

Identification of coal characteristics by near-infrared spectroscopy: Machine learning predictions and experimental validations

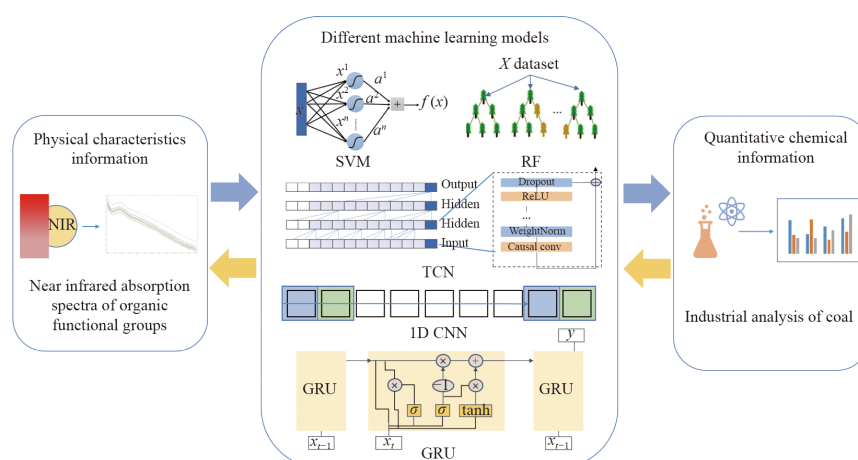
AN Haiquan^{*,*}, LIU Zhen, LI Ye, PENG Baozi^{#,*}

National Institute of Clean and Low-Carbon Energy, Beijing 102209, China

Abstract: In this study, near-infrared spectroscopy was used along with various machine learning algorithms to develop a fast and reliable framework for identifying coal components, which significantly improved the accuracy and efficiency of predictions. Support vector machine (SVM), random forest (RF), temporal convolutional network (TCN), one-dimensional convolutional

neural network (1D CNN), and gated recurrent unit (GRU) algorithms were used for regression modeling of coal properties. Among the various models considered herein, the TCN model was shown to provide the lowest Mean Absolute Error (MAE) value, which was 0.505. In addition, the RF model demonstrated the lowest Root Mean Square Error (RMSE) value of 0.618, indicating its robustness in prediction on the test set. The TCN model, on the other hand, had the lowest Coefficient of Variation (CV) value of 0.042, which demonstrated its good generalization ability and stability. Furthermore, the developed single-output models revealed that different models have different predictive effects on coal components. For example, it was found that the TCN model is optimal for predicting ash content, while the RF model is the most accurate model for predicting fixed carbon and low calorific value. Taking into account the quantity of model parameters, computation time, and accuracy analysis, the results showed that the RF, 1D CNN and TCN models are more efficient. Finally, the influence of key characteristic peaks on prediction results was revealed through SHAP value analysis. The results showed a relationship between absorptivity intensity and composition of organic and inorganic values in coal to near-infrared light, suggesting important characteristic spectral ranges for ash, moisture, and fixed carbon. This research not only provides a new method for rapid detection in the coal industry, but also serves as a reference for component analysis in other fields.

Key words: machine learning; coal characteristics; NIR; rapid testing



The importance of rapid testing of coal properties lies in its multifaceted contributions to the coal industry^[1–2]. Primarily, it enables precise process optimization in coal-fired power plants and industrial facilities by providing real-time quality data to adjust combustion parameters, thereby

enhancing efficiency, reducing emissions, and realizing cost savings. Moreover, rapid testing ensures stringent quality control, enabling producers to maintain consistency in product quality and comply with customer specifications. Additionally, rapid testing facilitates cost reduction by

Received: 2025-09-10; Revised: 2025-10-13

[#]: These authors contributed equally to this work.

^{*} Corresponding author. E-mail: haiquan.an.c@ceic.com, pbz1029@163.com.

DOI: 10.1016/S1872-5813(25)60611-6 CSTR: 32157.14.2097-213X.2025.JFCT.0157

本文的英文电子版由 Elsevier 出版社在 ScienceDirect 上出版 (<http://www.sciencedirect.com/science/journal/18725813>)

identifying low-quality coal early in the supply chain and optimizing blending processes. Efficient management of the coal supply chain is further facilitated by timely information on coal, enhancing logistics, storage, and blending processes^[3]. Ultimately, by optimizing energy production efficiency and minimizing environmental impacts, rapid testing of coal properties contributes to the sustainability and profitability of the coal industry.

In recent years, the advancement and application of rapid coal property testing technologies have significantly contributed to the efficiency and sustainability of the coal industry. Among the prominent techniques, Raman spectroscopy, Laser-Induced Breakdown Spectroscopy (LIBS), X-ray Fluorescence (XRF) spectroscopy, and Near-Infrared (NIR) spectroscopy have emerged as key methodologies for rapid coal analysis. Raman spectroscopy^[4], characterized by its ability to probe molecular vibrational modes, offers high sensitivity and specificity in identifying molecular structures in coal samples. This technique provides valuable insights into properties such as aromaticity and degree of carbonization. However, the need for sophisticated instrumentation and expertise to interpret the data may pose challenges in its widespread application. LIBS^[5], which uses laser-induced plasma emission to determine elemental composition, is notable for its rapid multi-elemental analysis capability. By providing non-destructive and minimally sample-preparation-dependent measurements, LIBS excels in elemental characterization. Nevertheless, limitations in accurate trace element quantification and sensitivity to surface changes may impact its accuracy in complex coal matrices. XRF spectroscopy^[6], which is based on fluorescent X-rays emitted by an X-ray irradiated sample, enables rapid and non-destructive elemental analysis. Its notable advantages include rapid data acquisition and suitability for elemental quantification. However, the efficiency of XRF may be limited by challenges in accurately assessing certain parameters, such as moisture content, particularly in heterogeneous coal samples. NIR spectroscopy^[7–8], which works on the principle of NIR light absorption, offers rapid, non-destructive, and real-time monitoring capabilities to assess coal properties. Its versatility allows the determination of various parameters including moisture content, calorific value, and sulfur content. However, limitations may arise in the quantification of certain parameters, especially at lower concentrations. These technologies have contributed to the advancement of coal properties testing. Among them, compared with other technologies, NIR technology has the advantages of high equipment safety, low cost and fast detection. However, coal has a complex composition as well as strong absorption and scattering effects on NIR spectra, and it is difficult for traditional modeling methods to meet the requirements of high-precision prediction.

The integration of NIR technology with advanced machine learning algorithms can significantly improve the speed and accuracy of NIR-based detection models.

Machine learning techniques, such as artificial neural networks and support vector machines, can improve accuracy by learning complex spectral patterns and correlations from large datasets. An enhanced partial least squares (PLS) regression method, based on support vector machine (SVM) classification, was proposed in Ref. [9] for rapid analysis of coal properties using NIR reflectance spectroscopy (NIRS). This method was tested on 199 coal samples and demonstrated a significant improvement in the predictive ability of the post-classification model. This improvement was primarily attributed to SVM classification prior to PLS regression, which allowed the model to better adapt to different types of coal. Consequently, this method enhanced the accuracy of NIRS for the rapid analysis of coal^[9]. A fast method for analyzing coal properties based on deep learning and visible-near-infrared spectroscopy was proposed in Ref. [10] using 158 coal samples. Subsequently, a new two-layer limit learning machine (DF-RTLM) algorithm was proposed and used to construct a coal characterization model. Experimental results showed that this model can effectively predict the composition of coal. Also, for testing 158 coal samples, the method proposed in Ref. [10] required only about 5 h and \$100, while traditional proximate and ultimate analysis methods required approximately 300 h and \$6000^[10]. A coal proximate analysis method based on spectroscopy and deep learning was proposed in Ref. [11], and a deep learning model, named DR_TELM, was constructed. The experimental results showed that the DR_TELM model can achieve high prediction accuracy and fast test speed in determining coal proximate analysis parameters such as moisture, ash, volatile matter, fixed carbon and heating value. Specifically, the DR_TELM model achieved an RMSE of 0.533 and R² of 0.981 for water content, an RMSE of 1.833 and R² of 0.989 for ash, an RMSE of 1.111 and R² of 0.990 for volatile matter, and an RMSE of 1.808 and R² of 0.990 for fixed carbon^[11].

The capability and accuracy of predicting coal properties through NIR spectroscopy have been enhanced by the aforementioned machine learning algorithms. However, the correlation between coal attributes and spectral data is not adequately accounted for by these algorithms. As a result, the prediction of coal properties is still not accurate enough. In this paper, chemometrics and multiple machine learning algorithms are used to develop a framework for predicting coal properties based on NIR spectroscopy and machine learning models for rapid and accurate measurement of coal composition^[12]. In addition, specific components of coal produce characteristic peak spectra when irradiated in the NIR, and their contribution to model predictions has been investigated by machine learning interpretation methods, thus allowing the mechanistic nature of the prediction models to be demonstrated. Specifically, a NIR spectrometer was used to collect coal spectra, and then the collected laboratory measurements were merged to create a comprehensive database of experimental results on coal spectra. On this basis, algorithms such as SVM,

Random Forest (RF), TCN, One Dimensional Convolutional Neural Network (1D CNN), and Gated Recursive Unit (GRU) were used as processing methods for sequence data, aiming to develop an efficient model for coal detection. Through rigorous model comparison and experimental validation, the machine learning framework combining NIR spectroscopy with sequence data processing algorithms was demonstrated to provide significant improvements. This approach not only simplifies the modeling challenges but also maximizes the prediction accuracy. In addition, employing model analysis techniques also helps to delve into the properties of the model and coal, thus improving the interpretability and generalizability of the model. In conclusion, the framework presented in this study for the rapid detection of coal characteristics, leveraging near-infrared spectroscopy and machine learning techniques, offers both precision and speed in assessing coal composition, demonstrating notable scalability.

1 Method

Chemical detection methods are mainly used by traditional coal enterprises for coal analysis, which are known for their high accuracy and reliability. However, systematic sampling and preparation of coal samples are necessary for these methods, which leads to increased analysis time and reduced timeliness of coal detection. This does not match the current demands of enterprise intelligence and low-carbon development. In recent years, NIR spectroscopy technology has gained importance as a rapid detection technique for liquids and solids. It is characterized by advantages such as straightforward implementation, fast analysis speed, non-destructive sampling, and minimal equipment investment and maintenance costs. However, conventional modeling methods like principal component analysis (PCA) and partial least squares (PLS) regression struggle to accurately

predict coal data with complex compositions and high absorption. Therefore, a rapid coal detection framework is proposed in this paper, based on NIR spectroscopy, a coal database, and machine learning algorithms, as depicted in Figure 1. By simultaneously performing laboratory analysis and spectral measurements on coal samples, a spectral coal database is first established. The spectra are then preprocessed and feature peaks are screened using chemometrics. After that, a variety of machine-learning algorithms are used to model and select the best prediction model to achieve rapid detection of coal indicators.

1.1 Coal sample and spectral dataset

After a long period of testing and data collection, the coal characterization data and corresponding NIR spectroscopy results of 268 coal samples were summarized into a coal spectroscopy dataset. The coal sample used in this experiment is a bituminous coal, sourced from the Shaanxi province in China. First, coal was ground into test samples with a diameter of 0.2 mm. And then, a portion of the coal samples was subjected to laboratory testing, in accordance with the Chinese standard “GB/T 212—2008 Industrial Analysis Methods for Coal”. The resulting data from this analysis provided a detailed breakdown of the component composition. The rest of samples were tested for spectral acquisition using a ball-integrated measurement fast Fourier transform near infrared spectrometer (model GALAXY-Q3000). Specifically, the equipment collected absorption spectra ranging from 4000 to 10000 cm^{-1} at wavenumber 1559. The sample filling and spectral acquisition process was repeated three times and their average values were taken as the spectra of the samples. Finally, the collected coal samples were spectroscopically measured using laboratory measurements to obtain coal characteristic indices, which, together with the spectral data, comprise the coal spectral-characterization index dataset.

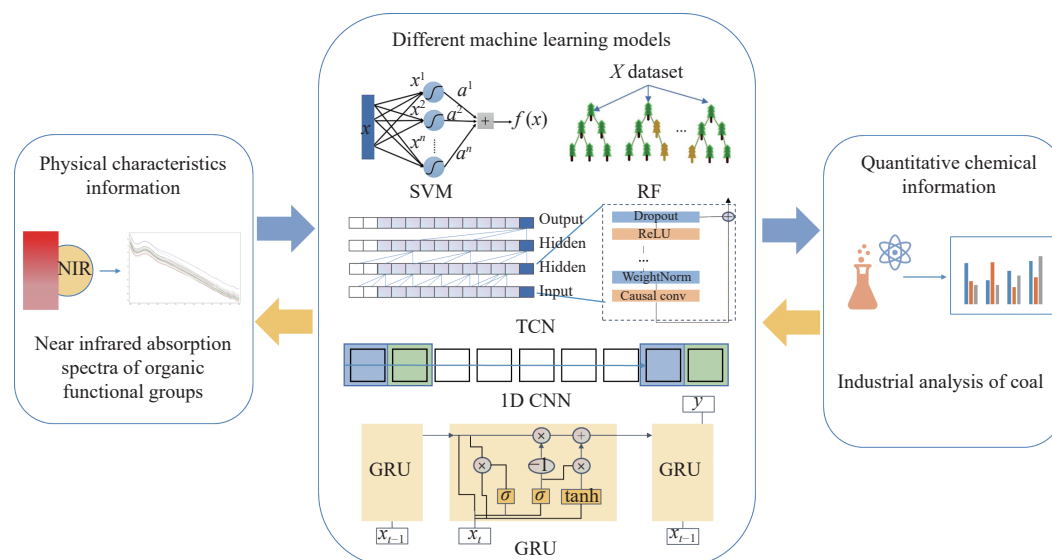


Fig. 1 The work flow of rapid coal detection with NIR spectra and machine learning models

Unlike other datasets commonly used for machine learning modeling, the NIR spectroscopy dataset of coal is distinguished by its limited sample size, extensive feature set, inherent error in coal composition metrics, and the potential influence of spectral temporal ordering on prediction results.

The ranges of key coal characterization data, such as moisture with air received basis (M_{ar}), ash content with air dried basis (A_{ad}), fixed carbon with dry ash free basis (FC_{daf}) and low calorific value with air received basis (Q_{net}), are shown in Figure 2. In this paper, they are modeled and validated as output parameters of the machine learning model.

As shown in Figure 2, the data distributions of A_{ad} , M_{ar} , FC_{daf} and Q_{net} exhibit relative symmetry, with means close to their medians, indicating the absence of pronounced skewness. The symmetry in the distributions of A_{ad} and M_{ar} suggests that the dataset is balanced, with values evenly distributed around the mean. The means of these datasets are closely aligned with their respective medians, indicating a lack of significant skewness. This is advantageous for many statistical analyses and machine learning models that assume normality or symmetric distributions.

The interquartile ranges (IQR) for A_{ad} and M_{ar} were 1.59 and 1.42, respectively, suggesting moderate variability in these datasets. Such variability implies a certain level of data spread without being overly broad, potentially indicating stable and predictable variable characteristics. On the other hand, it can be observed that FC_{daf} and Q_{net} have

more focused distributions, since their IQR is smaller than one (0.8 and 0.82, respectively). This indicates that most of the data for these two variables are gathered tightly around the central values, which in turn suggests less dispersion and possibly more stable performance/behavior.

Although the skewness is not very high and variability is moderate, the existence of outliers (not visible in the violin plots but suggested by the long tails) could still impact modeling. Outliers may unduly influence the results of statistical analyses and machine learning models; therefore, their removal or treatment should be thoughtfully addressed.

1.2 Machine learning models

Based on the thorough analysis of the objectives of this research, SVMs, Tree Models, Convolutional Neural Networks and Recurrent Neural Networks were chosen as modeling tools^[13–15]. Among them, the CNN and RNN algorithms could effectively solve the modeling and prediction problems of graphical data and sequence data. Specifically, SVM, RF, TCN, 1D CNN and GRU algorithms were used for regression modeling of coal properties. The best prediction results were selected by comparing these algorithms.

The SVM is a robust supervised learning model, known for its superior generalization capacity and suitability for small-sample data.

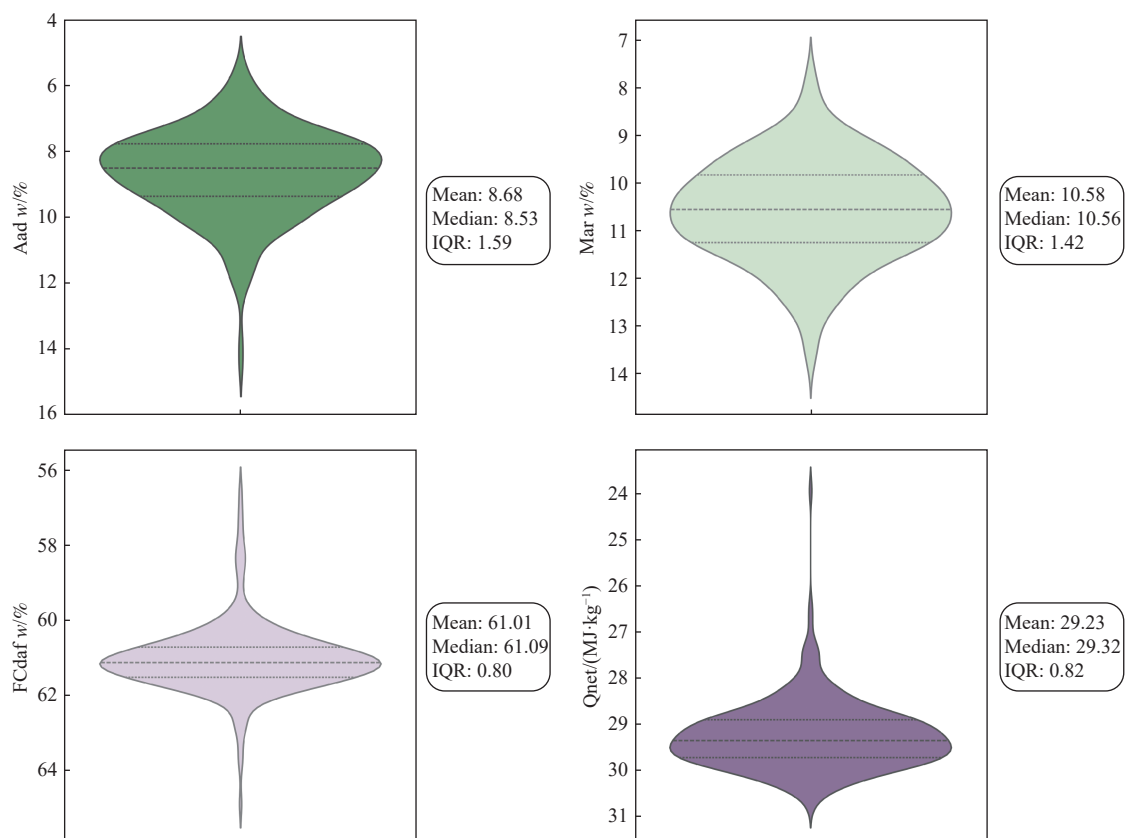


Fig. 2 Violin plots of coal characteristics

The SVM performs data classification by identifying the optimal hyperplane, thereby effectively handling high-dimensional data and reducing the risk of overfitting. Furthermore, the ability of SVM to handle nonlinear problems through kernel function mapping ensures that the model maintains robust performance on the test set. RF is a comprehensive learning method that enhances prediction accuracy by constructing multiple decision trees and voting on them. This method can handle a significant number of features, perform feature selection automatically, and exhibit robustness against outliers and noise. The strengths of RF include its superior generalization ability, its capacity to avoid overfitting, and its outstanding performance in both classification and regression tasks. The 1D CNN is a deep learning algorithm commonly used for computer vision tasks such as image recognition, classification, detection, and segmentation. The properties of CNN, such as parameter sharing and local connectivity, enhance its computational efficiency and effectiveness in capturing local image features, such as texture and edges. CNNs can learn hierarchical features of an image by incorporating multiple convolutional and pooling layers, moving from low-level features at the base to high-level features at the top^[15–16]. This paper utilizes a 1D CNN model to perform spectral analysis due to the characteristics of NIR spectra. Additionally, a TCN model is adopted to perform the prediction of coal spectra by considering the sequence of spectral results^[17]. TCN is a neural network optimized for sequential data processing, particularly adept at handling time series data. By using convolutional layers, the TCN can capture local features and patterns across different time scales, making it effective for modeling and predicting time series data. Compared with traditional Recurrent Neural Networks (RNNs), the TCN shows better performance and stability when handling long sequences of data. This effectively addresses the issues of gradient disappearance and computational inefficiency that are commonly associated with RNNs.

To evaluate the performance of ML models, the coefficient of mean absolute error (MAE), root mean square error (RMSE) and coefficient of variation (CV) were calculated with prediction dataset. MAE is the mean absolute error and better reflects the reality of the errors in the predicted values. RMSE measures the root-mean-square difference between the predicted and true values and represents the average degree of deviation between the predicted and true values. The coefficient of variation is an absolute value, serving as a measure of data dispersion. The magnitude of this coefficient is influenced not only by the degree of variability in the variable values but also by the mean size of these values. Notably, as the CV value approaches 1, the model error diminishes. This concept differs from the conventionally employed coefficient of variation in statistical analyses. The equations used to calculate these performance metrics are as follows^[18–19]:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |Y_i - y_i| \quad (1)$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_i - y_i)^2} \quad (2)$$

$$\text{CV} = 1 - \frac{\text{RMSE}}{\bar{Y}} \quad (3)$$

where Y_i indicates the true value of the output result in the dataset, y_i indicates the prediction value of the output result in the dataset, $\bar{Y}$ indicates the average value of Y_i and i indicates the i^{th} sample in the dataset.

Machine learning models are often viewed as a black box, and the processes and principles underlying their output indicators are challenging to elucidate. This paper employs the SHAP method to interpret the single-output 1D CNN model, in conjunction with spectral peak analysis, to identify the most influential feature peaks that affect the predictions of the model on individual output indicators. SHAP (SHapley Additive exPlanations) is an explanatory framework for machine learning model predictions, grounded in the Shapley value from game theory^[19–20]. It assigns an impact score to each feature, thereby enhancing the interpretability of model outputs and facilitating a deeper understanding of model predictions.

In this study, although the number of coal samples is limited, the employed NIR-machine learning framework demonstrates significant scalability. Models such as TCN and RF exhibit robust fitting capabilities for spectral variations, and SHAP analysis has elucidated the physical correlation between spectra and coal quality. Consequently, this framework possesses methodological universality. In future research, it can be adapted to other coal types through sample expansion or transfer learning without necessitating model reconstruction.

1.3 Hyperparameter tuning

Hyperparameter tuning is a central technique in the field of machine learning and deep learning. This process entails the adjustment of unfixed parameters during model training, which significantly influence the behavior and performance of the model. Hyperparameters are predefined parameters in the model architecture that are created before starting model training and subsequently remain fixed during the training phase. Unlike model parameters, such as weights and biases in neural networks, the values of hyperparameters are typically not learned directly through the model training process. Instead, they necessitate empirical selection or application of a specific strategy.

In this paper, a stochastic search was applied to find the best combination of hyperparameters, and a cross-validation approach was chosen to assess the impact of hyperparameters on the generalization ability of the model.

To enhance data utilization, mitigate overfitting, and ensure consistent performance estimation, the K-fold cross-validation for model assessment is employed. Given the limited sample size in this study, a smaller K value for K-fold cross-validation was to select to minimize data

distribution bias, manage model complexity, and conserve computational resources.

2 Results and discussion

2.1 Data set preprocessing and feature spectrum screening

Spectral dataset of coal samples is presented in Figure 3. Preprocessing steps are necessary before building a machine learning model based on NIR spectroscopy. Strategies such as spectral range selection, baseline correction, scatter correction, normalization, data standardization, and outlier detection and management are employed to enhance the predictive accuracy of the model. Specifically, a second-order polynomial Savitzky-Golay filter is used to remove noise and normalize the spectra, while the Mahalanobis distance method is employed to remove outliers from the spectral dataset. The optimal Savitzky-Golay filter window size is 81, which represents a range of 5% to 15% of the total data points, and yields superior denoising results. For outlier removal, the chi-square distribution is employed to set the Mahalanobis distance threshold, with a chosen confidence interval of 95%. These methods ensure the quality of the dataset, thereby providing a robust data foundation for subsequent modeling.

2.2 Comparison of results of multi-output prediction models

In order to predict the multiple coal indicators, a multi-output machine learning model has been developed. The prediction results of these models on the test set are presented in Table 1. As can be seen, the difference between different models is small concerning regression evaluation metrics. A detailed analysis reveals that: First, the TCN model outperforms the others in terms of MAE, recording an MAE value of 0.505. This result highlights that the TCN model exhibits a reduction in the absolute deviation between predicted and actual values, thereby indicating superior fitting accuracy. Second, in terms of RMSE, the RF model achieves the highest performance with an RMSE value of 0.618. This implies that the RF model gives better prediction accuracy compared to other models in the test set and does not show significant deviations between the predicted and actual values, which indicates the robustness of the model. Finally, in relation to CV, the TCN model also emerges as the best, with a CV value of 0.042. These results suggest that the predictions made by the TCN model on the test set have lower dispersion, which reflects the generalization ability and commendable stability of the model. In conclusion, it can be inferred that among the machine learning models, the RF model exhibits superior prediction performance; whereas among the deep learning models, the Time Convolutional Network (TCN) model demonstrates the most outstanding performance.

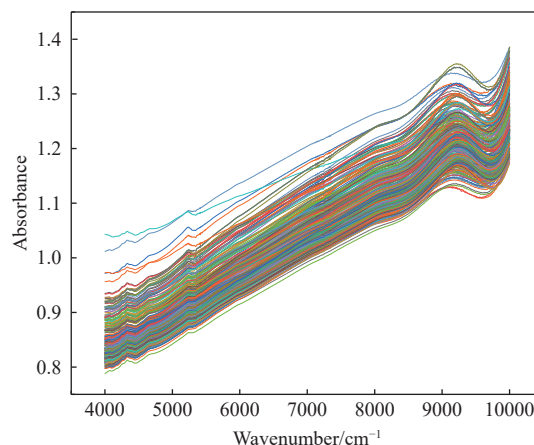


Fig. 3 Spectral dataset of coal samples

Table 1 Multi-output machine learning model evaluation metrics

Model	MAE	RMSE	CV
Multi-output SVM	0.521	0.652	0.047
RF	0.509	0.618	0.047
1D CNN	0.506	0.633	0.045
GRU	0.507	0.637	0.046
TCN	0.505	0.633	0.042

Upon analyzing various machine learning and deep learning models, it is evident that the latter possesses a notable ability to automatically extract features. In tasks involving the modeling of long-term dependency relationships in spectral sequences, deep learning models exhibit a significant advantage over their machine learning counterparts. This superiority contributes to the enhanced stability of predictions, consequently reducing CV. Additionally, the extensive parameter set of deep learning models enables them to effectively manage complex non-linear mapping relationships, decrease the linear accumulation of prediction bias, and achieve a smaller MAE. On the other hand, the RF, a quintessential tree-structured model, leverages the Bagging ensemble method to significantly reduce model variance. Notably, even in scenarios with limited sample data, RF successfully avoids grossly erroneous predictions, making it exceptionally proficient in terms of RMSE performance.

In summary, the TCN model demonstrates an advantage in evaluating the predictive performance. This suggests that NIR spectral data of coal exhibit a certain degree of sequence dependence with wavelength. However, considering that the spectral feature peaks are primarily related to the chemical composition of the coal, the TCN model does not significantly outperform the other models. Furthermore, machine learning methods such as multi-output SVM and RF models exhibit lower resource consumption and increased calculation speed during both training and prediction stages. These tasks can be completed within seconds. Conversely, deep learning models including 1D CNN, GRU, and TCN require more resources for training and prediction. Their computational efficiency

decreases when the number of model parameters reaches approximately 100000. In this study, the limited amount of accumulated coal spectral data and the comprehensive preprocessing of the dataset make the machine learning models significantly more efficient. However, anticipating future practical applications, it is expected that as coal spectral datasets continue to accumulate and expand, and given the limitations of thorough preprocessing for spectral data collected in real time during online prediction, deep learning models, particularly TCN models, are likely to exhibit superior overall performance and efficiency.

2.3 Comparison of results of single-output prediction models

The use of multi-output prediction models can simplify the modeling process and decrease the number of required models. As evidenced by the results of the previous section, these multi-output models yield superior predictive outcomes. However, the evaluation metrics for multi-output models are tailored to multiple output indicators, which does not allow for clear visualization of the impact of model predictions on individual output indicators. Consequently, single-output prediction models are constructed in this section for each output indicator and an analysis of their predictive efficacy is conducted.

The comparative evaluation of the single-output prediction model for four distinct output indicators is presented in Table 2. It should be noted that the predictive efficacy of multiple models varies across different output indicators, as elaborated upon below.

As demonstrated in Table 2, various models exhibit distinct prediction outcomes on different indicators. However, there is no universal rule governing these effects, and the overall differences in prediction performance are marginal. In particular, for the prediction of ash content, the TCN model performs better than the others. Also, for the prediction of moisture content, both 1D CNN and TCN models show superior results, while for the prediction of FC_{daf} and Q_{net} , the RF model provides the most accurate performance. Consequently, in the endeavor of concurrently modeling and predicting via NIR spectroscopy and machine learning models, it is advisable to choose different machine learning and deep learning models for distinct prediction indicators to achieve superior results.

Furthermore, the model evaluation parameters for each

predicted value do not differ significantly and maintain the same order of magnitude. However, due to the significant variations in the dimensionless values of each predicted value, it suggests that the prediction accuracy of the model is different for each value. For instance, the value of FC_{daf} is approximately five times that of A_{ad} , but the Mean Absolute Error (MEA) of the model for predicting FC_{daf} is lower. This indicates that the prediction performance of the model on FC_{daf} surpasses its prediction on A_{ad} . Further examination reveals that light within the NIR region corresponds to the vibrational frequencies of certain substances, particularly those with hydrogen atoms such as O–H, N–H and C–H bonds. These chemical bonds can absorb specific wavelengths of NIR light during their vibration, leading to distinct absorption spectra. Consequently, inorganic substances like metal oxides and hydrous minerals may also exhibit absorption characteristics in NIR spectroscopy, although these features are typically less pronounced than those of organic molecules. The composition of coal primarily consists of ash content, which is a complex mixture of various metal oxides and characteristic inorganic substances. The inherent moisture within coal exists in forms such as free water, hydrates, and internal water. The presence of FC_{daf} often coincides with elements or functional groups like hydrogen and oxygen, significantly impacting the Q_{net} value of coal. In conclusion, models utilizing near-infrared spectroscopy demonstrate greater accuracy in predicting FC_{daf} and Q_{net} compared to A_{ad} and M_{ar} .

2.4 Analysis of model efficiency

In the preceding sections, we presented and analyzed prediction indicators for both multi-output prediction and single-output prediction models. This analysis was done with the aim of ascertaining the effectiveness, accuracy and predictability of the model. However, it is crucial to note that the algorithmic complexities can vary significantly during the modeling process. Such variations lead to distinct differences in model structure and parameters, which in turn influence the time required for training and prediction. All experiments related to timing were conducted on a standard laptop with the following specifications: an Intel i5-8250U processor, 8 GB of RAM, Windows 10 version 22H2, Python version 3.8.8, TensorFlow version 2.10, and Jupyter Notebook version 6.5.7.

Table 2 Single-output machine learning model evaluation metrics

Model	$A_{ad}/\%$			$M_{ar}/\%$			$FC_{daf}/\%$			$Q_{net}/(MJ \cdot kg^{-1})$		
	MAE	RMSE	CV	MAE	RMSE	CV	MAE	RMSE	CV	MAE	RMSE	CV
SVM	0.725	0.907	0.107	0.482	0.591	0.056	0.444	0.545	0.009	0.433	0.481	0.016
RF	0.701	0.880	0.103	0.592	0.692	0.066	0.438	0.536	0.009	0.365	0.426	0.015
1D CNN	0.689	0.867	0.102	0.461	0.567	0.054	0.456	0.552	0.009	0.398	0.454	0.015
GRU	0.712	0.888	0.104	0.470	0.578	0.055	0.457	0.553	0.009	0.398	0.454	0.015
TCN	0.665	0.843	0.100	0.461	0.567	0.054	0.466	0.556	0.009	0.398	0.454	0.015

The CPU, with four cores and eight threads, was operated at an average usage rate of approximately 75%. Despite the modest specifications of the computer used, a thorough examination of the model parameters, training and prediction times, and prediction indicators provides valuable insights into the efficiency and performance of the model. Notably, when using a similar machine learning or deep learning algorithm, the structure and parameters maintain a consistent order of magnitude for both multi-output and single-output value modeling. Consequently, our discussion and analysis will focus solely on the performance results of the multi-output prediction model, as illustrated in Figure 4.

As shown in Figure 4, the M-SVM model has the fewest parameters and provides the fastest training and prediction times. However, its predictive accuracy is comparatively subpar. The RF model, with a parameter count of 9450, outperforms the GRU model in terms of predictive performance and significantly reduces the computation time, indicating superior efficiency. As a deep learning model, the 1D CNN model exhibits the highest computational speed, even surpassing the RF model, while also providing the best predictive performance. Despite having the most node parameters, the computation time of the TCN model remains lower than that of the GRU model, which is attributed to the use of convolution and pooling operations. In conclusion, for near-infrared spectroscopy coal quality index prediction models, RF, 1D CNN and TCN emerge as the more efficient models.

Furthermore, compared with the 5 h required by DBN-DF-RTELM for 158 samples in Ref. [10], the present RF/1D CNN/TCN framework completes training in 14–63 s and predicts in <1 s per sample on a consumer laptop. This 2–3-order-of-magnitude acceleration is ascribed to (i) rigorous spectral pre-processing that reduces data redundancy and (ii) lightweight architectures tailored to spectroscopic signals, eliminating the need for deep, parameter-intensive networks.

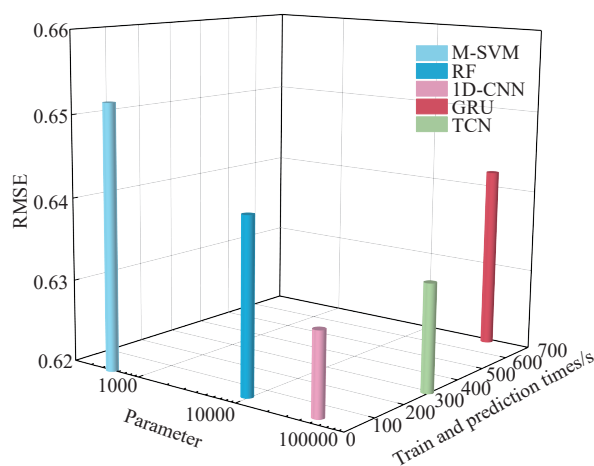


Fig. 4 The efficiency results of the multi-output prediction model

2.5 Analysis of characteristic peak with SHAP

To elucidate the near-infrared spectroscopy-coal index prediction model, the SHAP method was introduced in this study. As mentioned earlier, different organic and inorganic components of coal absorb near-infrared light to varying degrees, producing absorption spectra. This section employs NIR spectral feature analysis techniques to identify spectral peaks. Subsequently, the SHAP method is utilized to compute and illustrate the SHAP values of these characteristic peaks in relation to the prediction results, thereby examining the correlation between these peaks and the predictive metrics of the model. Given that multiple prediction indicators exhibit different levels of NIR light absorption, a single-output 1D CNN model was chosen for simple analysis in this section.

The differential spectrum method was used for spectral analysis of near infrared spectra in the database and identification of characteristic peaks. This process yielded 53 sets of characteristic peak results. The sequence of these characteristic peaks is aligned with the order of spectral wavenumbers in the dataset. By considering the relationship between wavenumber and wavelength, the wavelength range of the characteristic peaks was ultimately established.

In the single-output 1D CNN model, SHAP value analysis was conducted on characteristic peaks for three specific indicators: A_{ad} , M_{ar} and FC_{daf} . These indicators represent the organic and inorganic components in the coal that directly influence the formation of these characteristic peaks. The results of this analysis are presented in Figure 5.

At first, as illustrated in Figure 5, ten characteristic peaks with the highest SHAP values for each of the three output indicators are identified. Subsequently, their corresponding wavelength ranges are calculated and presented in Table 3.

Figure 5 and Table 3 reveal that the characteristic peak 52 has the highest SHAP value across all three single-output models, indicating its major influence on their predictive results. This implies a pronounced absorption of NIR light by ash, moisture, and fixed carbon components in coal within the wavelength range of 1141.35 to 1157.65 nm. The SHAP values indicate that the impact of characteristic peak 6 on predicting A_{ad} and M_{ar} results is only surpassed by peak 52, while its influence on FC_{daf} predictions is behind peaks 52 and 10. Notably, Figure 5 and Table 3 demonstrate significant overlaps in the range of absorption intensities for organic and inorganic substances in coal concerning NIR light. These overlapping peaks are crucial in prediction accuracy. Additionally, the SHAP value for the characteristic peak 46 is notably higher for the A_{ad} model compared to the other models, implying that the ash exhibits a stronger absorption effect on NIR light with wavelengths between 1173.89 and 1179.23 nm. Conversely, other substances display reduced absorption in this range. Consequently, this wavelength band can be identified as a distinct absorption characteristic spectrum for ash.

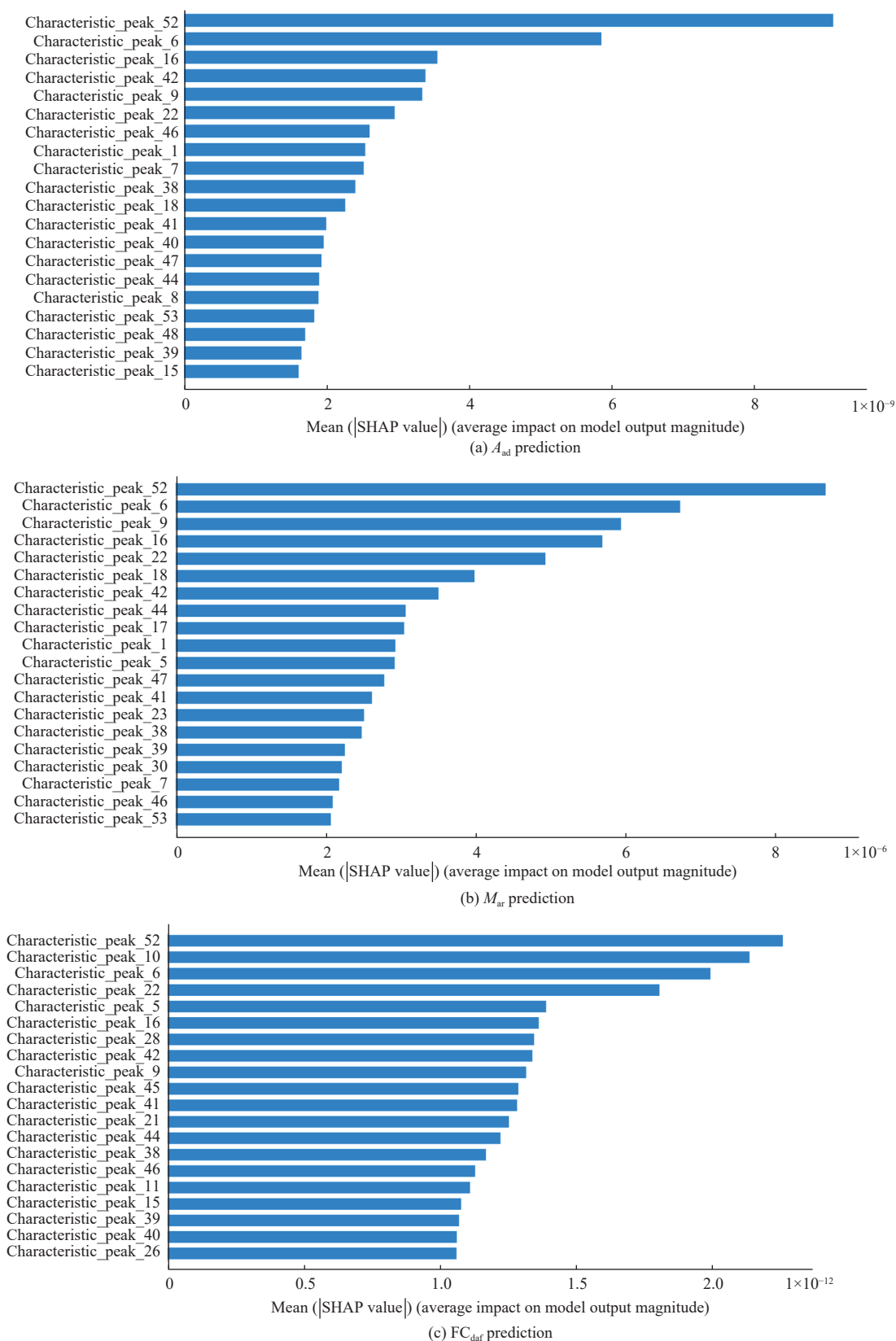


Fig. 5 The characteristic peak with SHAP value for coal prediction

In a similar vein, the characteristic peak 18 (wavelength: 1382.19–1388.10 nm) is uniquely associated with moisture, while the characteristic peaks 5 and 28, with wavelengths of 1914.79–1920.46 nm and 1267.38–1270.48 nm, respectively,

are distinctively linked to fixed carbon. Finally, coal, a prevalent energy and industrial raw material, has diverse applications. However, its complex structure has continuously posed challenges for comprehensive understanding.

Through the analysis of the NIR spectral features presented in this section, combined with existing research on the molecular structure of coal, there is potential to accurately predict the composition of coal materials in the future.

Further comprehensive research was undertaken, amalgamating the analysis of coal composition with the absorption bands of NIR^[21]. The characteristic peak 52 was identified as potentially the secondary harmonic of the aromatic C–H stretching vibration of alkylbenzene and the combination peak of non-bonded hydroxyl. Notably, the principal characteristic peaks influencing FC_{daf} were consistently located proximal to the secondary harmonic region of compounds such as acyl chloride C=O, carboxylic acid C=O, and ethyl acetate C=O. In predicting A_{ad} , the characteristic peaks were primarily distributed around the

absorption peaks of silanol O–H stretching vibration and organic phosphine P–H stretching vibration. A dominant role was played by free O–H characteristic peaks and their combination peaks in the prognostic analysis of M_{ar} .

It is imperative to acknowledge that the preceding analysis primarily hinges on the absorption bands of isolated or collective molecular chemical bonds, as discerned through NIR. Considering coal's heterogeneous and unstable composition as a solid, a comprehensive exploration of the relationship between its spectral characteristics and composition becomes paramount. Consequently, future research endeavors will be concentrated on delving deeper into this area to elucidate the precise correlations between coal's spectral features and its compositional constituents.

Table 3 Characteristic peak wavelength range that mainly affects the prediction results

Characteristic peak	Wavenumber range/cm ⁻¹		Wavelength range/nm	
52	8638.20	8761.57	1141.35	1157.65
46	8480.14	8518.69	1173.89	1179.23
45	8457.01	8476.28	1179.76	1182.45
44	8418.46	8453.15	1182.99	1187.87
42	8314.37	8364.48	1195.53	1202.74
38	8198.71	8225.70	1215.70	1219.70
28	7871.02	7890.29	1267.38	1270.48
22	7308.16	7350.57	1360.44	1368.33
18	7204.07	7234.91	1382.19	1388.10
17	7180.94	7200.21	1388.85	1392.58
16	7107.69	7177.08	1393.32	1406.93
10	5473.08	5488.50	1821.99	1827.12
9	5392.12	5457.66	1832.29	1854.56
7	5315.02	5338.15	1873.31	1881.46
6	5226.35	5303.45	1885.56	1913.38
5	5207.07	5222.50	1914.79	1920.46
1	3996.54	4069.79	2457.13	2502.16

3 Conclusions

In this study, NIR spectroscopy technology was effectively integrated with various machine learning algorithms to develop a fast and accurate framework for coal component prediction. The experimental results demonstrated a significant improvement in accuracy and efficiency by this framework for coal component prediction. Notably, the TCN model exhibited the highest fitting accuracy among all models, as evidenced by its lowest MAE value of 0.505. Conversely, the RF model showcased robustness in prediction accuracy on the test set, as indicated by its minimum RMSE value of 0.618, thereby implying minimal deviation between predicted and actual values. Furthermore, the TCN model also showed the lowest CV value of 0.042, suggesting that its predictions on the test set were highly consistent, thus indicating commendable generalization ability and stability of the

model. In addition, through independent modeling of multiple output indicators, it was observed that different models exhibit different predictive performances for distinct output indicators. For instance, the TCN model demonstrated optimal performance in predicting ash content. Both the 1D CNN and TCN models exhibited superior results in predicting moisture content, while the RF model was the most accurate in predicting fixed carbon and low calorific value. A comprehensive analysis considering model parameters, training and prediction times, along with prediction indicators, identified the RF, 1D CNN and TCN models as more efficient.

The SHAP value analysis identified the characteristic peaks that had the greatest impact on the model prediction results. Peak 52 exhibited the highest SHAP value in all three single-output models, which emphasizes its important role in influencing the prediction results. This suggests that the absorption of NIR light by ash, moisture, and fixed carbon components in coal is particularly pronounced at

wavelengths ranging from 1141.35–1157.65 nm. In addition, SHAP values showed distinct relationships. For example, peak 46 (wavelength: 1173.89–1179.23 nm) was uniquely associated with ash, peak 18 (wavelength: 1382.19–1388.10 nm) was specifically associated with moisture, and peaks 5 and 28 (wavelength: 1914.79–1920.46 nm and 1267.38–1270.48 nm, respectively) were clearly related to fixed carbon. Through the analysis of NIR spectral characteristics, in conjunction with existing research on the molecular structure of coal, we can accurately predict the composition of coal substances in the future. This study not only offers a novel method for rapid detection in the coal industry, but also serves as a reference for rapid component analysis in other sectors. Future research will focus on further optimizing the model structure, enhancing prediction accuracy, and investigating additional machine learning algorithms to predict coal properties.

Declaration of Competing Interest

The authors declare no conflict of interest.

References

- [1] LI Q W, YAO G H. Improved coal combustion optimization model based on load balance and coal qualities[J]. *Energy*, 2017, 132: 204–212.
- [2] WANG K C, ZHANG J, SHANG C, et al. Operation optimization of Shell coal gasification process based on convolutional neural network models[J]. *Appl Energy*, 2021, 292(17): 116847.
- [3] AN H Q, LIU Z, SUN K D, et al. A machine learning framework for intelligent prediction of ash fusion temperature characteristics[J]. *Fuel*, 2024, 362: 130799.
- [4] WEI Q L, ZHANG L D, SONG C F, et al. Quantitative detection of dithiocarbamate pesticides by surface-enhanced raman spectroscopy combined with exhaustive peak-seeking method[J]. *Anal Methods*, 2021, 13(12): 1479–1488.
- [5] ZHAO Y, ZHANG L, ZHAO S X, et al. Review of methodological and experimental LIBS techniques for coal analysis and their application in power plants in China[J]. *Front phy*, 2016, 11(6): 217–230.
- [6] WEN Z P, LIU H T, ZHOU M Q, et al. Explainable machine learning rapid approach to evaluate coal ash content based on X-ray fluorescence[J]. *Fuel*, 2023, 332: 125991.
- [7] ZHAO Z, LI B, LIU W, et al. A qualitative recognition method based on Karhunen–Loeve decomposition for near-infrared (NIR) and mid infrared (MIR) spectroscopy analysis[J]. *Anal Methods*, 2020, 12: 112–121.
- [8] WANG S H, ZHAO Y, HU R, et al. Analysis of near-infrared spectra of coal using deep synergy adaptive moving window partial least square method based on genetic algorithm[J]. *Chin J Anal Chem*, 2019, 47(4): e19034–e19044.
- [9] WANG Y S, YANG M, WEI G, et al. Improved PLS regression based on SVM classification for rapid analysis of coal properties by near-infrared reflectance spectroscopy[J]. *Sens Actuators, B*, 2014, 193: 723–729.
- [10] XIAO D, LE B T. Rapid analysis of coal characteristics based on deep learning and visible-infrared spectroscopy[J]. *Microchem J*, 2020, 157: 104880.
- [11] XIAO D, YAN Z L, LI J, et al. Rapid proximate analysis of coal based on reflectance spectroscopy and deep learning[J]. *Spectrochim Acta, Part A*, 2023, 287: 122042.
- [12] BÜYÜKKANBER K, HAYKIRI-ACMA H, YAMAN S. Calorific value prediction of coal and its optimization by machine learning based on limited samples in a wide range[J]. *Energy*, 2023, 277: 127666.
- [13] MATIN S S, CHELGANI S C. Estimation of coal gross calorific value based on various analyses by random forest method[J]. *Fuel*, 2016, 177: 274–278.
- [14] FENG Q H, ZHANG J Y, ZHANG X M, et al. Proximate analysis based prediction of gross calorific value of coals: A comparison of support vector machine, alternating conditional expectation and artificial neural network[J]. *Fuel Process Technol*, 2015, 129: 120–129.
- [15] ZHOU G Z, GUO Z Q, SUN S M, et al. A CNN-BiGRU-AM neural network for AI applications in shale oil production prediction[J]. *Appl Energy*, 2023, 344: 121249.
- [16] LEE J W, LEE J H. Simultaneous extraction of intra- and inter-cycle features for predicting lithium-ion battery's knees using convolutional and recurrent neural networks[J]. *Appl Energy*, 2024, 356: 122399.
- [17] ZHU J J, SU L C, LI Y W. Wind power forecasting based on new hybrid model with TCN residual modification[J]. *Energy and AI*, 2022, 10: 100199.
- [18] SUN X T, ZHANG P, SHEN Z W, et al. Investigation on spectral standardization among multi-channel of an on-line near-infrared spectrometer[J]. *Vib Spectrosc*, 2021, 113: 103206.
- [19] RUSSO C, CASTRO A, GIOIA A, et al. Improving the sediment and nutrient first-flush prediction and ranking its influencing factors: An integrated machine-learning framework[J]. *J Hydrol*, 2023, 616: 128842.
- [20] HOMAFAR A, NASIRI H, CHELGANI S C. Modeling coking coal indexes by SHAP-XGBoost: Explainable artificial intelligence method[J]. *Fuel Commun*, 2022, 13: 100078.
- [21] BRUNO T J, SVORONOS P D N. CRC Handbook of Fundamental Spectroscopic Correlation Charts[M]. Boca Raton: CRC Press, 2005.

基于近红外光谱预测煤质特性: 机器学习预测与实验验证

安海泉^{#,*}, 刘 臻, 李 焱, 彭宝仔^{#,*}

(北京低碳清洁能源研究院, 北京 102209)

摘 要: 在本研究中, 基于近红外光谱与多种机器学习算法, 构建了一套快速、可靠的煤质组分预测框架, 显著提升了预测的精度与效率。研究采用支持向量机(SVM)、随机森林(RF)、时间卷积网络(TCN)、一维卷积神经网络(1D CNN)以及门控循环单元(GRU)对煤质指标进行回归建模。在所用模型中, TCN 模型的平均绝对误差(MAE)最低, 为 0.505; RF 模型的均方根误差(RMSE)最低, 为 0.618, 表明其在测试集上的预测稳健性最佳; TCN 模型的变异系数(CV)最低, 为 0.042, 显示出良好的泛化能力与稳定性。此外, 单输出模型结果显示, 不同指标对应最优算法各异: TCN 对灰分预测最佳, RF 则在固定碳和低位发热量预测中精度最高。综合考虑参数量、计算耗时与精度, RF、1D CNN 与 TCN 模型效率更优。通过 SHAP 值分析揭示了关键特征峰对预测结果的影响, 表明煤中有机与无机组分对近红外光的吸收强度与其含量相关, 明确了灰分、水分及固定碳的重要光谱区间。该研究不仅为煤炭工业提供了新的快速检测方法, 也为其他领域的组分分析提供了参考。

关键词: 机器学习; 煤质特性; 近红外光谱; 快速检测

中图分类号: TQ533; O657.33; TP181

文献标识码: A

作者简介:



安海泉(1988), 博士, 高级工程师, 主要从事煤化工智能化技术研究工作, haiquan.an.c@ceic.com。



彭宝仔(1978), 博士, 正高级工程师, 主要从事煤和生物质的高效低碳利用研究工作, pbz1029@163.com。