DOI 10.1007/s10812-021-01220-5
Journal of Applied Spectroscopy, Vol. 88, No. 3, July, 2021 (Russian Original Vol. 88, No. 3, May–June, 2021)

INTELLIGENT PROXIMATE ANALYSIS OF COAL BASED ON NEAR-INFRARED SPECTROSCOPY

W. Liu,[a] B. Peng,[a] X. Liu,[a] F. Ren,[a] and L. Zhang[b,*]

UDC 543.42

The proximate analysis of coal, which aims to estimate the moisture, volatile matter, and caloric value, is of great importance for coal processing and evaluation. However, traditional methods for proximate analysis in the laboratory are not only time-consuming and labor-intensive but also expensive. The near-infrared spectroscopy (NIRS) technique provides a rapid and nondestructive method for coal proximate analysis. We exploit two regression methods, random forest (RF) and extreme learning machine (ELM), to model the relationships among spectral data and proximate analysis parameters. In addition, given the poor stability and robustness caused by the random selection of parameters in ELM, we employ the particle swarm optimization algorithm (PSO) to optimize the structure of ELM (PSO–ELM). A total of 384 coal samples from Inner Mongolia are collected for model training and validation. The experimental results show that the proposed PSO–ELM algorithm achieves the best performance in terms of accuracy and efficiency, which indicates that NIRS combined with PSO–ELM has significant potential for accurate and rapid proximate analysis.

Keywords: *coal, proximate analysis, random forest, extreme learning machine, particle swarm optimization.*

Introduction. The world has witnessed rapid growth in coal and coke consumption, especially after the beginning of industrialization [1]. Coal provides affordable and reliable electricity to millions of people in developing and emerging economies. With the explosive development in industrial fields such as cement making and electricity generation [2], the traditional utilization of fossil fuels is unsatisfactory not only for its low-grade energy utilization but also for increasing pollution [3, 4].

To utilize coal more wisely and efficiently, accurate estimation of coal parameters in the proximate analysis is essential, which is always the first stage of coal assessment. The main parameters of proximate analysis include the moisture, volatile matter, and caloric value. Moisture content is the water that exists in the coal. Moisture absorbs heat, and therefore, the high moisture content in coal decreases the relative efficiency of combustion. In addition, the weight added by moisture is an unfavorable factor in coal transportation. Water in coal also affects its sorption of gas, including methane [5] and carbon dioxide [6]. Volatile matter (VM), by definition, is the measure of nonwater gases formed from a coal sample when heated to 950^{o}C in an oxygen-free environment. VM is the product of coal thermal decomposition and mainly consists of hydrogen, methane, carbon monoxide, carbon dioxide, and other complex organic compounds. This product is a key safety concern since coals with high VM have an increased risk of spontaneous combustion. Calorific value is a dominant factor in coal pricing [7] since approximately half of the coal is used by power plants. This parameter is determined by measuring the heat produced by the complete combustion of coal in oxygen.

Determining moisture, volatile matter, and calorific value in the laboratory with traditional testing methods ought to be performed under rigidly controlled conditions. In the sample preparation process, all samples need to be pulverized to 250 μm. In addition, the measurement of coal parameters involves many instruments, such as drying ovens, capsules, balance, electric furnaces, and platinum crucibles. Each step in the experimental procedure may result in remarkable accumulative errors.

Near-infrared (NIR) spectroscopy, a fast and nondestructive analytical method that uses near-infrared spectroscopy with an electromagnetic spectrum from approximately 780 to 2526 nm, has been deployed for chemical investigation. These existing studies reveal its advantages in component detection over traditional elemental analytical methods [8, 9].

[a]School of Information and Control Engineering at China University of Mining and Technology, Xuzhou, China; [b]School of Law at Tsinghua University, Beijing, China; email: zhangli19@mails.tsinghua.edu.cn. Abstract of article is published in Zhurnal Prikladnoi Spektroskopii, Vol. 88, No. 3, p. 501, May–June, 2021.

To determine the coal parameters in proximate analysis effectively and efficiently, much attention has been given to constructing a high-quality model. For instance, Jonathan P. Mathews reviewed the correlations of coal properties with elemental composition [10], and several models were established to study the coal characteristics based on the NIR technique. Combined with partial least square regression (PLS), the predictive ability of the supportive vector machine (SVM) is improved [11]. Although previous works have achieved good research results in prediction accuracy, the efficiency of these studies, which is also important in the practice and application of portable NIR scanners, is unsatisfactory. In addition, previous studies were conducted on relatively small datasets, and inhomogeneity may exist, which may lead to overfitting and low generalization ability of the established models. The random forest (RF) algorithm is a statistical model that can be regarded as an ensemble of various decision trees [12–14]. It provides nonlinear statistical approximation to handle high-dimensional data instead of transforming the data with kernel functions. Even when trained with limited samples, this algorithm can achieve high prediction accuracy. The extreme learning machine (ELM), another feed-forward neural network, has come into practice since 2015. Unlike other network-based algorithms, the weight of hidden nodes in ELM can be randomly assigned and never changed [15, 16]. Apart from the good prediction accuracy, the most attractive highlight of ELM is that it outperforms popular machine learning algorithms in training and fitting time on many publicly available benchmark datasets [17–20]. However, since the input weight and offset parameters of ELM are randomly set, ELM also has the disadvantages of low robustness and stability.

Considering the abovementioned issues, we expanded the data volume of samples, totaling 384 samples. To optimize the structure of the ELM, we employed particle swarm optimization algorithm to obtain the optimized parameter set, including input weights and bias. The experimental results show that PSO–ELM yields the best prediction performance in terms of accuracy and time.

Materials and Methods. A reliable dataset is essential for modeling to obtain the desired performance. A total of 384 coal samples from Inner Mongolia was collected and prepared according to the 'Method for Preparation of Coal Sample' (GB474-2008) [21] in the National Laboratory of the Import and Export Quarantine Inspection Bureau. All coal samples were scanned by an Antaris II Fourier transform near-infrared spectrometer (Fig. 1), and 1609 wavelength points were obtained for the spectrum of each sample. The Michelson interferometer is mainly composed of a fixed mirror, moving mirror, light source, beam splitter, sample stage, and detector. Light from the source is split into two parts by a beam splitter, resulting in a difference in the length of the optical path. After interference, the light is transformed into an electrical signal when it reaches the detector. Fourier transform is performed by the amplifier and A/D converter when the interferogram is generated, amplified, and converted to digital. The spectrum of the sample is plotted with the x axis of wavelength and y axis of absorption.

Figure 2a shows all spectra of 384 coal samples (each color represents one sample) with wavelengths ranging from 3799.0793 to 10,001.0283 cm^{-1}. The sampling and proximate analysis process strictly follows the guidelines for the standard testing method for moisture, volatile matter, and calorific value. Four-fifths of the samples in the dataset was randomly used for training, and one-fifth was used for testing in 5-fold cross-validation. The training and testing sets remain exactly the same for all models. Real values of moisture, volatile matter, and calorific value are plotted in Fig. 2b. The color map is used here for demonstration, and each sample is represented with one unique color.

Developed from decision trees, the RF model is an ensemble algorithm of multiple decision tree regressors and achieves good performance in classification and regression tasks (Fig. 3). In regression tasks, the original dataset is broken down into smaller and smaller subsets.

Given a dataset $D = \{(x_1, y_1), \dots, (x_i, y_i), \dots, (x_n, y_n)\}$, where x_i and y_i denote the input and output values, respectively, and $x_1, x_2, \dots, x_k$ are the features in x_i, it is partitioned into M regions $R_1, R_2, \dots, R_m$. These regions R_m are defined by a series of binary splits. In the fitting process, the aim is to minimize the squared error as follows:

$$\sum_{x_i \in R_m} (f(x_i) - y_i)^2 , \tag{1}$$

where $f(x_i)$ is the predicted value and $f(x_i)$ is the average of y:

$$f(x_i) = c_m = \text{ave}\,(y_i \,|x_i \in R_m) . \tag{2}$$

The decision tree model automatically decides the splitting variables and split points. We define a pair of half-planes

$$R_1(j, s) = \{x : x_j \le s\} , \; R_2(j, s) = \{x : x_j > s\} . \tag{3}$$

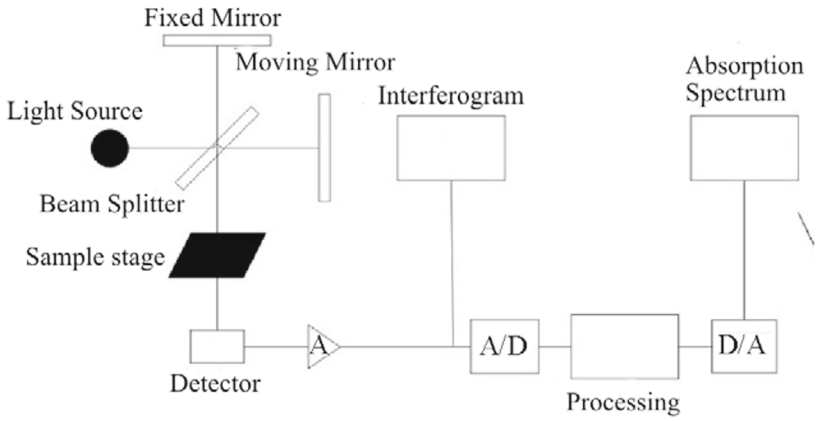

Fig. 1. Layout of the NIR system.

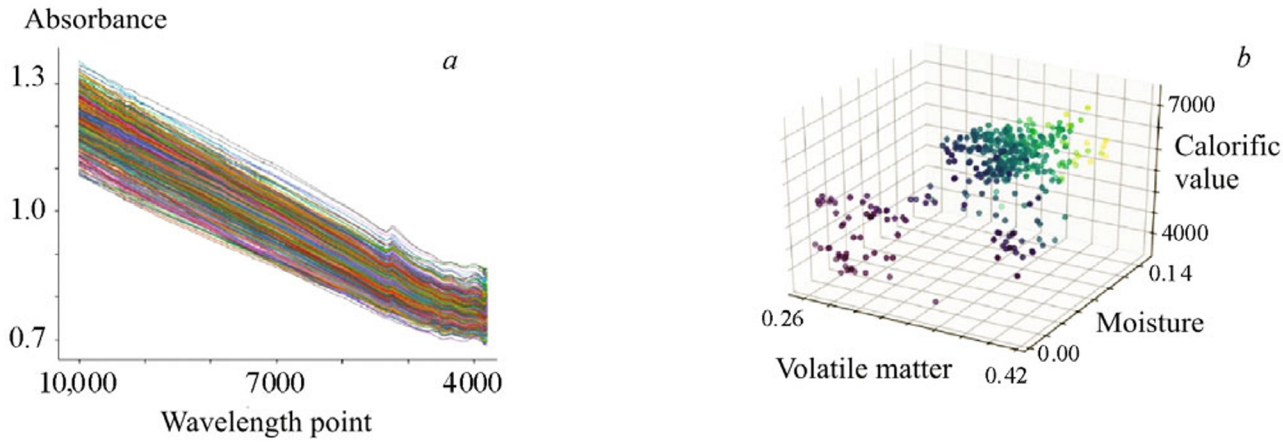

Fig. 2. Dataset perspective: a) spectrum of coal samples (features); b) ranking parameters (labels).

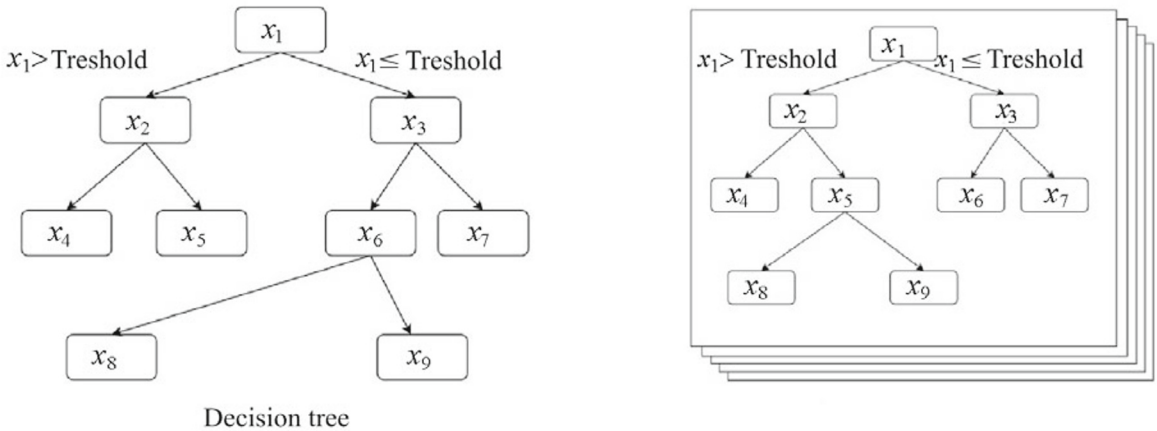

Fig. 3. Demonstration of a decision tree and random forest structure.

Denote a splitting variable as j and split point as s; the objective function for searching the optimal j and s can be written as

$$\min_{j,s} \left(\min_{c_1} \sum_{(x,y)\in R_1(j,s)} (y - c_1)^2 + \min_{c_2} \sum_{(x,y)\in R_2(j,s)} (y - c_2)^2 \right), \tag{4}$$

where $c_1 = \text{ave}\,(y_i | x_i \in R_1(j,s))$ and $c_2 = \text{ave}\,(y_i | x_i \in R_2(j,s))$.

There are a finite number of dimensions in x_i, and thus, we can search over all of them. Given a fixed splitting variable j, the best splitting point can be determined in a short time. Checking all the input variables makes the pair (j, s) feasible for seeking splitting variables and split points. The process of seeking is repeated on all resulting subsets. The model based on the decision tree algorithm is easy to interpret. However, decision tree regressors are limited by low generalization capability. Overfitting trees are created when the minimum number of samples required at a leaf node or the maximum depth of the tree are not set appropriately. What makes the RF model different from the traditional decision tree is that the original decision tree algorithm divides the input data based on all feature parameters, whereas the RF model splits on the subset of randomly picked features, also known as bootstrapping. The process of bootstrapping can avoid overfitting effectively, thus improving the performance of random forest. The general idea of bagging can be described as:

Step 1. Create multiple subdatasets drawn with replacement from the training set;

Step 2. Build multiple decision tree regressors on each bootstrap sample;

Step 3. Combine decision trees.

Based on the bagging method, each tree in the random forest model is grown with a bootstrap set from the training set. Another source of randomness comes from the random input vectors: the best split point is found over a random sample of features to minimize the residual sum of squares (RSS)

$$\text{RSS} = \sum_{i=1}^{n} (y_i - y_{i,t})^2, \tag{5}$$

where y_i is the true value and $y_{i,t}$ is the predicted value of y_i.

As an ensemble model, the random forest model fits the input data in a shorter time due to the independence of each decision tree, making parallel computing and modeling possible. One trick employed in RF is the setting of the maximum number of potential predictors. Thus, the situation of a strong predictor being a root node repeatedly is avoided. The RF model provides unbiased estimation with generalization. In addition, RF can rank the importance of all features in the regression task. The extreme learning machine is a fast machine learning method composed of feedforward neural networks. It is proposed to speed up the training and fitting process of machine learning. Another advantage of the ELM model is that it can provide a universal approximation, implying that it can tackle most regression tasks with desired prediction accuracy.

In the ELM model, there is only one nonlinear neural layer, and all variables are taken by the input layer without any additional computation. Here W, the weight of the input layer, and b, the value of bias, are fixed in the training progress as they are set out of random beforehand; f represents nonlinear functions that operate transformations on input data. As the only nonlinear layer of the entire model, the nonlinear function improves the prediction performance of ELM (transformation function can be linear or nonlinear). Here H, the hidden layer, can be constrained to multiple nonlinear functions to calculate the weight of the output layer (as shown in Fig. 4). The standard single-layer feedforward network can be described as

$$H\beta = Y, \tag{6}$$

where $H = f(WX + b)$ and β is the output weight.

The target is to minimize $H\beta - Y$ to achieve the smallest training error [22]. This equation has strictly proven that i) a model with randomly selected input weight W and hidden layer bias b can learn with the desired small error [23, 24] and ii) the minimum norm solution of $\min\|H\beta - Y\|$ is an inverse of H [25]. Thus, the remaining work aims to find specific β by least squares fit as $\hat{\beta} = H^{-1}Y$. As stated, in most cases, we have $\tilde{n} \le k$ ($\tilde{n}$ is the number of hidden neurons); therefore, H is a non-square matrix. Therefore, the smallest norm least-squares solution can be found when H^{-1} is replaced by H^{+}, the Moore–Penrose generalized inverse of H. Finally, the extreme learning machine method can be summarized as follows:

Step 1. Select input weight W and hidden layer bias b out of random;

Step 2. Calculate hidden layer output matrix H;

Step 3. Calculate output weight β.

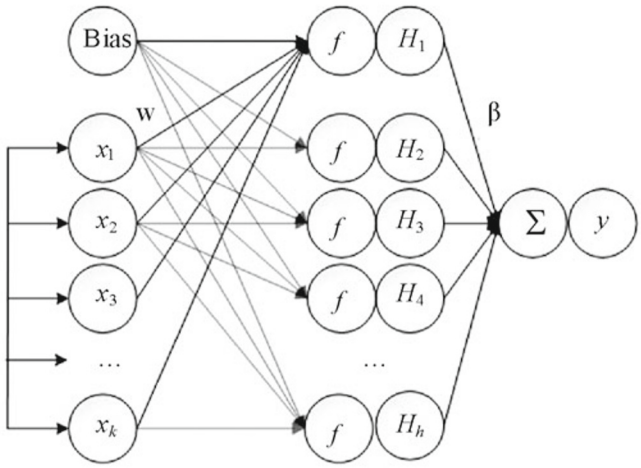

Fig. 4. Demonstration of ELM structure.

In most cases, the ELM algorithm can produce a good effect, but as a result of the threshold and weights being randomly selected, it inevitably leads to modeling the redundancy of hidden layer neurons to issue such differences in unknown input parameter identification ability, thus reducing the stability and accuracy of the model. Thus, the extreme learning machine was optimized by using the PSO algorithm in the threshold and the weights to further enhance the stability and accuracy of the model. PSO is a global optimization strategy based on the social behavior of animal species. During the optimization process, we initialize q particles in the group $s = (s_1, s_2, \dots, s_q)$, and each particle represents a potential solution of the ELM model. In this study, the dimension of the search space is set as D. The ith particle is assigned a random position represented as $s_i = (s_{i1}, s_{i2}, \dots, s_{iD})^{\mathrm{T}}$ and a random velocity represented as $v_i = (v_{i1}, v_{i2}, \dots, v_{iD})^{\mathrm{T}}$. Similarly, to the objective function, the fitness function of the optimization process is defined as follows.

When particle i searches the solution space, the optimal parameter $P_i = (p_{i1}, p_{i2}, \dots, p_{iD})^{\mathrm{T}}$ and the optimal parameter $P_g = (p_{g1}, p_{g2}, \dots, p_{gD})^{\mathrm{T}}$ of the population are saved. The particle speed and particle position are adjusted once per iteration and updated according to the following formula:

$$\begin{aligned} v_{id}^{t+1} &= wv_{id}^{t} + c_1 r_1 (p_{id}^{t} - x_{id}^{t}) + c_2 r_2 (p_{gd}^{t} - x_{gd}^{t}) , \\ s_{id}^{t+1} &= s_{id}^{t} + v_{id}^{t+1} , \end{aligned} \tag{7}$$

where w is the initial weight, and t is the current round; c_1 and c_2 are positive acceleration constants, and r_1 and r_2 are uniformly distributed random numbers, ranging from 0 to 1.

The details of the improved ELM model are as follows:

Step 1. Initialize the initial population;

Step 2. Set the mean square error of the network as the fitness function;

Step 3. Calculate the fitness value of each particle and save the current global best solution;

Step 4. Update the velocity and position of each particle and update the personal best solution;

Step 5. Determine whether it has obtained the maximum number of iterations or the required error. If it has, then end, otherwise return to Step 3.

Results and Discussion. Spectral data are used as input features, and proximate analysis parameters (moisture, volatile matter, and calorific value) are employed as regression targets. In this experiment, 5-fold cross-validation is employed, where 4 folds are used to fit the model, and the remaining fold is used to test the model. This process is repeated until every fold serves as the test set.

The normalized root mean, square error (NRMSE), fitting time, and prediction time are used to evaluate the performance. The NRMSE is defined as follows:

$$\mathrm{NRMSE} = \frac{\sqrt{\dfrac{1}{n}\sum_{i}^{\hat{n}} (\hat{y}_i - y_i)^2}}{y_{\max} - y_{\min}} , \tag{8}$$

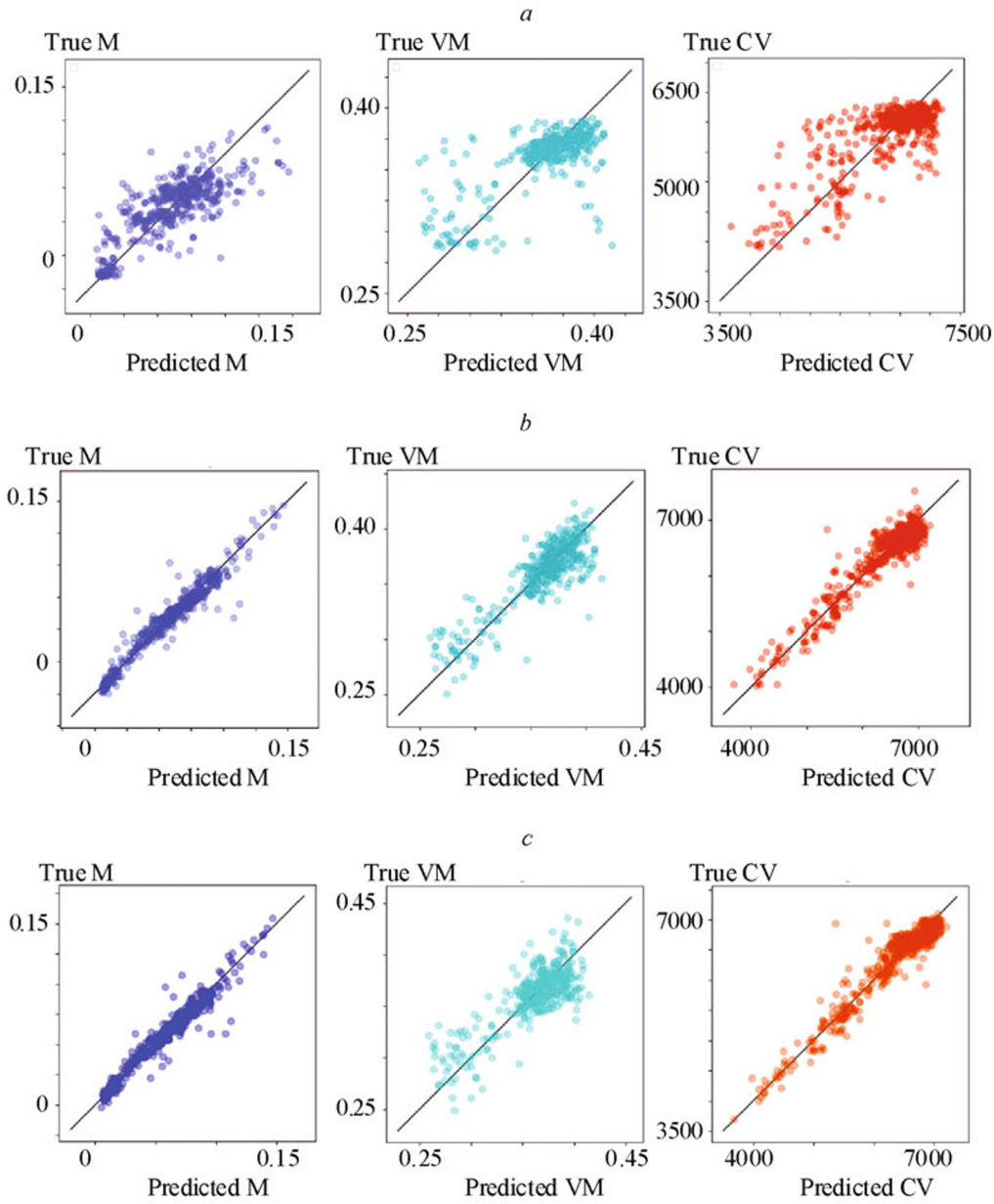

Fig. 5. Performance of (a) RF, (b) ELM, and (c) PSO-ELM modeling.

where $\hat{y}_i$ is the predicted value, y_i is the ground truth, and $\hat{n}$ is the total number of test samples; $y_{\max}$ and $y_{\min}$ are the maximum and minimum values of the testing set, respectively.

RF modeling. Based on bootstrap aggregation, the random forest is an estimator that fits each tree to various subsamples of the dataset. Figure 5a shows a comparison between the predicted value of the RF model and the true property value. The performance of the RF model is generally unsatisfactory, and samples with either small or large reference values are not predicted well in the RF model. We suspect that one potential reason is the ineffective estimation of variable importance. Variable importance in the RF model is computed by permuting features of the current variable and averaging the out-of-bag error, and therefore, it is sensitive to noise. In addition, the features of each sample spectrum are interrelated since, in the area of coal chemistry, the range of chemical elements (carbon, nitrogen, sulfur, and hydrogen) is not discontinuous for overlapping transition energy absorption. Therefore, when a corelated variable is chosen as an indicator, the importance of other variables decreases to different degrees, so it is misleading to improperly evaluating predictors even if they have very similar responses. The measurements of execution time and the prediction error of the RF model are summarized in

TABLE 1. Performance Comparison of Different Models

Property	Model	Fitting time, s	Prediction time, s	NRMSE
Moisture	RF	0.893	0.105	0.12394
	ELM	0.188	0.012	0.05605
	PSO–ELM	0.049	0.007	0.01297
Volatile matter	RF	0.921	0.105	0.13083
	ELM	0.233	0.009	0.11248
	PSO–ELM	0.121	0.005	0.04261
Calorific value	RF	0.982	0.104	0.13123
	ELM	0.205	0.009	0.06880
	PSOELM	0.089	0.002	0.02579

Table 1. The NRMSEs of predicting moisture, volatile matter, and calorific value are 0.12394, 0.13083, and 0.13123, respectively, which indicates that there is much room for improvement to achieve a more accurate proximate analysis of coal.

ELM and PSO–ELM modeling. To make better predictions, ELM is introduced in this paper due to its effectiveness and efficiency. Note that the data splitting criteria used in this section are the same as those in the RF model. Figure 5b plots the straightforward comparison of the reference value and the predicted value. It shows that the overall performance has improved significantly. The statistics of the ELM model are listed in Table 1. Taking moisture as an example, the NRMSE yielded by the ELM model is 0.05605, compared with 0.12394 for the RF model. In addition, the fitting time and prediction time are approximately 1/5 and 1/9, respectively, of the RF model. There is a similar case in terms of volatile matter and calorific value. Experimental results show that the ELM model performs better than the RF model in predicting the three properties of coal.

Considering that the input weights and hidden biases are generated randomly, the ELM model suffers from low robustness and generalization. To optimize the structure of the ELM, PSO is introduced in this paper to obtain the optimized parameter set, including input weights and biases. Figure 5c shows the prediction performance of the proposed PSO–ELM model. Compared with ELM, PSO–ELM performs significantly better in predicting all properties except volatile matter. For an overall insight, a comparison of performance parameters from those discussed models is shown in Table 1. It shows that the PSO–ELM model is superior to the other two models, RF and ELM. For simplicity, taking the calorific value as an example, the PSO–ELM model yields outstanding performance with an NRMSE of 0.02579 and prediction time of 0.002 during the testing phase, which is nearly 1/3 and 1/4 of the ELM model. In addition, the proposed model is effective for predicting moisture and volatile matter. These results indicate that the developed PSO–ELM algorithm has good generalization capability for proximate analysis of coal.

Conclusions. Three machine learning models, RF, ELM, and PSO–ELM, are established to predict moisture content, volatile matter, and calorific value based on NIRS. The extreme learning model optimized by the particle swarm optimization algorithm is expected to be of practical use on portable near-infrared examiners for coal trading and quality inspection. We evaluate the performance of the proposed models on 386 coal samples originating from Inner Mongolia. The fitting time, prediction time, and NRMSE are used to evaluate the regression models. Experimental results demonstrate that the PSO–ELM model can be considered a reliable tool in intelligent proximate analysis.

REFERENCES

1. Y. Wolde-Rufael, *Appl. Energ.*, **87**, 160–167 (2010).
2. L. Pérez-Lombard, J. Ortiz, and C. Pout, *Energ. Buildings*, **40**, 394–398 (2008).
3. M. M. Alam, M. W. Murad, A. H. M. Noman, and I. Ozturk, *Ecol. Indic.*, **70**, 466–479 (2016).
4. Y. Jafari, J. Othman, and A. H. S. M. Nor, *J. Policy Model.*, **34**, 879–889 (2012).
5. J. I. Joubert, C. T. Grein, and D. Bienstock, *Fuel*, **52**, 181–185 (1973).

6. M. Švábová, Z. Weishauptová, and O. Přibyl, *Fuel*, **92**, 187–196 (2012).
7. U. Lorenz and Z. Grudziński, *Appl. Energ.*, **74**, 271–279 (2003).
8. Y. Hu, L. Zou, X. Huang, and X. Lu, *Sci. Rep. Uk*, **7** (2017).
9. Y. Hongfu, C. Xiaoli, L. Haoran, and X. Yupeng, *Fuel*, **85**, 1720–1728 (2006).
10. J. P. Mathews, V. Krishnamoorthy, E. Louw, A. H. N. Tchapda, F. Castro-Marcano, V. Karri, D. A. Alexis, and G. D. Mitchell, *Fuel Process. Technol.*, **121**, 104–113 (2014).
11. Y. Wang, M. Yang, G. Wei, R. Hu, Z. Luo, and G. Li, *Sens. Actuat. B: Chem.*, **193**, 723–729 (2014).
12. L. Breiman, *Mach. Learn.*, **45**, 5–32 (2001).
13. P. Probst, A. Boulesteix, and B. Bischl, *J. Mach. Learn. Res.*, **20** (2019).
14. L. Zou, Q. Huang, A. Li, and M. Wang, *China Life Sci.*, **55**, 618–625 (2012).
15. S. Tamura and M. Tateishi, *IEEE Trans. Neural Net.*, **8**, 251–255 (1997).
16. G. B. Huang, *IEEE Trans. Neural Net.*, **14**, 274–281 (2003).
17. A. Akusok, K. Bjork, Y. Miche, and A. Lendasse, *IEEE Access*, **3**, 1011–1025 (2015).
18. G. Huang, H. Zhou, X. Ding, and R. Zhang, *IEEE Trans. Syst. Man Cybern. B: Cybern.*, **42**, 513–529 (2012).
19. J. Tang, C. Deng, and G. Huang, *IEEE T. Neur. Net. Learn.*, **27**, 809–821 (2016).
20. G. Huang, Q. Zhu, and C. Siew, *Neurocomputing*, **70**, 489–501 (2006).
21. M. Lei, Z. Rao, M. Li, X. Yu, and L. Zou, *Appl. Sci.,* **9**, 1111 (2019).
22. P. L. Bartlett, *IEEE T. Inform. Theory*, **44**, 525–536 (1998).
23. G. Huang, Q. Zhu, and C. Siew, *Neurocomputing*, **70**, 489–501 (2006).
24. G. B. Huang, Q. Y. Zhu, and C. K. Siew, *Neural Networks*, **2**, 985–990 (2004).
25. P. Lancaster and M. Tismenetsky, *The Theory of Matrices with Application*, Elsevier (1985).