

Intelligent analysis of direct coal liquefaction diesel components by near-infrared spectroscopy

WANG Xiwu^{1,2}, LI Haowei¹, QI Zhendong^{1,2}, WANG Xingbao^{1,*}, FENG Jie^{1,*}, ZHU Yimeng³, LI Wenyong¹

¹ State Key Laboratory of Clean and Efficient Coal Utilization, Taiyuan University of Technology, Taiyuan 030024, China;

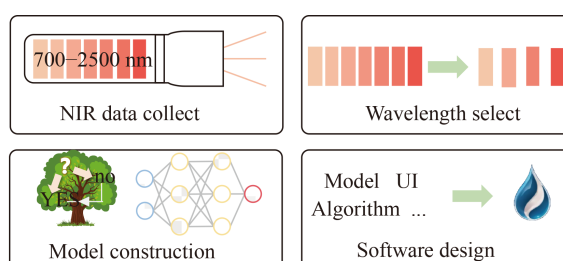
² Ordos Branch of China Shenhua Coal-to-Liquid Chemical Co., Ltd., Ordos 017209, China;

³ Kyland Technology Co., Ltd., Beijing 100144, China

Abstract: Diesel accounts for over 60% of the products derived from direct coal liquefaction (DCL). Compared to petroleum-based diesel, DCL diesel exhibits a cetane number ranging from 30 to 40, which fails to meet the automotive diesel standard requirement of ≥ 45 . Therefore, rapid and accurate analysis of its chemical composition is crucial for property optimization to meet fuel specifications by component blending. Though traditional methods like gas chromatography offer high accuracy, they are

unsuitable for rapid online analysis under industrial conditions. Near-infrared (NIR) spectroscopy can provide advantages in rapid, non-destructive analysis. Its application however, is limited by the complexity of spectral data interpretation. Machine learning (ML) is effective method for extracting valuable information from spectra and establishing high-precision prediction models. This study integrates NIR spectroscopy with ML to construct a spectral-composition database for DCL diesel. Feature extraction was performed using the correlation coefficient and mutual information methods to screen key wavelength variables and reduce data dimensionality. Subsequently, the predictive performance of three ML models—Lasso, SVR and XGBoost—was compared. Results indicate that excluding spectral data with absorbance greater than 1 significantly enhances model accuracy, increasing the test set R^2 from 0.85 to 0.96. After feature extraction, the optimal number of wavelength variables was reduced to 177, substantially improving computational efficiency. Among the models evaluated, the SVR-MI-0.9 model, based on mutual information feature selection, demonstrated the best performance, achieving training and test set R^2 values both exceeding 0.98. This model enables precise prediction of paraffin, naphthene, and aromatic hydrocarbon contents. This research provides a robust methodology for intelligent online quality monitoring. An intelligent NIR spectroscopy data analysis software was independently developed based on the established model. Compared with comprehensive two-dimensional gas chromatography, the software reduced the analysis time by over 98%, with an absolute prediction error below 0.2%. Thus, rapid analysis of DCL diesel components was successfully realized.

Key words: direct coal liquefaction diesel; real-time spectral detection; machine learning; feature extraction; component prediction



Direct coal liquefaction (DCL) diesel is characterized by its low sulfur and nitrogen compound content, as well as a low condensation point. However, challenge of DCL diesel is existent due to the lower cetane number than the standard for automotive diesel^[1-2]. The combustion performance of direct coal liquefaction diesel is determined by its

composition^[3-5]. Currently, the compositional analysis of direct coal liquefaction diesel primarily relies on traditional laboratory testing methods, such as gas chromatography and mass spectrometry^[6-7]. Although accurate, these methods suffer from long analysis times, high costs, and difficulties in achieving online monitoring^[8-9]. Near-infrared (NIR)

Received: 2025-08-12; Revised: 2025-09-08

Foundation items: Supported by National Natural Science Foundation of China (U24B6018, 22178243).

* Corresponding author. Tel & Fax: +86-0351-6018453, E-mail: wangxingbao@tyut.edu.cn, fengjie@tyut.edu.cn.

DOI: 10.1016/S1872-5813(25)60620-7 CSTR: 32157.14.2097-213X.2025.JFCT.0171

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>)

spectroscopy technology with wavelength 780–2526 nm has demonstrated significant potential in fuel property detection due to its rapidity, non-destructiveness, and thus is very suitable for online analysis^[10–14]. However, NIR spectroscopy also exhibits challenges such as weak absorption signals, band overlapping, and complex spectral interpretation^[15–16]. Moreover, NIR requires establishing a reliable model between spectra and target parameters, a task that becomes challenging and prone to inaccuracy with insufficient sample data.

In recent years, machine learning (ML) has made significant progress in spectral data analysis. By establishing high-dimensional nonlinear models, ML can more accurately correlate spectral data with chemical composition^[17]. For instance, methods such as support vector machines (SVM), random forests (RF), and deep learning (DL) have been successfully applied to determine the compositional information of petroleum products^[18–21]. Al-Kaf et al.^[22] employed a single-layer neural network model combined with NIR spectroscopy to predict six key properties of diesel, including cetane number, total aromatic content, and density, achieving small errors between predicted and measured values. These ML-based NIR models facilitate rapid, high-precision prediction of diesel properties, overcoming the limitations of traditional methods and laying a technical foundation for intelligent online analysis. Nevertheless, current research primarily focuses on petroleum-based and biodiesel^[23–25]. Studies on machine learning-assisted NIR analysis for direct coal liquefaction diesel are still in their early stages. In particular, systematic methodologies for deciphering the complex relationship between NIR spectra and the chemical composition of DCL diesel have not yet been established.

This study aims to integrate NIR spectroscopy with ML to establish a rapid prediction model for the composition of DCL diesel. The relative contribution of spectral information to each chemical component of the diesel is evaluated using both the correlation coefficient and mutual information methods, thereby complementing traditional property evaluation with quantitative composition analysis.

By constructing a NIR spectral-composition database for DCL diesel, key features are screened through relevant feature extraction methods. ML algorithms are then adopted to predict the group composition of the diesel based on the NIR spectral data with ultimate goal of rapid analysis of DCL liquid products. The research outcomes can be applied to monitor the production process and control the quality of DCL diesel, enabling a transition from offline laboratory testing to industrial online analysis. Furthermore, by monitoring the diesel composition in real time, process parameters can be dynamically adjusted to optimize product performance, thereby enhancing production efficiency and product quality.

1 Data collection and modeling methodology

1.1 Data collection

The DCL diesel used in this study was obtained from the Ordos coal-to-liquids plant of Shenhua coal-to-liquids and Chemicals Co., Ltd. The diesel fuel was first fractionated using a PILODIST fully automatic true boiling point distillation unit. The distillation was carried out in 10 °C increments per fraction, and the detailed data from the fractionation are presented in Table 1. To further expand the dataset, fractions with larger quantities were sub-sampled at 400 mL intervals. The suffixes (1), (2) and (3) in the procedural numbering represent samples from the lower, middle, and higher temperature ranges of a given fraction, respectively. Comprehensive two-dimensional gas chromatography coupled with mass spectrometry (GC×GC-MS) was employed to analyze the various fractions of the direct coal liquefaction diesel. High-purity nitrogen was used as the carrier gas. The oil samples were diluted with dichloromethane (0.1 µL of sample in 1 mL) prior to GC×GC-MS analysis. The chromatographic system was equipped with a DB-1 (15 m) and a BPX-50 (2.7 m) column. The injector temperature was set at 280 °C, with an injection volume of 1 µL and a split ratio of 10:1.

Table 1 Composition data of various distillation fractions in direct coal liquefaction diesel

Distillation range /°C	Volume /mL	Alkanes /%	Cycloalkanes /%	Aromatics /%	Distillation range /°C	Volume /mL	Alkanes /%	Cycloalkanes /%	Aromatics /%
140–150	86	—	—	—	200–210(1)	400	7.0	91.4	1.6
150–160	125	10.7	88.1	1.2	200–210(2)	400	7.3	91.1	1.6
160–170	313	9.4	89.6	1.0	200–210(3)	136	11.0	86.8	2.2
170–180(1)	400	7.3	91.7	1.0	210–220	396	14.9	82.4	2.7
170–180(2)	145	10.4	88.8	0.7	220–230	531	12.6	84.7	2.7
180–190(1)	400	7.6	91.7	0.7	230–240	396	20.0	76.5	3.5
180–190(2)	184	6.0	93.2	0.8	240–250	324	22.7	74.7	2.6
190–200	194	6.1	92.4	1.5	250–260	237	29.1	68.8	2.1

The temperature program was as follows: initial temperature held at 60 °C for 5 min, ramped to 120 °C at 3 °C/min and held for 10 min, then ramped to 200 °C at 2 °C/min and held for 2 min, and finally ramped to 280 °C at 4 °C/min and held for 10 min. The FID detector was operated under the following conditions: hydrogen flow rate of 40 mL/min, air flow rate of 400 mL/min, and a detector temperature of 285 °C. The MS detector parameters were: ion source temperature 230 °C, interface temperature 280 °C, electron energy 70 eV, acquisition rate 50 Hz, with a 5 min solvent delay applied. The modulator used a 300 ms hot pulse and a 6 s modulation period. Data were processed using GC Image software, and mass spectral matching was performed using the NIST database. The quantitative information for each fraction was based on the FID detector signal, and the relative content percentages of the components were determined by normalization.

Near-infrared (NIR) spectra were collected using an Agilent Cary 5000 UV-vis-NIR spectrophotometer. The wavelength of the spectra is ranged of 700–2500 nm with a 1 nm intervals. The baseline was corrected using air in dual-beam mode, and undiluted fractionated oil samples were analyzed directly using a 2–4 mL quartz cuvette. To construct a mixed mixed-sample dataset, a base mixture was first prepared combining 1 mL of each of two different oil fractions. Then, three other fractions were sequentially added to this base mixture in 0.5 mL increments for spectral measurement. A total of 238 NIR spectra were tested in this work.

1.2 Model establishment and evaluation methods

The ML models were constructed with the absorbance at each NIR wavelength as the input variable and the oil group composition as the output variable to establish a mapping relationship for the intelligent analysis of diesel group composition. Given the regression task characteristics involving continuous target variables, three representative machine learning models were selected for comparative analysis in this study: Lasso regression^[26], SVR^[27–28] and eXtreme Gradient Boosting (XGBoost)^[29]. Among them, Lasso regression can enhance generalization capability by introducing L1 regularization to reduce redundant features and noise interference. SVR, based on kernel functions that map data to a high-dimensional space, is particularly suitable for handling small-sample, high-dimensional nonlinear regression problems. XGBoost integrates multiple decision trees using a gradient boosting framework, combined with regularization methods and parallel computing, which demonstrate excellent robustness when processing high-dimensional, complex nonlinear data.

A systematic comparison of the three models was conducted to identify the most suitable method for NIR spectral analysis. When building models using datasets obtained from different feature selection methods, the hyperparameter tuning ranges were kept consistent across

all models. This approach eliminated the influence of the hyperparameter search space and allowed for an objective evaluation of the wavelength selection methods. The Lasso and SVR models were implemented using the scikit-learn package, with the radial basis function (RBF) used as the kernel for SVR. The XGBoost model was built using the XGBoost package. During the model training process, the optimal combination of hyperparameters was identified via an exhaustive grid search over the parameter space defined in Table 2.

The dataset was partitioned into training and test sets in an 8:2 ratio using a random splitting method, designated for model training and performance evaluation. In ML model construction, four metrics are commonly employed, including mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2). Among them, MSE reflects the model's error by calculating the square of the difference between predicted and true values. MSE amplifies the impact of points with large prediction errors during evaluation, thereby imposing a stricter penalty on significant model deviations. RMSE is the square root of MSE, sharing the same units as the target variable, which facilitates intuitive interpretation. However, RMSE may perform less favorably for models with uniformly small errors. As to MAE, it directly computes the absolute difference between predicted and true values, making it insensitive to outliers. Yet, MAE cannot distinguish between large and small errors and is non-differentiable during optimization, preventing the use of gradient-descent-based algorithms. R^2 is commonly used to measure the closeness of the model's predictions to the true values, which can provide an intuitive indication of the model's goodness-of-fit. Given the continuously differentiable mathematical property, MSE is highly suitable for gradient descent optimization algorithms. Its use during optimization directs the model's focus towards significant errors, thus helping to improve the overall accuracy of the model. Therefore, MSE was selected as the loss function for the model training process. As a dimensionless, standardized metric, R^2 clearly and intuitively indicates the proportion of variance in the data explained by the model. It is suitable for comparing performance across different models or datasets, providing a comprehensive evaluation of the goodness-of-fit for the model. Consequently, R^2 was chosen as the primary metric for evaluating model quality.

Table 2 Hyperparameter optimization ranges by model

Model	Hyperparameter
Lasso	alpha: 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000
SVR	c: 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000 epsilon: 0.01, 0.1, 0.2, 0.3 gamma: 0.001, 0.01, 0.1, 1, 10, 100
XGBoost	n_estimators: 10, 20, 30, 50, 100, 200, 300 max_depth: 1, 3, 5, 7, 9 learning_rate: 0.001, 0.01, 0.1, 0.2, 0.3

1.3 Wavelength selection methods

The raw NIR spectral data are characterized by high-dimensional, comprising absorbance values at thousands of wavelengths. This data includes both effective information reflecting molecular vibrations and irrelevant signals such as instrument noise and environmental interference^[17]. Directly using the raw data to build ML models can increase model complexity and computational costs due to the high-dimensional features, which is disadvantageous for real-time monitoring. Besides, this treatment may also reduce the model's prediction accuracy and generalization capability due to noise interference^[30]. To address these issues, this study employed two methods, the correlation coefficient and the mutual information methods, to screen the NIR spectral wavelengths, thereby extracting informative features and improving model performance.

The correlation coefficient method^[31], also known as the Pearson correlation coefficient, can quantify the linear relationship between two variables. Its value denoted by r , ranges between -1 and 1 , where $|r|$ value closer to 1 indicates a stronger linear relationship between the two variables, and the sign indicates a positive or negative direction. It can be calculated using Equation (1).

$$r = \frac{\text{cov}(X, Y)}{\sigma_X \cdot \sigma_Y} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \cdot \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (1)$$

where $\text{cov}(X, Y)$ represents the covariance of variables X and Y ; σ_X and σ_Y are the standard deviations of variables X and Y , respectively; n is the sample size; X_i and Y_i are the i -th observed values of variables X and Y , respectively; and $\bar{X}$ and $\bar{Y}$ are the mean values of variables X and Y , respectively.

The mutual information method^[32] measures the nonlinear relationship between two random variables by calculating their mutual information. It can assess the correlation between the target variable and feature variables. The mutual information method, unlike the correlation coefficient one, can capture not only linear relationships but also effectively identify nonlinear dependencies between variables. It can be calculated using Equation (2).

$$\text{MI}(X; Y) = \int_x \int_y p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dx dy \quad (2)$$

where X represents the target variables of the direct coal liquefaction diesel: the content of alkanes, cycloalkanes, and aromatics; Y represents the feature wavelength; $\text{MI}(X, Y)$ is the mutual information value; $p(x, y)$ is the joint probability density of X and Y ; and $p(x)$ and $p(y)$ are the marginal probability densities of X and Y , respectively.

1.4 Software establishment

To further enable the intelligent analysis of NIR spectral data and enhance the usability of the methods and models developed in this study, a dedicated software was developed using PySide6. The software interface, algorithms, and models were subsequently packaged into an executable application using auto-py-to-exe.

2 Results and discussion

2.1 Preliminary screening of NIR spectral data

The raw data consists of NIR spectral absorption intensities across the 700–2500 nm range. However, due to its high-dimensional nature and the presence of substantial redundancy and noise, direct use for modeling would increase computational burden and reduce prediction accuracy. Wavelength screening can effectively simplify the model structure, reduce computational complexity, and simultaneously enhance prediction performance and generalization capability. Figure 1 shows the NIR spectra of various fractions of DCL diesel. Saturated signals with absorbance values equal to 10 are observed around 1700 nm and within the 2200–2500 nm range. These spectral bands contain limited information and exhibit elevated noise levels, which may adversely affect the accuracy of machine learning models. A comparison of Figures 1(b) and 1(c) reveals significant spectral differences in the 1100–1500 and 1780–2260 nm wavelength ranges, suggesting these may be key spectral regions for characterizing the group composition of the diesel.

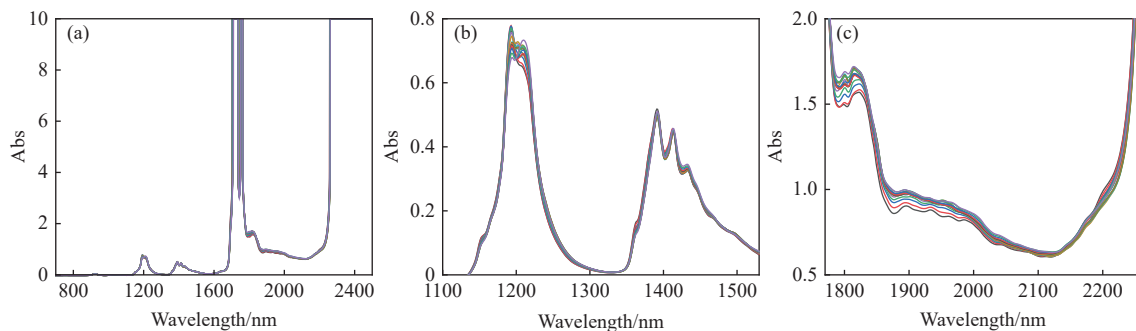


Fig. 1 The near-infrared spectra of various distillation fractions of direct coal liquefaction diesel at wavelength ranges of 700–2500 nm (a), 1100–1500 nm (b) and 1780–2260 nm (c)

Therefore, this study first applied an absorbance-threshold-based method for preliminary band screening, followed by further optimization of feature wavelength selection using the correlation coefficient method and mutual information method.

The predictive performance of the SVR model using the raw data was compared with that using data processed with absorbance elimination thresholds (1) and (2). Hyperparameter optimization ranges for the SVR model were kept consistent to ensure a fair comparison. The prediction results of the test set and training set for the three tasks are presented in Figures 2–4, where NIR-O represents the original data, and NIR-1 and NIR-2 represent the data after eliminating absorbance values greater than (1) and (2), respectively. Overall, the SVR model trained on raw data exhibited noticeable overfitting. In contrast, models built on data filtered by absorbance thresholds demonstrated

significantly improved performance. For the alkane content prediction task (Figure 2), the raw data caused noticeable overfitting due to noise interference, evidenced by a 0.144 difference in R^2 between the training and test sets. Applying an elimination threshold of 2, the model performance is improved. The percentage of data points with an absolute error <1 in the training set can reach 99.5%, superior to 97.9% with raw data. The percentage of test set data with errors <0.5 increased to 75%, superior to 58.3% with raw data. And the percentage of data with errors >1 was reduced to 2.1%, superior to 20.8% with raw data. Optimal performance was achieved with a threshold of 1: 99.5% of the training set data had errors <0.5 with no data points having errors >1 , 87.5% of the test set data had errors <0.5 , and the percentage of data with errors >1 remained stable at 2.1%.

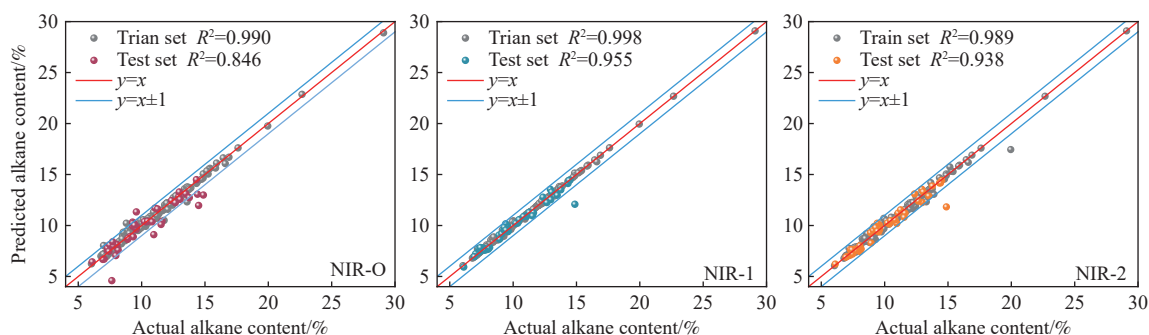


Fig. 2 Prediction of alkane relative content under different inputs

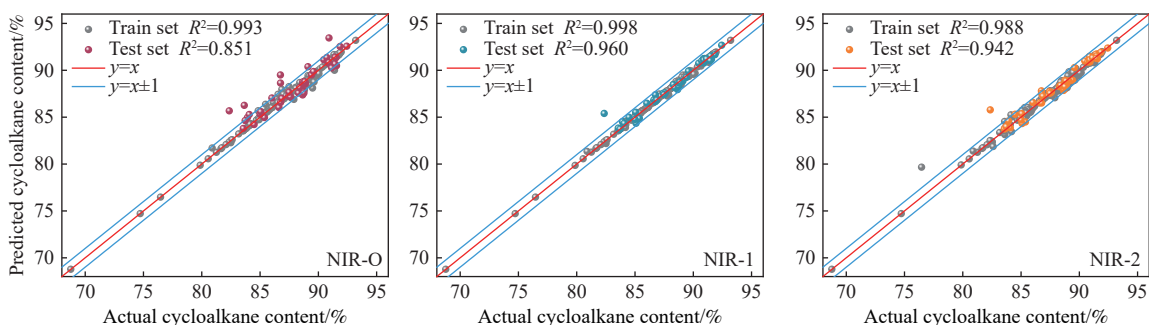


Fig. 3 Prediction of cycloalkane relative content under different inputs

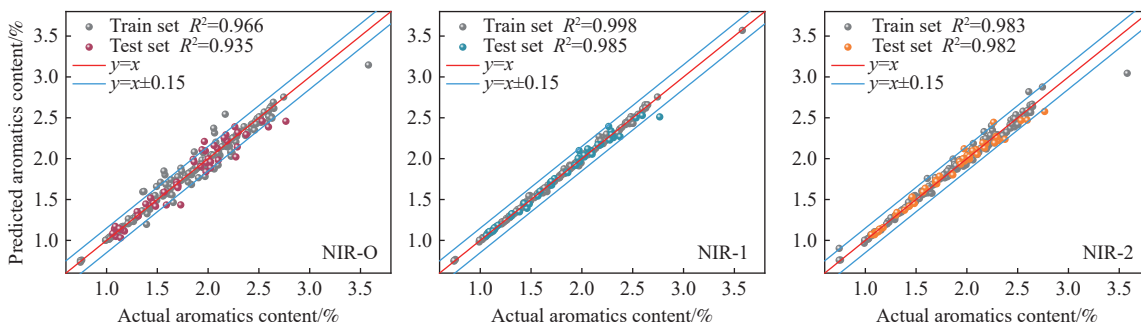


Fig. 4 Prediction of aromatics relative content under different inputs

For the cycloalkane content prediction task (Figure 3), the model performance was optimal with an absorbance threshold of 1, followed by a threshold of 2, while the use of raw data yielded the poorest results. Although the distributions of predicted data points for thresholds 1 and 2 appeared similar, a detailed statistical comparison revealed distinct performance differences. On the training set, the model with threshold 1 achieved absolute errors all below 0.5 (accounting for 98.4% of data), with no data points exhibiting errors greater than 1. This significantly outperforming the threshold 2 model (where errors <0.5 accounted for 88.4% of data, and data points with errors >1 were present). On the test set, the model with threshold 1 achieved errors <0.5 in 85.4% of data, which is again superior to threshold 2 (70.8%). The performance of model trained on raw data is comparable to the threshold-2 model. However, its errors increased substantially, with data points having errors >1 accounting for 25%, demonstrating clear overfitting.

For the aromatics content prediction task (Figure 4), the models under all three input conditions performed well. The model using the threshold of 1 is the optimal, which is consistent with the trends observed for the alkane and cycloalkane predictions. On the training set, the model with the threshold of 1 achieved a percentage of data points with prediction errors >0.05 reaching 95.2%. This is significantly outperformed the results from the raw data (71.4%) and the threshold of 2 (75.7%). On the test set, the model using raw data showed the poorest performance, with data points having errors >0.05 and >0.1 accounting for 37.5% and 35.4%, respectively. In contrast, the percentage of data points with errors >0.05 for the model with the threshold of 1 was improved to 81.3%, representing the best performance.

A systematic comparison of SVR model performance using raw data versus data processed with absorbance thresholds of 1 and 2 demonstrated that the model achieved optimal predictive performance for alkanes, cycloalkanes, and aromatics when the threshold was set to 1. Although some improvement over raw data was observed at a threshold of 2, the enhancement was limited. This indicates that data with absorbance values greater than 2 predominantly interference or noise rather than useful chemical information and should therefore be discarded. Furthermore, eliminating data with absorbance values greater than 1 effectively removes instrument-related errors while preserving critical spectral information. Consequently, this approach significantly enhances model accuracy compared to using raw data directly.

2.2 NIR spectra selection

The initial screening of wavelengths based solely on absorbance (NIR-1) produced a model with the satisfactory predictive capability for the content of alkanes, cycloalkanes, and aromatics. However, the SVR predictions

for alkanes and cycloalkanes still exhibited a certain degree of overfitting. Given that minor compositional variations in industrial production can significantly impact economic outcomes, a secondary, optimized wavelength screening was conducted to enhance the model's generalization ability. This secondary screening integrated the initial absorbance-based approach with correlation-based feature selection methods. The correlation coefficient method is a common feature selection technique used to assess linear relationships between variables. However, considering that NIR spectral data may exhibit wavelength synergy effects where the combined influence of certain wavelength pairs on the group composition far exceeds the linear contribution of any single wavelength. Additionally, some spectral information may have a nonlinear relationship with the target properties. Therefore, the correlation coefficient method alone might inadvertently remove effective features. To address these issues, this study employed both the correlation coefficient and the mutual information methods for feature selection and model construction. This combined approach facilitated an analysis of the importance assigned to each wavelength from two perspectives, thereby improving the effectiveness of the wavelength selection process.

Based on the NIR-1 data, the importance of wavelengths was analyzed using these two methods. The correlation coefficients between each component (such as alkanes, cycloalkanes, and aromatics) and the NIR spectra were calculated using the correlation coefficient method (Figure 5). As shown in Figure 5(a), the alkane content showed a positive correlation with absorbance in the 1420–1460 and 1990–2190 nm wavelength bands, and a negative correlation with the 1190–1290 nm band. From Figure 5(b), cycloalkane exhibited an opposite trend, showing positive correlations around 900–910, 1180–1205 and 1490–1510 nm, but negative correlations in the 1300–1480, 1520–1700 and 1990–2200 nm ranges. As seen in Figure 5(c), the spectral response of aromatics was similar to that of alkanes, but the influence of aromatic content was stronger near 1200 nm and weaker around 1440 nm.

The mutual information values between the content of alkanes, cycloalkanes, aromatics and NIR spectra were calculated (Figure 6). It is found that the importance trends of the three group compositions based on mutual information were similar to those obtained by the correlation coefficient method. However, the importance of wavelengths in the 2000–2200 nm interval was more prominent in the mutual information method. This is primarily because mutual information can better capture nonlinear dependencies, whereas the correlation coefficient method primarily reflects linear relationships.

Feature wavelengths were selected using the correlation coefficient method and the mutual information method with specific thresholds. For the correlation coefficient, absolute correlation coefficient values $|r|$ were > 0.8, 0.9 and 0.95, and denoted as P-0.8, P-0.9 and P-0.95, respectively.

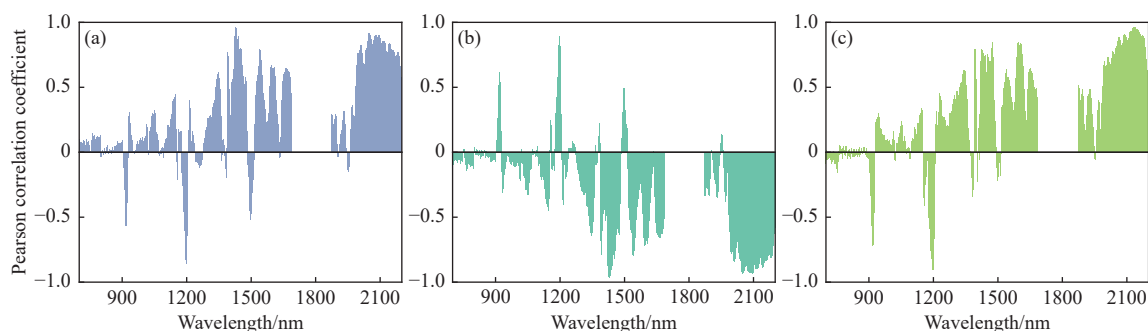


Fig. 5 The pearson correlation coefficient analysis of alkanes (a), cycloalkanes (b) and aromatics hydrocarbons (c)

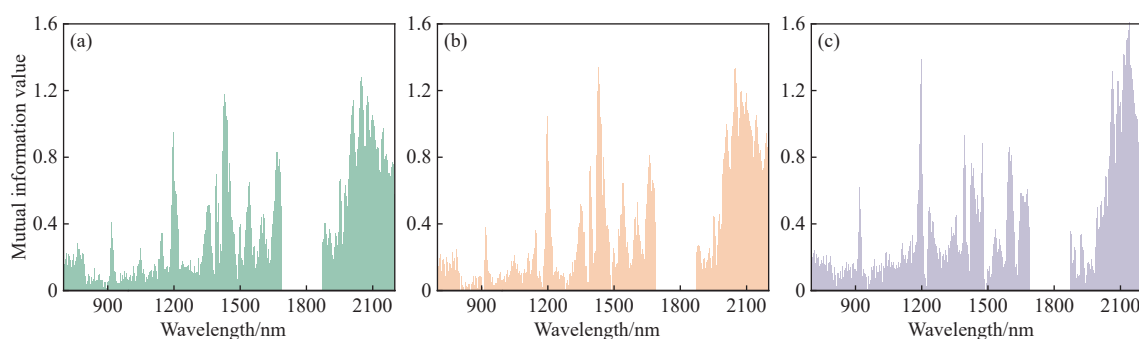


Fig. 6 Mutual information analysis for alkanes (a), cycloalkanes (b) and aromatic hydrocarbons (c)

For mutual information, values of MI were > 0.8 , 0.9 and 0.95 , and denoted as MI-0.8, MI-0.9 and MI-0.95, respectively. SVR models were then constructed based on these selected features. Figure 7 shows the number of features in the datasets resulting from each selection method and the performance of the corresponding models. As seen in Figure 7(a), feature selection significantly reduced data dimensionality with both methods. Figures 7(b)–(d) present the performance of the SVR models built on the respective datasets, where the black dots represent the test set results. From the Figures, it is evident that the model based on P-0.95, which retained only 20 wavelengths, showed a significant decrease in R^2 values on the test set for predicting alkanes, cycloalkanes, and aromatics (0.882 , 0.900 and 0.969 , respectively) compared to the training set. This indicates that overly stringent screening leads to loss of information, which is detrimental to predicting the group composition of DCL diesel. The models constructed using the other datasets all demonstrated varying degrees of improvement. Overall, the models based on P-0.8, MI-0.9 and MI-0.95 performed best. These models achieved good results on the test set for predicting each group composition, with only small discrepancies compared to the training set performance (average differences around 0.015). This demonstrates that moderate screening (e.g., $|r| > 0.8$) can effectively reduce dimensionality while optimizing generalization capability, achieving an optimal balance between prediction accuracy and model robustness.

2.3 Model performance comparison

To identify the optimal modeling approach, this study

compared the predictive performance of three ML models—Lasso, SVR and XGBoost—on three different datasets: P-0.8, MI-0.9 and MI-0.95. The results in Table 3 indicated a consistent performance ranking across the three datasets. SVR outperformed Lasso, while XGBoost demonstrated the lowest predictive accuracy. Specifically, in the alkane content prediction task, the SVR model based on the MI-0.9 dataset achieved the best performance, with an R^2 of 0.998 on the training set and 0.980 on the test set, resulting in a difference of 0.018 . This indicates that the SVR model successfully captured essential features without compromising its generalization ability. In contrast, the XGBoost model based on the MI-0.95 dataset achieved an R^2 of 0.999 on the training set, but its test set R^2 dropped to 0.903 , yielding a difference of 0.096 , clearly indicating overfitting. A similar trend was observed in the prediction of cycloalkane and aromatics content. This is primarily because XGBoost, as a complex model, is generally more suitable for tasks with large datasets and significant nonlinear relationships. However, the NIR spectral dataset of DCL diesel in this study was limited in sample size, leading to overfitting of the training data and impair its ability to extract key features for XGBoost. This thereby reduce the model's generalization performance. These findings provide important insights for model selection in NIR analysis of DCL diesel, indicating that under limited sample conditions, relatively simpler models like SVR are more advantageous than complex ensemble methods.

The distribution of absolute errors for three optimal models (Lasso-MI-0.95, SVR-MI-0.9, XGBoost-P-0.8) on both the training and test sets were shown in Figure 8.

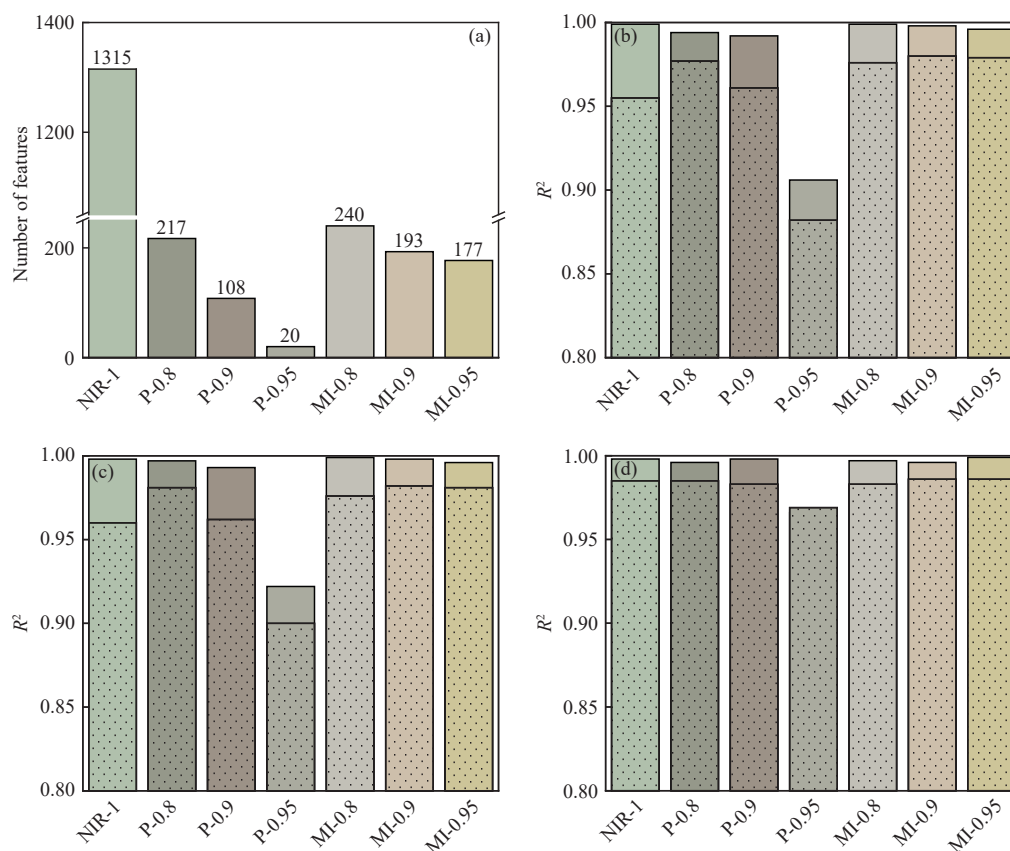


Fig. 7 (a) Number of features in each dataset; Performance of models: (b) alkane, (c) cycloalkane and (d) aromatic

Table 3 Performance of optimal models by Lasso, SVR and XGBoost at P-0.8, MI-0.9 and MI-0.95

Data set	Target	Training set R^2			Test set R^2		
		Lasso	SVR	XGBoost	Lasso	SVR	XGBoost
P-0.8	Alkane	0.991	0.998	0.985	0.954	0.977	0.916
	Cycloalkane	0.992	0.997	0.984	0.958	0.981	0.936
	Aromatics	0.991	0.996	0.994	0.981	0.985	0.957
MI-0.9	Alkane	0.990	0.998	0.994	0.959	0.980	0.905
	Cycloalkane	0.991	0.998	0.981	0.966	0.982	0.928
	Aromatics	0.991	0.996	0.998	0.981	0.986	0.951
MI-0.95	Alkane	0.989	0.996	0.999	0.963	0.979	0.903
	Cycloalkane	0.991	0.996	0.983	0.967	0.982	0.925
	Aromatics	0.991	0.999	0.989	0.978	0.986	0.948

For the alkane and cycloalkane prediction tasks, absolute errors were categorized into three classes: Class A (<0.5%), Class B (0.5%–1%) and Class C (>1%). For the aromatics prediction task, which has a smaller content range, a stricter classification standard was adopted: Class A (<0.05%), Class B (0.05%–0.1%) and Class C (>0.1%). From Figure 8 (a), all three optimal models presented satisfactory predictive performance on the training set, with Class A data points accounting for over 75% and Class C data points below 5% in each case. However, a comparison with Figure 8 (b) reveals a decline in performance on the test set for all models, though the SVR-MI-0.9 model maintained a clear advantage. On the test set (Figure 8(b)), the SVR-MI-0.9

model demonstrated superior performance. For alkane content prediction, the SVR-MI-0.9 model performed best, with Class A errors accounting for 89.6% and no Class C errors. This is significantly outperforming Lasso-MI-0.95 (79.2% Class A, 2.1% Class C) and XGBoost-P-0.8 (64.6% Class A, 6.25% Class C). Similarly, for cycloalkane prediction, the SVR-MI-0.9 model led with 85.4% Class A errors, which presented an improvement of 8.3% over Lasso-MI-0.95. In aromatic prediction, the SVR-MI-0.9 model achieved 77.1% Class A errors, exceeding the combined total of Class A and B errors (72.9%) for XGBoost-P-0.8, while the latter exhibited a high error rate with 25% Class C errors.

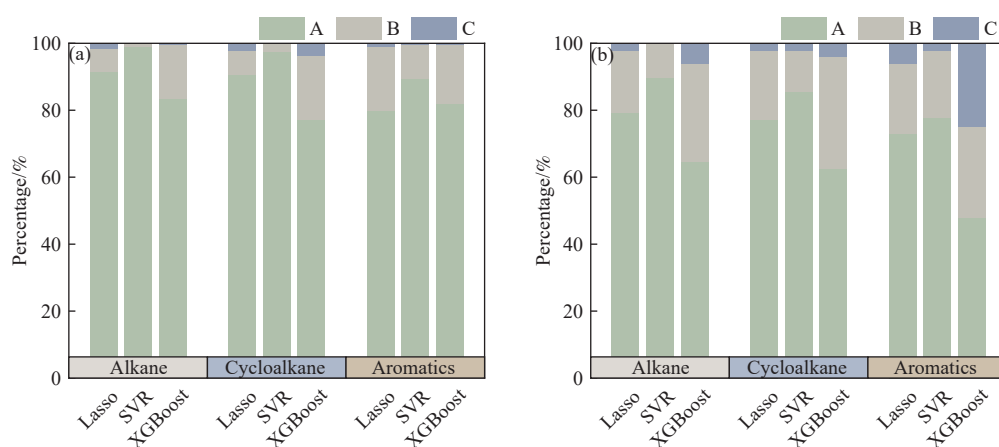


Fig. 8 Performance of optimal Lasso, SVR and XGBoost models on the training set (a) and test set (b)

By these comprehensive comparisons, the SVR-MI-0.9 model performed excellently on both the training set ($R^2 > 0.99$) and the test set ($R^2 > 0.98$). The proportion of its prediction errors falling into Class A was $\geq 77\%$, while Class C errors were $\leq 3\%$, which was also significantly surpassing the other models. This fact further confirmed that the SVR-MI-0.9 model was highly accurate and reliable to predict the group composition of DCL diesel using NIR data.

2.4 Software development

To enable the automated analysis of DCL diesel components, specialized software was developed utilized the PySide6 framework and Qt Designer tool. The software integrates the three trained models from this research—Lasso-MI-0.95, SVR-MI-0.9 and XGBoost-P-0.8. It is capable of directly reading tabular data generated by NIR spectrometers and processing the data through the relevant algorithms to provide content information for each group composition.

The software interface is shown in Figure 9. The operation procedure is as follows. Users first select the spectral file for import. The software then provides a notification indicating whether the file was successfully read. After a successful import, users can select the desired model to initiate the spectral data analysis. Once processing

is complete, a dialog box displays the prediction results.

2.5 Effectiveness of analysis software

For successful analysis of NIR spectroscopic using the software, the input NIR spectral data file must be in CSV or Excel format and contain rows for wavelength and absorbance, as illustrated in Table 4. Specifically, the required NIR spectral data must ensure that the first row consists of Arabic numbers decreasing sequentially from 2500 to 700, representing the wavelengths from 2500 to 700 nm in the NIR range. The second row should contain the corresponding absorbance values for the wavelengths listed in the first row. The data file is read by the software via the Pandas package. Then the dataset is automatically split based on the specific model employed. The model file is then loaded using the Joblib package to perform predictions.

To validate the speed and accuracy of the developed software and its underlying models, three samples of the blended fraction oils were prepared and tested in this study. The results were compared with those obtained from the software to validate the speed and accuracy of the software and its model. When analyzing the three blended fraction oils by GC×GC, the chromatographic run time was 107 min, with an average data processing time of about 30 min per chromatogram. In contrast, NIR spectroscopic analysis required only about 2 min for spectral acquisition.

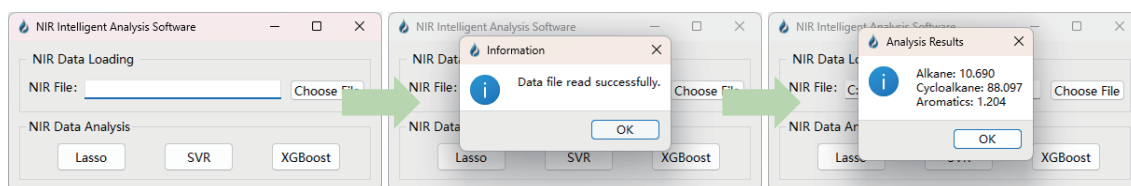


Fig. 9 Operation workflow of the NIR spectroscopy intelligent analysis software

Table 4 Example of NIR data format required for intelligent analysis software

2500	2499	...	Wavelength n	...	701	700
10	10	...	Abs	...	0.25	0.12

This dramatic difference demonstrates that the use of NIR spectroscopy combined with the software enables a much faster determination of hydrocarbon group composition in oil samples compared to the GC×GC

method. Furthermore, the analytical results obtained by the two methods (Table 5) indicate that the software provides high accuracy and reliably in determination of the hydrocarbon group composition of the oils.

Table 5 Comparison of analysis results between comprehensive two-dimensional gas chromatography and NIR analysis software

Sample	Type	GC×GC result/%	NIR result/%	Absolute error/%
Sample 1	Alkane	7.7	7.5	−0.2
	Cycloalkane	91.1	91.3	+0.2
	Aromatics	1.1	1.2	+0.1
Sample 2	Alkane	7.7	7.7	0.0
	Cycloalkane	91.2	91.2	0.0
	Aromatics	1.1	1.1	0.0
Sample 3	Alkane	8.0	7.8	−0.2
	Cycloalkane	91.0	91.1	+0.1
	Aromatics	1.0	1.1	+0.1

3 Conclusions

This study established a predictive model for the component content of DCL diesel through ML and NIR analysis. The data processing and modeling methods were also systematically optimized. The following results and conclusions were obtained:

An initial wavelength screening was performed based on the magnitude of NIR spectral absorbance, followed by further feature wavelength extraction using the correlation coefficient method and mutual information method. Compared with using the raw data, the value of R^2 by the model developed in this work increased from 0.85 to 0.96. The number of feature wavelengths was reduced from about 1800 to approximately 200, leading to improved accuracy and computational efficiency.

Among the three ML models developed and compared, SVR-MI-0.9 demonstrated the best performance, with R^2 values exceeding 0.98 for both the training and test sets. It proved highly effective in predicting the contents of alkanes, cycloalkanes, and aromatics in DCL diesel, and thus provide an effective method for the rapid analysis of the hydrocarbon group composition;

An intelligent analysis software was developed with the features of rapid spectral interpretation capability based on the established model. This software can provide a reliable foundation for subsequent online detection system development.

Declaration of Competing Interest

The authors declare no conflict of interest.

References

[1] LIN X, YIN J, DING X, et al. Study on the transformation characteristic of heteroatoms during liquefaction of

Naomahu coal[J]. J Fuel Chem Technol, 2021, 49(5): 656–663.

[2] LIU H, JIANG S S, GUO H S, et al. A new kind of pour point depressant: Diesel from direct coal liquefaction[J]. Fuel Process Technol, 2016, 149: 285–289.

[3] COOKSON D J, LLOYD C P, SMITH B E. Investigation of the chemical basis of diesel fuel properties[J]. Energy Fuels, 1988, 2(6): 854–860.

[4] XIONG C, TIAN G, REN L. Correlation of physico-chemical properties with hydrocarbon groups of diesel[J]. Acta Pet Sin, 2010, (4): 551–555.

[5] QI Z, LI L, WANG X, et al. Analysis of influencing factors on the properties of coal-to-direct liquefied diesel[J]. J Fuel Chem Technol, 2025, (6): 827–835.

[6] BAI L, SMUTS J, SCHENK J, et al. Comparison of GC-VUV, GC-FID, and comprehensive two-dimensional GC-MS for the characterization of weathered and unweathered diesel fuels[J]. Fuel, 2018, 214: 521–527.

[7] ZHANG X, WANG A, WANG X, et al. Theoretical analysis of hydrogen solubility in direct coal liquefaction solvents[J]. Int J Coal Sci Technol, 2024, 11(1): 28.

[8] WU Y W, LI B N, LIU L L, et al. Research progress of analysis of mineral oil hydrocarbons using on-line high performance liquid chromatography coupled with gas chromatography[J]. Chin J Anal Chem, 2021, 49(3): 341–349.

[9] LIU S, WANG S, HU C, et al. Development of a new NIR-machine learning approach for simultaneous detection of diesel various properties[J]. Measurement, 2022, 187: 110293.

[10] YAN H, DONG C, ZHANG J, et al. Moisture prediction of biomass fuel based on near-infrared spectroscopy and deep learning algorithm[J]. Energy Fuels, 2024, 38(7): 6062–6071.

[11] LESTANDER T A, SAMUELSSON R. Prediction of resin and fatty acid content of biorefinery feedstock by on-line near-infrared (NIR) spectroscopy[J]. Energy Fuels, 2010, 24(9): 5148–5152.

[12] LESTANDER T A, JOHNSON B, GROTHAGE M. NIR techniques create added values for the pellet and biofuel

- industry[J]. Bioresour Technol, 2009, 100(4): 1589–1594.
- [13] BLANCO M, VILLARROYA I. NIR spectroscopy: A rapid-response analytical tool[J]. TrAC, Trends Anal Chem, 2002, 21(4): 240–250.
- [14] CHU X, CHEN P, LI J, et al. Progresses and perspectives of near infrared spectroscopy analytical technology[J]. J Instrum Anal, 2020, (10): 1181–1188.
- [15] DE OLIVEIRA R R, DE LIMA K M G, TAULER R, et al. Application of correlation constrained multivariate curve resolution alternating least-squares methods for determination of compounds of interest in biodiesel blends using NIR and UV–visible spectroscopic data[J]. Talanta, 2014, 125: 233–241.
- [16] VENTURA M, DEUS W B, SILVA J R, et al. Determination of the biodiesel content in diesel/biodiesel blends by using the near-infrared thermal lens spectroscopy[J]. Fuel, 2018, 212: 309–314.
- [17] CHU X. Chemometric Methods in Modern Spectral Analysis[M]. Beijing: Chemical Industry Press, 2022.
- [18] FENG F, WU Q, ZENG L. Rapid analysis of diesel fuel properties by near infrared reflectance spectra[J]. Spectrochim Acta, Part A, 2015, 149: 271–278.
- [19] ALVES J C L, HENRIQUES C B, POPPI R J. Determination of diesel quality parameters using support vector regression and near infrared spectroscopy for an in-line blending optimizer system[J]. Fuel, 2012, 97: 710–717.
- [20] ZHANG W, KASUN L C, WANG Q J, et al. A review of machine learning for near-Infrared spectroscopy[J]. Sensors, 2022, 22(24): 9764.
- [21] LIU Q, CHU X, CHEN P, et al. Rapid prediction of hydrocarbon molecular composition of naphtha based on near infrared spectroscopy[J]. Pet Process Petrochem, 2022, (1): 86–92.
- [22] AL-KAF H A G, CHIA K S, ALDUAIS N A M. A comparison between single layer and multilayer artificial neural networks in predicting diesel fuel properties using near infrared spectrum[J]. Pet Sci Technol, 2018, 36(6): 411–418.
- [23] LESTANDER T A, SANDSTRÖM L, WIINIKKA H, et al. Characterization of fast pyrolysis bio-oil properties by near-infrared spectroscopic data[J]. J Anal Appl Pyrolysis, 2018, 133: 9–15.
- [24] PAUL A, BRÄUER B, NIEUWENKAMP G, et al. A validated near-infrared spectroscopic method for methanol detection in biodiesel[J]. Meas Sci Technol, 2016, 27(6): 065002.
- [25] SALES R, DA SILVA N C, DA SILVA J P, et al. Handheld near-infrared spectrometer for on-line monitoring of biodiesel production in a continuous process[J]. Fuel, 2019, 254: 115680.
- [26] ZHU J, DENG J, MENG F, et al. Quantitative analysis of petroleum derivatives in vegetable oils based on vibration spectrum data fusion and multi-space optimized LASSO feature selection model[J]. Measurement, 2025, 255: 118010.
- [27] ELEN A, BAŞ S, KÖZKURT C. An adaptive gaussian kernel for support vector machine[J]. Arabian J Sci Eng, 2022, 47(8): 10579–10588.
- [28] SANTOS F D, SANTOS L P, CUNHA P H P, et al. Discrimination of oils and fuels using a portable NIR spectrometer[J]. Fuel, 2021, 283: 118854.
- [29] CHEN T, GUESTRIN C. XGBoost: A scalable tree boosting system[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016: 785–794.
- [30] RINNAN Å, BERG F VAN DEN, ENGELSEN S B. Review of the most common pre-processing techniques for near-infrared spectra[J]. TrAC, Trends Anal Chem, 2009, 28(10): 1201–1222.
- [31] JIANG W, LU C, ZHANG Y, et al. Molecular spectroscopic wavelength selection using combined interval partial least squares and correlation coefficient optimization[J]. Anal Methods, 2019, 11(24): 3108–3116.
- [32] DURAND A, DEVOS O, RUCKEBUSCH C, et al. Genetic algorithm optimisation combined with partial least squares regression and mutual information variable selection procedures in near-infrared quantitative analysis of cotton–viscose textiles[J]. Anal Chim Acta, 2007, 595(1): 72–79.

近红外光谱智能分析煤直接液化柴油组分

王喜武^{1,2}, 李皓玮¹, 齐振东^{1,2}, 王兴宝^{1,*}, 冯 杰^{1,*}, 祝一蒙³, 李文英¹

(1. 太原理工大学 省部共建煤基能源清洁高效利用国家重点实验室, 山西 太原 030024;

2. 中国神华煤制油化工有限公司鄂尔多斯煤制油分公司, 内蒙古 鄂尔多斯 017209;

3. 北京东土科技股份有限公司, 北京 100144)

摘 要: 煤直接液化产品中柴油占比超过 60%, 与石油基柴油相比, 煤直接液化柴油十六烷值为 30–40, 无法满足 ≥ 45 车用柴油标准。快速准确地分析煤直接液化柴油化学组成, 通过组分调和来实现燃料性质满足柴油标准意义重大。气相色谱等传统分析方法虽精确但也存在分析速度慢、维护复杂且成本高等缺点, 难以满足快速实时控制的需求。近红外光谱技术具有快速、无损检测优势, 但光谱数据解析的复杂性限制了其应用; 机器学习方法能够有效挖掘光谱信息, 建立高精度预测模型。本研究结合近红外光谱与机器学习技术, 构建煤直接液化柴油的光谱-组成数据库, 并采用相关系数法和互信息法进行特征提取, 筛选关键波长变量, 降低数据维度。在此基础上, 对比 Lasso、SVR 和 XGBoost 三种机器学习模型

的预测性能。结果表明,剔除吸光度大于1的数据可显著提升模型精度,测试集 R^2 从 0.85 提高至 0.96。特征提取后,最优波长变量减少至 177 列,大幅提高计算效率。其中,基于互信息特征选择的 SVR-MI-0.9 模型表现最佳,训练集和测试集 R^2 均超过 0.98,能精准预测链烷烃、环烷烃和芳烃含量,从而为智能化在线质量监控提供了可靠方法。基于所构建的智能分析模型,自主开发了近红外光谱数据智能分析软件,与全二维气相色谱分析柴油组分结果对比,解析时间节约 98% 以上的软件分析,预测的绝对误差不超过 0.2%,完全实现了煤直接液化柴油组分的快速解析。

关键词: 煤直接液化油; 光谱实时检测; 机器学习; 特征提取; 组分预测

中图分类号: TQ529.1

文献标识码: A

作者简介:



王喜武(1980), 硕士, 正高级工程师, 主要研究方向为煤化工、煤直接液化, 10511443@ceic.com。



王兴宝(1985), 博士, 副教授, 研究方向为煤炭清洁高效转化利用, 主要从事煤直接液化及智能化工相关研究, wangxingbao@tyut.edu.cn。



李皓玮(2000), 硕士, 研究方向为煤炭清洁高效转化利用, 主要从事智能化工相关研究, lihaowei1022@link.tyut.edu.cn。



冯杰(1968), 博士, 研究方向为煤/生物质高附加值资源化利用研究, 主要从事生物质转化、煤转化基础理论及其气化反应器设计/工艺的改进方面的研究, fengjie@tyut.edu.cn。