Soft-sensor of Carbon Content in Fly Ash based on LightGBM

Liu Junping[1], Luo Hairui[2], Huang Xiangguo[3], Peng Tao[4], Zhu Qiang[5], Hu XinRong[6], He Ruhan[7]
Hubei Provincial Engineering Research Center for Intelligent Textile and Fashion
Engineering Research Center of Hubei Province of Clothing Information
School of Computer Science and Artificial Intelligence
Wuhan Textile University, Wuhan, China[1, 2, 4, 5, 6, 7]
Hubei Provincial Exchange, Wuhan, China[3]

Abstract—The soft-sensor method of carbon content in fly ash is to predict and calculate the carbon content of boiler fly ash by modeling the distributed control system (DCS) data of thermal power stations. A novel data-driven soft-sensor model that combines data pre-processing, feature engineering and hyperparameter optimization for application in the carbon content of fly ash is presented. First, extract steady-state data by data mining technology. Second, twenty characteristics that may affect the carbon content in fly ash are identified as variables by feature engineering. Third, a LightGBM prediction model that captures the relation between the carbon content in fly ash and various DCS parameters is established and improves the prediction accuracy by the Bayesian optimization (BO) algorithm. Finally, to verify the prediction accuracy of the proposed model, a case study is carried out using the data of a coal-fired boiler in China. Results show that the proposed method yielded the best prediction accuracy and closely approximates the non-linear relationships between variables.

Keywords—LightGBM; carbon content; fly ash; soft-sensor; feature engineering; Bayesian optimization

I. INTRODUCTION

The unburned carbon content in fly ash reflects the combustion efficiency of a coal-fired boiler. The combustion condition of coal can be better evaluated by analyzing the unburned carbon content in fly ash [1]. Real-time monitoring of carbon content in fly ash helps keep the carbon content in fly ash within a reasonable range, thus reducing the cost of power generation and improving the economy of generating units.

Currently, the methods for detecting the carbon content of fly ash are divided into two categories: physical measurement methods and soft-sensor methods. Physical methods commonly include the loss-on-ignition method [2], Laser-induced breakdown spectroscopy [3], microwave absorption method. [4] etc. Physical solutions are not widely available due to technical or cost reasons [5-6]. Machine learning methods have been widely used in human life, industrial production and power generation [7-9]. Distributed control system (DCS) is a computer control system for centralised management and decentralised control of the production process [10]. The distributed control system contains many sensors, which record the information of system operation. By analyzing this information, we can predict the operation status of the system [11]. The soft-sense method organically combines the production process knowledge through mechanism analysis, which can quickly and accurately reflect the carbon content in fly ash under different working conditions, and has a high economy.

Currently, there are three main problems with soft-sensor methods for the carbon content in fly ash:

1) The boiler combustion process is a multivariable variable, nonlinear and highly coupling thermal process [12]. For example, the DCS records variables such as air volume, air pressure, and air temperature for each coal mill outlet. These variables are highly correlated with boiler combustion prediction modeling, resulting in a certain amount of variable redundancy, affecting the model estimation accuracy, and increasing the computational complexity. Therefore, it is necessary to apply feature engineering to reduce the impact caused by redundant variables.

2) Most current research tests have limited data and working conditions. They do not effectively represent the complete operational status of the boiler.

3) The accuracy of these algorithms is limited.

II. RELATED WORK

Zhou et al [13]. established an artificial neural networks (ANN)-based soft-sensor model for the carbon content in the fly ash of a 300MW utility tangentially firing coal burned boiler and verified the validity of the model by multi-state thermal experiments. Wang et al [14]. proposed building a prediction model with support vector regression(SVR) for carbon content in fly ash and showed through experiments that the carbon content in fly ash model using SVR has reliable generalization and is suitable for online modeling. In machine learning, finding appropriate data processing methods, such as removing noise data and extracting suitable features, will help to improve the accuracy of prediction [15]. To address these issues, Zhu et al [16] performed a sensitivity analysis of the related features for the carbon content in fly ash, using the Garson algorithm for variables selection before modeling. Wang [12] collects the factors influencing the carbon content in fly ash constitute the initial variables, and the optimal variables are selected by the random forest-based variable selection method. The machine learning model contains many super parameters, such as penalty, learning rate and loss function. A suitable combination can effectively improve the

prediction accuracy of the model [17]. Feng [18] improves the model generalization ability by using the genetic algorithm to optimize the values of each neural network parameter. Peng [19] proposed an adaptive perturbation quantum particle swarm optimization algorithm (AQPSO) with a support vector machine to jointly predict the carbon content in fly ash and improve the prediction accuracy of the SVR model by ADQPSO.

LightGBM [20] is an ensemble learning algorithm, Developed by Microsoft in 2017. It is an advanced implementation of the distributed gradient boosting decision tree (GBDT) framework. The GBDT [21] algorithm is the core of LightGBM, which iteratively sums weak estimators to generate robust estimators by computing the negative gradient of the loss function. Lightgbm integrates GOSS (Gradient-based One-Side Sampling) algorithm and EFB (Exclusive Feature Bundling) algorithm based on GBDT. GOSS algorithm can lead to a more accurate gain estimation than uniformly random sampling, and the EFB algorithm provides a nearly lossless approach to reduce the number of effective features [20]. LightGBM algorithm extensively applied in many regression problems [22-23].

Hyperparameters play a valid role in the accuracy of regression prediction algorithms. In practice, it is necessary to continuously adjust the hyperparameters, train the model under different sets of hyperparameters, and determine the best hyperparameters by comparing the model's performance. Therefore, finding the appropriate hyper-parameters has become a critical issue in machine learning [24].

Bayesian optimization (BO) is a very effective global optimization algorithm. BO is very suitable for solving highly complex optimization problems. Their objective functions could not be expressed, or the functions are non-convex, multimodal, and computational expensive [25]. BO can actively select appropriate evaluation points according to the relevant results of the current unknown function to avoid unnecessary sampling. At the same time, Bayesian optimization can use historical search information to improve search efficiency. [26]. BO has achieved better results than other hyperparameter tuning methods in the Black-Box Optimization Challenge 2020 [27].

III. Proposed Work

In this study, a new soft-sensor method for measuring the carbon content of fly ash is proposed by analyzing and experimenting with a total of 3,272,872 DCS data from an electric boiler from October 23 to November 30, 2020. The method combines data mining, feature engineering, LightGBM, and BO algorithm. A flowchart of the applied methodology is proposed in Fig.1. By comparing with other feature selection methods, and prediction models, experiments show that the prediction results of the presented approach are closer to the actual working conditions of the carbon content of fly ash, which improves the soft-sensor accuracy and ensures the reliability and accuracy of the soft-sensor method.

The data processing part is the operation of apparent outlier removal and re-sampling of the acquired DCS data.

A. Apparent Outlier Removal

First, the raw data was examined, and remove the data are outside the reasonable range. For example, the actual load recorded by the DCS has some invalid data at the beginning due to plant shutdown, etc. As shown in Figure 2, the data in the red area are unreasonable. By removing apparent outliers, the natural distribution of the variables can be captured.

The load changes drastically since thermal power units need to adjust the power generation capacity according to the grid load during operation. The thermal power units are constantly changing the working conditions, such as steady-transition-steady. This will result in a miss correlation between data. This effect can be minimized by data-resampling the data in an appropriate period. In this study, the datasets were re-sampled into 6-minute intervals. The actual load's scatter plot, before and after re-sampling, is shown in Fig.3, Fig.4. The re-sampled data is smoother and more similar to regular operation, as is shown in Fig.4.

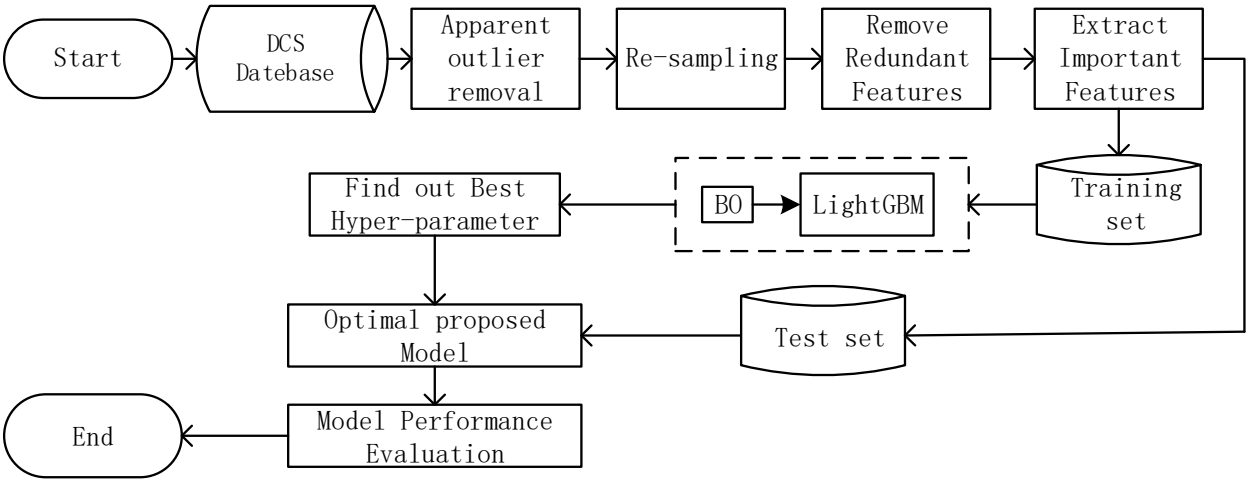

Fig. 1. Methodology Flowchart.

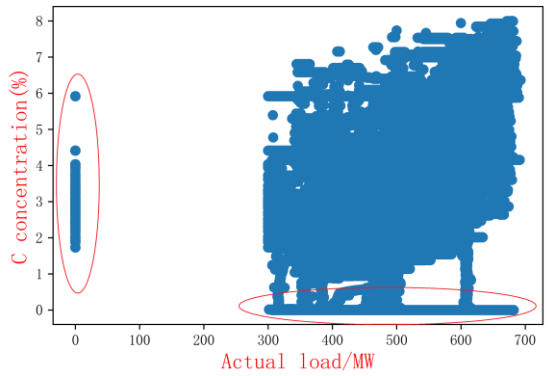

Fig. 2. Raw Load Data Scatter Plot.

B. Data Re-sampling

Feature engineering [28] is scoring each potential feature based on specific feature selection metrics and selecting representative variables from a given dataset to improve the final prediction. Feature engineering is crucial in the model design, as irrelevant or redundant data features will harm the model's performance. By reducing the number of variables, noisy and irrelevant data are removed, and the algorithm can run fast as the number of variables is reduced. There are generally three feature selection methods: filter method based on statistical information, wrapper method, and embedded method [29]. This study uses the correlation matrix(based on the filter method) and wrapper method to deal with variables.

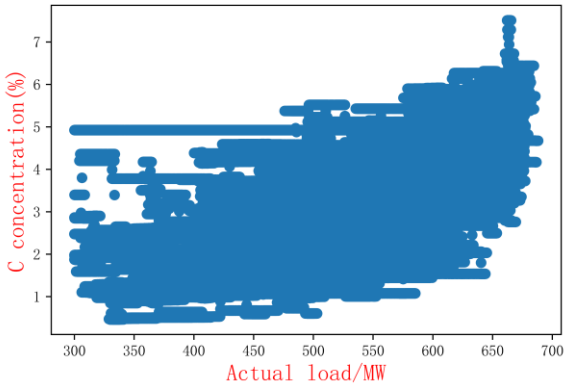

Fig. 3. Scatter Plot before Re-sampling.

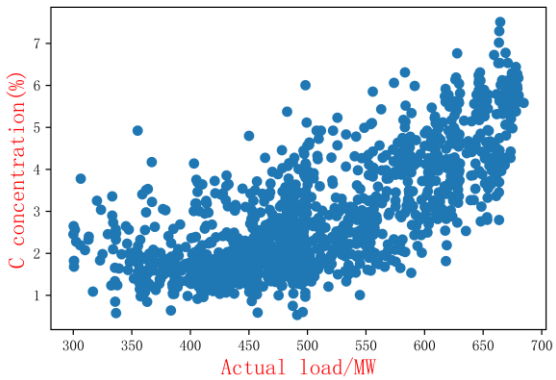

Fig. 4. Scatter Plot after Re-sampling.

According to the characteristics of multivariable variables, nonlinear and highly coupling thermal process, Firstly, the features with strong coupling are found through the correlation matrix, and the variables with low correlation with the carbon content of fly ash are eliminated. The essential variables are further extracted by the wrapper method to reduce the computational complexity of the model.

C. Remove Redundant Features

The correlation matrix (CM) is a table that is constructed to quantify the dependence between variables, as shown in equation (1), and the correlation coefficient indicates the positive or inverse relationship between the target variables [30]. The correlation matrix identifies and deletes redundant features in the dataset. Fig.5 shows one of the generated correlation matrices, which presents the correlation between the six features, The features from top to bottom are 'air temperature', 'air volume', 'steam temperature', 'steam temperature 2', 'air temperature 2', and 'air pressure'. If the correlation coefficient between the two variables is more significant than 0.95, they are compared with the carbon content in fly ash, and the variable with the smaller correlation coefficient is removed. For Fig.5, the features ' air temperature', 'air volume', and 'steam temperature ' were removed. In this way, the number of 71 DCS variables was reduced to 45.

$$r = \frac{\sum (x_i - x_{ave})(y_i - y_{ave})}{\sqrt{\sum (x_i - x_{ave})^2 \sum (y_i - y_{ave})^2}} \quad (1)$$

where r is the correlation coefficient, x_i is value of feature x, y_i is value of feature y, x_{ave} is mean value of the feature x, y_{ave} is mean value of the feature y.

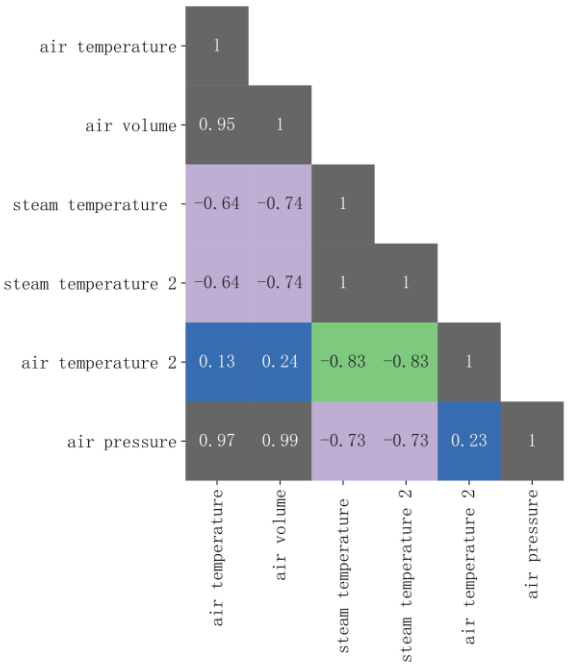

Fig. 5. Part of the Correlation Matrix.

D. Extract Important Features

The wrapper method is a feature selection method according to a specific prediction model, and this method uses recursive feature elimination (RFE). It is a greedy optimization algorithm that selects the best feature subset by repeated iteration. For the last step, the 45 variables selected by the correlation matrix are then used to determine the best performing 20 variables by the wrapper method.

E. Establishment of the Prediction Model

Before modeling, we will process features through correlation matrix and wrapper method, eliminate variables with highly coupling through correlation matrix, and select essential features subset by wrapper method, to reduce the computational complexity and further improve the expressiveness of the model.

LightGBM has many hyperparameters, and a reasonable choice of hyperparameters can improve prediction. While using the lighthGBM model to predict the carbon content of fly ash, the BO algorithm is used to continuously adjust the hyperparameters of the lighthGBM model to improve the prediction accuracy of the modelThe process of generating the optimal model BO_LightGBM is shown in Fig.6. During the model satisfaction assessment, cross-validation is set 5, the evaluation function is a root mean squared error.

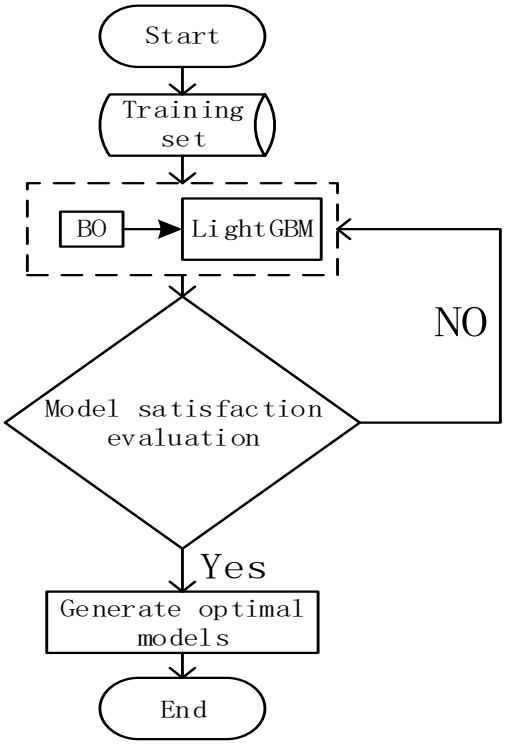

Fig. 6. Hyperparameter Tuning Flow Chart.

IV. Validation and Results

A. Performance Metrics

The regression evaluation indexes in regression analysis have mean squared error (MSE), root mean squared error (RMSE), mean absolute percentage error (MAPE), and coefficient of determination(R^2). $RMSE = \sqrt{MSE}$, In this study, RMSE, MAPE, and R^2 were selected as performance metrics, and these indices can be calculated as Eq. (3)(4)(5). R^2 is adapted to measure the approximation degree of the data to the prediction value, the closer R^2 is to 1, the better the fitting effect of the model. The smaller MSE, MAPE means the more accurate the prediction.

$$RMSE=\sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i-\hat{y}_i)^2} \tag{3}$$

$$MAPE=\frac{1}{m}\sum_{i=1}^{m}\left|\frac{y_i-\hat{y}_i}{\hat{y}_i}\right|\times 100\% \tag{4}$$

$$R^2=1-\frac{\sum_{i=1}^{m}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{m}(y_i-\bar{y}_i)^2} \tag{5}$$

where y_i are the actual values, $\hat{y}_i$ are the predicted values,and $\bar{y}_i$ is the mean of y_i (i=1,2,..., n).

B. Performance Comparison of Feature Processing Methods

Methods frequently used in feature processing are Random Forest (RF) [31] and Pearson correlation coefficient (PCC) [32]. The experiment compares the method proposed in this paper with RF and PCC. The optimal 20 features of the three methods are modeled for prediction while ensuring that the selected data are consistent with the boiler's steady-state operating conditions. In this process, the correlation matrix eliminates the features with high correlation, retains 45 features with low coupling, and then uses the packaging method to retain 20 features. The processed results of each of the three methods are used as input to the LightGBM model. The experimental results are listed in Table I.

The random forest method model is less effective, as shown in Table I. The model treated with PCC outperformed the RF. After using the correlation matrix to process the features, the R^2, MAPE, and RMSE were significantly optimized, and after further processing of features by wrapper method, the R^2 was improved to 0.822, MAPE was reduced to 16.5%, and RMSE was reduced to 0.509. It is proved that the fitting effect of the model using the method proposed in this paper is further enhanced, and the error is further reduced, which effectively solves the problem of high correlation and strong coupling between variables.

TABLE I. Prediction Results after Feature Processing

Method	R^2	MAPE	RMSE
RF	0.71	19.22%	0.644
PCC	0.78	17.47%	0.573
CM	0.814	16.69%	0.523
CM+Wrapper	0.822	16.50%	0.509

C. Comparison of Model Prediction Performance

To validate the superiority of the proposed method, three methods, including LM- Garson-BP [16] AQPSO-SVR [19] FPA-RF [12], the three latest methods are compared as benchmarks.

- LM-Garson-BP: The LM-Garson-BP methods used sensitivity analysis to select the feature variables and then used BP neural networks for predictive modeling and genetic algorithms to optimize the connection weights, number of neurons, and number of hidden layers.
- AQPSO-SVR: The AQPSO-SVR method first adds adaptive perturbation to the quantum particle swarm optimization (QPSO) algorithm and uses this improved algorithm to find the optimal hyper-parameters of the support vector regression (SVR).
- FPA-RF: The FPA-RF method first uses the random forest method to filter features, then uses the random forest as a prediction model and uses the flower pollination (FPA) algorithm to optimize the hyperparameters of the random forest.

MAPE, RMSE, and R^2 are selected as evaluation indexes. The experimental results are shown in Table II, and the prediction comparison results are listed in Fig. 7 and Fig. 8.

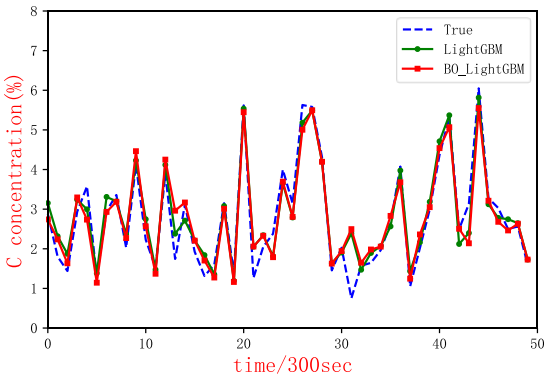

Fig. 7. Prediction Comparison of LightGBM Model and BO_LightGBM.

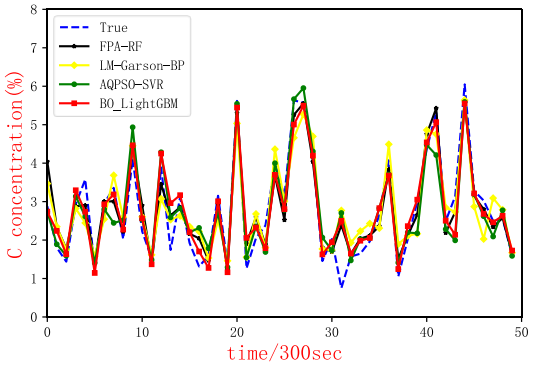

Fig. 8. Prediction Comparison of different Combined Models.

TABLE II. PREDICTION RESULTS OF DIFFERENT MODELS

Model	R^2	MAPE	RMSE
LM-Garson-BP [16]	0.722	19.69%	0.658
AQPSO-SVR [19]	0.786	18.22%	0.560
FPA-RF [12]	0.696	19.37%	0.652
LightGBM	0.822	16.50%	0.509
BO_LightGBM	0.831	16.02%	0.494

Table II shows the performance comparison between the proposed method and other methods. The obtained results are representative. The method proposed in this paper achieves the lowest MAPE, RMSE, and the highest R^2. The method in this paper reduces RMSE by 2.9%~24.9% and MAPE by 2.9%~18.6% compared with the above methods, indicating a further reduction of errors and improvement of measurement accuracy. The R^2 was improved by 1.1%~15.1%, indicating that the prediction curves were better fitted and the method in this paper was more accurate and reliable. Specifically, LM-Garson-BP, AQPSO-SVR, and FPA-RF all use heuristic algorithms for hyperparameter tuning and combine with regression models for prediction, which improves the prediction accuracy of the corresponding models to some extent.

From the perspective of hyper-parameter tuning, The BO algorithm can find the next evaluation position based on the information obtained for the unknown objective function when facing a complex optimization problem with hyperparameter tuning that is non-convex, multimodal, and computational, to reach the optimal solution the fastest [25]. The BO algorithm avoids the issues of ineffective use of iterative feedback information and the slow search speed of the algorithm. From the perspective of the prediction model, the LightGBM algorithm objective function adopts the second-order Taylor expansion, which can fully learn the model, add regular terms, reduce the complexity of the model, prevent overfitting, support parallel and distributed computing, and effectively improve the prediction accuracy. Therefore, the prediction results are better compared to the four models compared.

V. CONCLUSION

In this study, a data-driven approach integrating various machine learning algorithms and data mining techniques is used for the first time to analyze the relationship between the carbon content of fly ash and various operating parameters of boilers. This method has practical significance for guiding boiler production; collecting data for 37 days of complete working conditions and comparing our feature processing method with the PCC method, the RF method. The performance of the model is compared with LM-Garson-BP, AQPSP-SVR, and FPA-RF models. The results demonstrate that the method used in this paper exhibits better prediction results.

In future work, consider combining DCS data and coal type characteristics to improve accuracy by more data information. In addition, due to the high correlation between DCS data and time, it is worthwhile to study more deeply how to mine

valuable information from these unstructured time series data and find the intrinsic correlation between the time series data.

ACKNOWLEDGMENT

This research is supported by National Natural Science Foundation of China (52176110); Scientific Research Project of Education Department of Hubei Province (D20191708); Intellectual Property Promotion Project of Colleges and Universities of Hubei Province (GXYS2018009).

REFERENCES

[1] Li R, Wei K, Huang Q. A novel method for precise measurement of unburnt carbon in boiler fly ash by ECSA® based on TG-MS[J]. Fuel, 2020, 264: 116849.

[2] Brown R C, Dykstra J. Systematic errors in the use of loss-on-ignition to measure unburned carbon in fly ash[J]. Fuel, 1995, 74(4): 570-574.

[3] Liu R, Deguchi Y, Nan W, et al. Unburned carbon measurement in fly ash using laser-induced breakdown spectroscopy with short nanosecond pulse width laser[J]. Advanced Powder Technology, 2019, 30(6): 1210-1218.

[4] Liu H, Tan H, Gao Q. Microwave attenuation characteristics of unburned carbon in fly ash[J]. Fuel, 2010, 89(11): 3352-3357.

[5] CHENG Qi-ming, HU Xiao-qing, et al.Summary of Measurement Methods of Carbon Content in Fly Ash[J]. Journal of Shanghai University of Electric Power, 2011, 27(5): 519-524.

[6] H.Y. Liu, H.Z. Tan, Q.A. Gao, et al. Microwave attenuation characteristics of unburned carbon in fly ash, Fuel 89 (2010) 3352–3357.

[7] Xiao, Jianli. "SVM and KNN ensemble learning for traffic incident detection." Physica A: Statistical Mechanics and its Applications 517 (2019): 29-35.

[8] Development of soft sensors for isomerization process based on support vector machine regression and dynamic polynomial models[J]. Journal of Robotics & Machine Learning,2019, 149:95-103.

[9] Liu R, Chen P, Wang Z. Quantitative analysis of carbon content in fly ash using LIBS based on support vector regression[J]. Advanced Powder Technology, 2021, 32(8): 2978-2987.

[10] Amin S M. Smart grid: Overview, issues and opportunities. advances and challenges in sensing, modeling, simulation, optimization and control[J]. European Journal of Control, 2011, 17(5-6): 547-567.

[11] Yan W, Tang D, Lin Y. A data-driven soft sensor modeling method based on deep learning and its application[J]. IEEE Transactions on Industrial Electronics, 2016, 64(5): 4237-4245.

[12] WANG Fang1, MA Suxia1, WANG He. Prediction model of carbon content in fly ash using random forest variable selection method[J]. Thermal Power Generation, 2018, 47(11): 89-95.

[13] ZHOU Hao,ZHU Hong-bo, et al. Artificial neural network modelling on the unburned carbon in fly ash from utility boilers[J].Proceedings of the CSEE,2002(06):97-101.

[14] WANG Chun-lin, Zhou Hao, et al. Support vector machine modeling on the unburned carbon in fly ash[J]. Proceedings of the CSEE,2005(20):72-76.

[15] Zhang P. A novel feature selection method based on global sensitivity analysis with application in machine learning-based prediction model[J]. Applied Soft Computing, 2019, 85: 105859.

[16] Zhu Jinqi,Niu Xiaofan,Xiao Xianbin. Prediction models of the carbon content of fly ash in a biomass boiler based on improved BP neural networks [J]. Renewable Energy Resources, 2020,38(02):150-157(in Chinese).

[17] Wen X. Modeling and performance evaluation of wind turbine based on ant colony optimization-extreme learning machine[J]. Applied Soft Computing, 2020, 94: 106476.

[18] Feng Xugang, Qian Jiajun, Zhang Jiayan. Prediction method of unburned carbon content in fly ash based on genetic neural network with sensitivity analysis [J].Journal of electronic measurement and instrumentation,2016,30(7): 1083-1089(in Chinese).

[19] PENG Dao-gang, LI Dan-yang, et al.Research on Prediction of Carbon Content in Boiler Fly Ash Based on ADQPSO ─ SVR[J]. Computer Simulation,2020,37(03):72-77(in Chinese).

[20] Ke G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree[J]. Advances in neural information processing systems, 2017, 30: 3146-3154.

[21] J. Friedman. Greedy function approximation: a gradient boosting machine. Annals of Statistics, 29(5):1189–1232, 2001.

[22] Chen C, Zhang Q, Ma Q, et al. LightGBM-PPI: Predicting protein-protein interactions through LightGBM with multi-information fusion[J]. chemometrics and intelligent laboratory systems, 2019, 191: 54-64.

[23] Ju Y, Sun G, Chen Q, et al. A model combining convolutional neural network and LightGBM algorithm for ultra-short-term wind power forecasting[J]. Ieee Access, 2019, 7: 28309-28318.

[24] Probst P, Boulesteix A L, Bischl B. Tunability: importance of hyperparameters of machine learning algorithms[J]. The Journal of Machine Learning Research, 2019, 20(1): 1934-1965.

[25] Cui JX, Yang B. Survey on Bayesian Optimization Methodology and Applications. Journal of Software, 2018, 29(10): 3068-3090(in Chinese).

[26] Wu J, Chen X Y, Zhang H, et al. Hyperparameter optimization for machine learning models based on Bayesian optimization[J].Journal of Electronic Science and Technology, 2019, 17(1): 26-40.

[27] Turner R, Eriksson D, McCourt M, et al. Bayesian optimization is superior to random search for machine learning hyperparameter tuning: Analysis of the black-box optimization challenge 2020[J]. arXiv preprint arXiv:2104. 10201, 2021.

[28] Tsai C F. Feature selection in bankruptcy prediction[J]. Knowledge-Based Systems, 2009, 22(2): 120-127.

[29] Yvan S , I Iñaki, Pedro L . A review of feature selection techniques in bioinformatics[J]. Bioinformatics, 2007(19):2507-2517.

[30] Asuero A G, Sayago A, Gonzalez A G. The correlation coefficient: An overview[J]. Critical reviews in analytical chemistry, 2006, 36(1): 41-59.

[31] Breiman L. Random forests[J]. Machine learning, 2001, 45(1): 5-32.

[32] Benesty J, Chen J, Huang Y, et al. Pearson correlation coefficient[M]//Noise reduction in speech processing. Springer, Berlin, Heidelberg, 2009: 1-4.