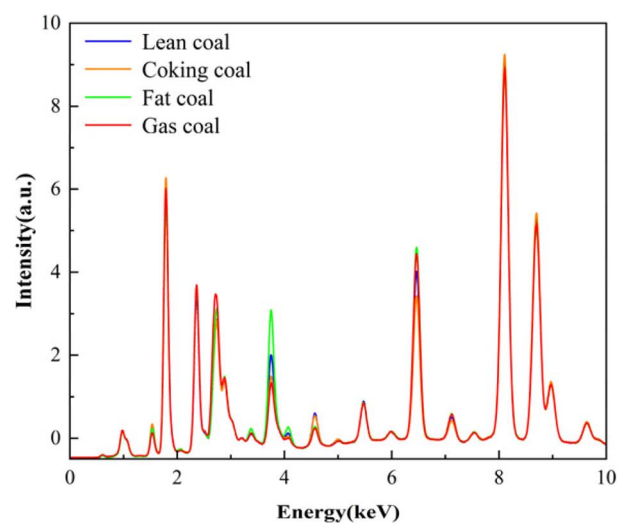

(a)
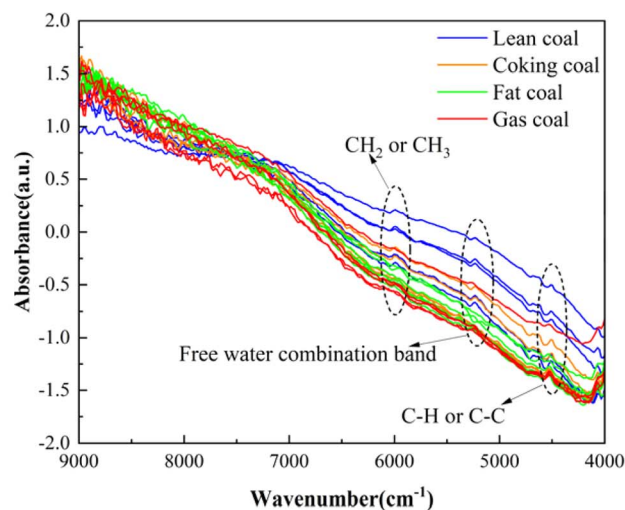
(b)

Fig. 2 NIRS spectra of lean, coking, fat and gas coal, (a) the raw NIRS spectra, (b) the NIRS spectra after pre-processing.

(a)

(b)

Fig. 3 XRF spectra of lean, coking, fat and gas coal, (a) the raw XRF spectra, (b) the XRF spectra after pre-processing.

This journal is © The Royal Society of Chemistry 2024

J. Anal. At. Spectrom., 2024, **39**, 2433–2442 | **2437**

$$\text{Precision} = \frac{TP}{TP + FP} \times 100\%, \quad (4)$$

(3) Recall: recall represents the proportion of samples correctly predicted as positive by the model among those that are actually positive, which was defined as:

$$\text{Recall} = \frac{TP}{TP + FN} \times 100\%, \quad (5)$$

(4) F_1 score: F_1 score is a metric that combines precision and recall, which was defined as:

$$F_1 = \frac{2 \times (\text{precision} \times \text{recall})}{\text{precision} + \text{recall}} \times 100\%, \quad (6)$$
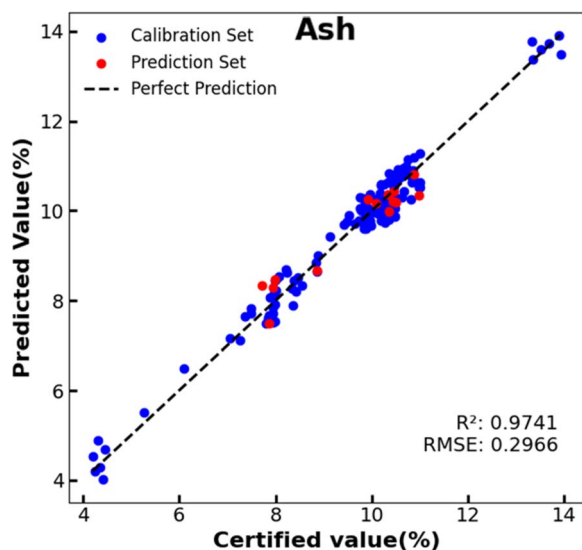
3 Results and discussion

3.1 NIRS and XRF spectra of coal

Fig. 2 and 3 display typical NIRS and XRF raw spectra of lean coal, coking coal, fat coal and gas coal, as well as the spectra after SG smoothing and SNV pre-processing. In Fig. 2, distinctive absorptions are evident in the NIRS spectra of each coal type at specific wavelengths, notably around 6000 cm^{-1}, 5200 cm^{-1}, and 4400 cm^{-1}. These peaks may originate from the asymmetric stretch of CH$_2$ or CH$_3$, free water combination band, and asymmetric stretch of C–H or symmetric bend of C–C.[24]

From Fig. 3, it can be observed that there are significant differences in the spectral line intensities of inorganic elements in the XRF spectra of these four types of coal. The spectral line intensities of Fe are higher in the gas coal spectrum, those of Ca and Fe are higher in the fat coal spectrum, those of Si, Cu, Zn, and Ti are higher in the coking coal spectrum, while the spectral line intensity of Ti is higher in the lean coal spectrum. Additionally, elements such as Mg, Al, K, Cr, and Co show relatively small differences in intensity in the spectra. These significant differences in the spectral line intensities of inorganic elements

Fig. 4 Comparison between the predicted and certified values of A$_{ad}$ by the overall prediction model.

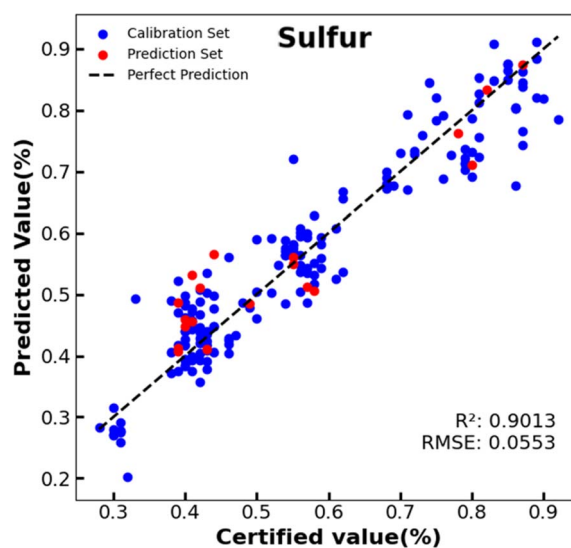**Fig. 6** Comparison between the predicted and certified values of S by the overall prediction model.

Fig. 5 Comparison between the predicted and certified values of V$_{ad}$ by the overall prediction model.

Table 3 Comparison of predictive performance of PLS models using different spectra

Method	Properties	R^2	RMSE$_P$ (%)
NIRS	Ash	0.768	0.887
	Volatile matter	0.754	3.1872
	Sulfur	0.800	0.471
XRF	Ash	0.861	0.738
	Volatile matter	0.873	0.943
	Sulfur	0.825	0.158
NIRS-XRF	Ash	0.974	0.297
	Volatile matter	0.942	0.757
	Sulfur	0.901	0.055

2438 | *J. Anal. At. Spectrom.*, 2024, **39**, 2433–2442

This journal is © The Royal Society of Chemistry 2024

reflect the different compositions of gas, fat, coking, and lean coal.

3.2 Results of the overall model

The NIRS-XRF fusion spectra of 130 training set coal samples were utilized as independent variables, while their certified values of A_{ad}, V_{ad}, and S were respectively recorded as dependent variables. The Partial Least Squares (PLS) method was employed to construct an overall model. Subsequently, the NIRS-XRF fusion spectra of the coal samples in the test set were inputted into the established overall model for prediction. The results are presented in Fig. 4–6. The findings indicate that R^2 of A_{ad}, V_{ad}, and S, based on the overall model, are 0.974, 0.942, and 0.901 respectively, with corresponding $RMSE_P$ of 0.297%, 0.757%, and 0.055%.

Compared to single-spectrum techniques, the combination of two complementary spectroscopic methods can provide more comprehensive and accurate information about samples. Using fused spectra can lead to improved classification or prediction results.[25] Table 3 compares the predictive performance of PLS models using single spectra *versus* fused spectra. It can be seen that the model using fused spectra shows significant improvements across all metrics compared to the single-spectrum model.

3.3 Results of the classification model

Fig. 7 shows the confusion matrices for the classification results of the coal samples into different coal types using SVM on the training set (a) and the test set (b). Here, 1 represents lean coal, 2 represents coking coal, 3 represents fat coal, and 4 represents gas coal. From this confusion matrix, it can be seen that 130

coal samples in the training set were correctly classified, with a classification accuracy of 100%; and 55 coal samples in the test set were correctly classified, with a classification accuracy of 96.49%. The evaluation metrics of SVM on the test set are shown in Table 4. The results indicate that high classification accuracy can be achieved using only spectral data, suggesting that spectral information indeed captures the differences between different coal types. In this experiment, coal was classified into four major categories for further analysis.

The comparison between the predicted values and the certified values of ash, volatile matter, and sulfur for the coal samples in the test set using the corresponding PLS classification models is shown in Fig. 8–10. The specific data are listed in Table 5. It can be observed that the average R^2 for ash, volatile matter, and sulfur predicted by these classification models for the four types of coal samples are 0.992, with corresponding average $RMSE_P$ of 0.062%, 0.080%, and 0.008%, respectively. Compared to the overall model, the R^2 of the classification models have increased by 1.85%, 5.31%, and 10.10%, while the $RMSE_P$ have decreased by 0.24%, 0.68%, and 0.05%. After classification, the prediction performance for lean coal is the best, with R^2 for ash, volatile matter, and sulfur increasing by

Table 4 Evaluation metrics of SVM on test set%

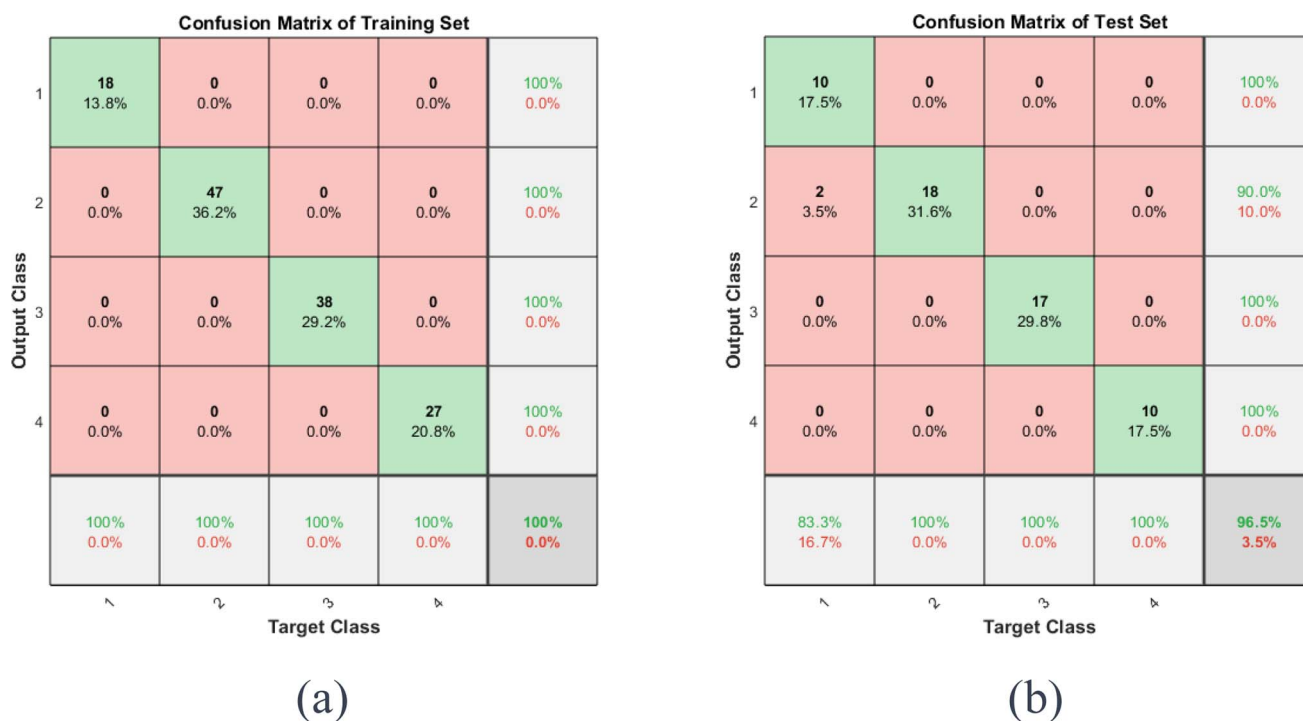| Classification algorithm	Accuracy	Precision	Recall	F_1 score
SVM	96.49	97.50	95.83	96.41

(a) (b)

Fig. 7 Classification confusion matrix for training set (a), test set (b) coal samples. (1–4 values correspond to lean coal, coking coal, fat coal, and gas coal).
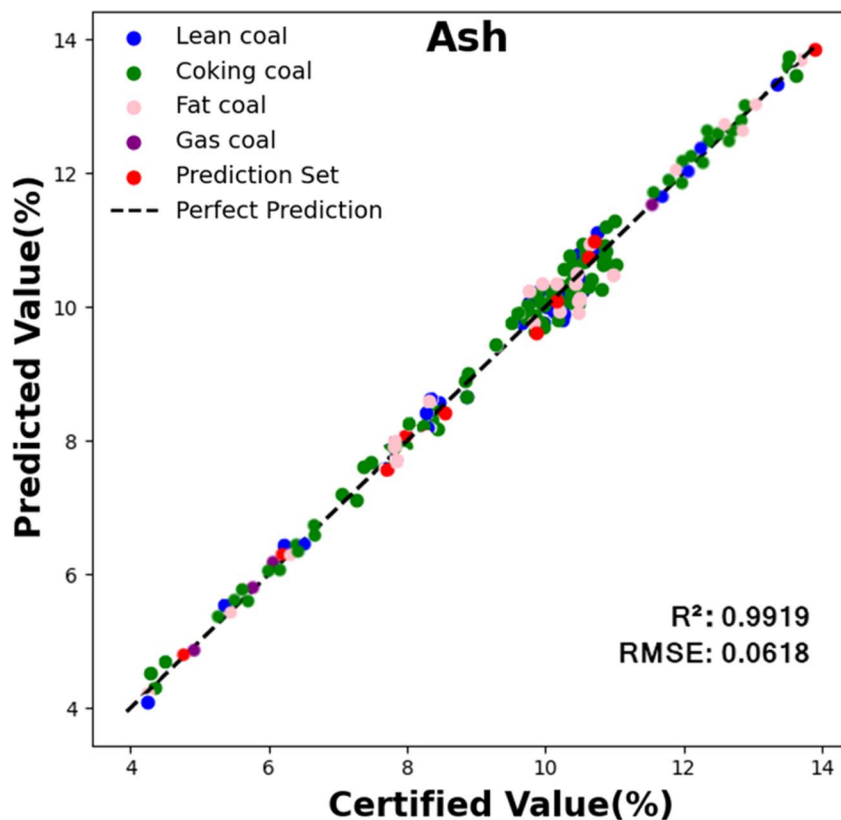
Fig. 8 The comparison between the predicted values and the certified values of A_{ad} by the classification model.

2.36%, 5.94%, and 10.65% compared to the overall model, and $RMSE_P$ decreasing by 0.25%, 0.71%, and 0.05%. These experimental results confirm that in coal quality analysis, the prediction accuracy of various indicators of coal quality is significantly improved by performing coal type identification and classification prediction compared to the overall model.

4 Summary

In this study, we investigated the modeling approach and classification algorithm for predicting properties of complex coal types based on NIRS-XRF coal quality analysis technology. Firstly, we employed PLS regression for overall modeling
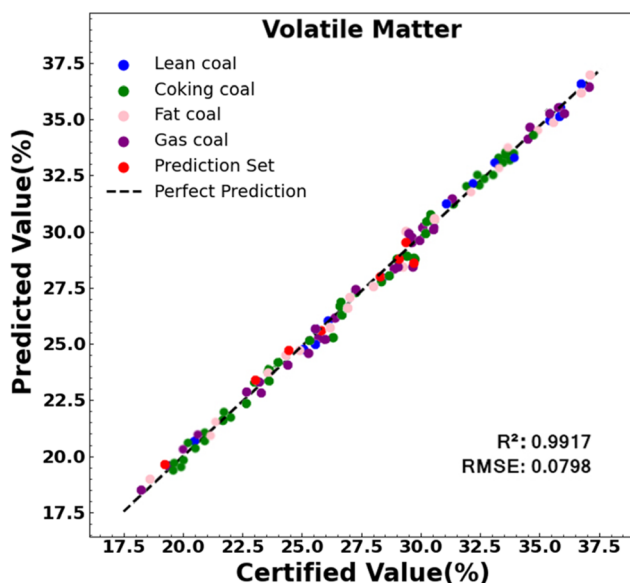

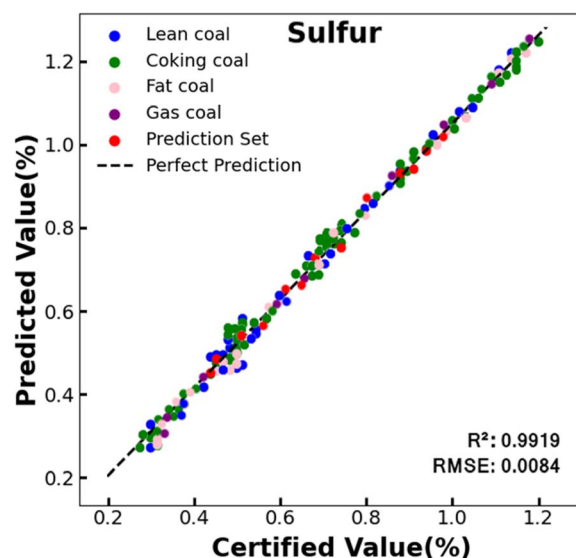**Fig. 9** The comparison between the predicted values and the certified values of V_{ad} by the classification model.

Fig. 10 The comparison between the predicted values and the certified values of S by the classification model.

2440 | *J. Anal. At. Spectrom.*, 2024, **39**, 2433–2442

This journal is © The Royal Society of Chemistry 2024

Table 5 Coal quality PLS analysis results after support vector machine classification

Model type	Classifications	Properties	R^2	RMSE$_P$ (%)
Overall model	—	Ash	0.974	0.297
		Volatile matter	0.942	0.757
		Sulfur	0.901	0.055
Classification model	Lean coal	Ash	0.997	0.045
		Volatile matter	0.998	0.048
		Sulfur	0.997	0.005
	Coking coal	Ash	0.976	0.044
		Volatile matter	0.986	0.113
		Sulfur	0.997	0.009
	Fat coal	Ash	0.995	0.066
		Volatile matter	0.996	0.084
		Sulfur	0.996	0.010
	Gas coal	Ash	0.999	0.092
		Volatile matter	0.986	0.075
		Sulfur	0.978	0.010

performance, with R^2 values of 0.974, 0.942, and 0.901 for ash, volatile matter, and sulfur, respectively, and corresponding RMSE$_P$ values of 0.297%, 0.757%, and 0.055%. Secondly, to explore better prediction results, we applied SVM for the classification of complex coal samples. SVM was applied to the fusion spectra, classifying coal into fat coal, coking coal, lean coal, and gas coal. The results showed that the classification accuracy of this method reached 96.49%. Finally, we validated the performance of this strategy using the test set samples. The classification model achieved R^2 values of 0.992 for ash, volatile matter, and sulfur, with corresponding RMSE$_P$ values of 0.062%, 0.080%, and 0.008%, respectively. Compared to modeling the complete sample set, the classification-based PLS regression model yielded better results, with R^2 values increasing by 1.85%, 5.31%, and 10.10%, and RMSE$_P$ values decreasing by 0.24%, 0.68%, and 0.05%. Among these, the prediction performance for lean coal was the best after classification. Therefore, applying SVM classification before PLS regression can improve the accuracy of the analysis. The study demonstrated that classifying coal types before PLS regression for coal quality prediction is feasible, and the predicted values of assay indicators obtained can meet the needs of coal chemical enterprises. In future work, we will further investigate other classification algorithms and modeling approaches based on NIRS-XRF fusion spectra analysis to be applicable to various types of coal, and expand the analysis properties, such as calorific value, to achieve more comprehensive and accurate coal quality analysis.

Data availability

The data supporting this study's findings are available from the corresponding author upon reasonable request.

Author contributions

Zhedong Zhang is the main author and is responsible for the conceptualization, methodology, software, data curation and formal analysis of the experiment, and the writing of original draft. Yan Zhang and Lei Zhang are responsible for conceptualization and methodology, while Lei Zhang is also responsible for writing review and editing and project management. Jiaxuan Li and Rui Gao is responsible for investigation, validation and supervision. Yang Zhao, Zefu Ye, Peihua Zhang and Zhujun Zhu are responsible for providing the necessary resources for the experiment. Wangbao Yin and Suotang Jia are responsible for project management and funding acquisition, while Wangbao Yin is also responsible for writing reviews and editing.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

References

1 Q. Yang, L. Zhang, S. Zou and J. Zhang, *Energy Pol.*, 2020, **139**, 111360.
2 K. Liu, C. He, C. Zhu, J. Chen, K. Zhan and X. Li, *TrAC, Trends Anal. Chem.*, 2021, **143**, 116357.
3 L. Cheng, L. Zhao and J. Wang, *Coal Economic Research*, 2022, **42**, 57–62.
4 P. Chen, *Chinese coal properties, classification and use*, Chemical Industry Press Pub, 2nd edn, 2007.
5 M. Shealy and J. P. Dorian, *Energy Pol.*, 2010, **38**, 2116–2122.
6 C. S. Lim and D. A. Abernethy, *Appl. Radiat. Isot.*, 2005, **63**, 697–704.
7 B. Yang, *Inner Mongolia Coal Economy*, 2021, **15**, 140–141.
8 S. Pan, *Science and Technology Innovation and Application*, 2014, vol. 20, pp. 69–70.

This journal is © The Royal Society of Chemistry 2024

J. Anal. At. Spectrom., 2024, **39**, 2433–2442 | **2441**

9 M. Borsaru and Z. Jecny, *Appl. Radiat. Isot.*, 2001, **54**, 519–526.

10 W. Gu, L. Zhang and M. Dong, *Plasma Sci. Technol.*, 2022, **24**, 080101.

11 S. Yao, J. Mo and J. Zhao, *Appl. Spectrosc.*, 2018, **72**, 1225–1233.

12 S. Sheta, M. S. Afgan and Z. Hou, *J. Anal. At. Spectrom.*, 2019, **34**, 1047–1082.

13 H. Yan, Z. Yang, H. Lu and Y. Ju, *Appl. Mech. Mater.*, 2014, **556**, 575–579.

14 S. Wang, Y. Zhao, R. Hu, Y. Zhang and X. Han, *Chin. J. Anal. Chem.*, 2019, **47**, e19034–e19044.

15 N. Begum, D. Chakravarty and B. S. Das, *Int. J. Coal Prep. Util.*, 2019, **42**, 979–985.

16 R. Gao, J. Li, S. Wang, Y. Zhang, L. Zhang, Z. Ye, Z. Zhu, W. Yin and S. Jia, *Anal. Methods*, 2023, **15**, 1674–1680.

17 D. Bolortuya, P. Zuzzan and M. V. Gustova, *Phys. Part. Nucl. Lett.*, 2013, **10**, 723–726.

18 D. G. Hicks, J. E. Reilly and D. W. Kopenaal, *American Institute of Physics Conference Series*, 1981, vol. 70, pp. 454–455.

19 K. F. Ma, *Coal Quality Technology*, 2019, **2**, 32–35.

20 J. Li, R. Gao, Y. Zhang, S. Wang, L. Zhang, W. Yin and S. Jia, *Chemosensors*, 2023, **11**, 363.

21 Y. Wang, M. Yang, G. Wei, R. Hu, Z. Luo and G. Li, *Sens. Actuators, B*, 2014, **193**, 723–729.

22 S. Zhong, *Min. Surv.*, 2020, **48**, 65–68.

23 T. Leng, F. Li, Y. Chen, L. Tang, J. Xie and Q. Yu, *Meat Sci.*, 2021, **180**, 108559.

24 S. A. Fysh, D. A. J. Swinkels and P. M. Fredericks, *Appl. Spectrosc.*, 1985, **39**, 354–357.

25 E. Borràs, J. Ferré, R. Boqué, M. Mestres, L. Aceña and O. Busto, *Anal. Chim. Acta*, 2015, **891**, 1–14.