IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 25 January 2024, accepted 21 February 2024, date of publication 5 March 2024, date of current version 11 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3372660

RESEARCH ARTICLE

Improved LS-SVM Boiler Combustion Model Based on Affinity Propagation

MING YAN[1], LIANG WANG[1], MEILING ZHANG[2], AND PAN SHI[1]

[1]Technical Supervision Center, Huadian Electric Power Research Institute Company Ltd., Hangzhou 310030, China
[2]Chongqing University, Chongqing 400044, China

Corresponding authors: Liang Wang (swu_wl@163.com) and Meiling Zhang (202303042140@stu.cqu.edu.cn)

This work was supported by the Science and Technology Project of China Huadian Corporation Ltd. ''Cavitation Monitoring and Economic Operation Research and Application of Large-Scale Hydroelectric Units'' under Grant CHDKJ20-02-153.

ABSTRACT In the global effort to promote green energy policies, understanding and optimizing boiler combustion processes in coal-fired power plants is crucial. During unit start-ups, shutdowns, and load deep peak regulation, significant energy-saving potential can be harnessed in boilers. This paper focuses on a 600MW supercritical coal-fired power unit and presents an improved Least Squares Support Vector Machine (LS-SVM) model with refined initial parameters. By combining the improved LS-SVM with Affinity Propagation (AP) clustering, a combustion efficiency model for boilers is constructed. The experimental results demonstrate that the AP-based improved LS-SVM model not only reduces computational complexity and training time but also enhances predictive accuracy and generalization performance.

INDEX TERMS Boiler efficiency, AP clustering algorithm, LS-SVM, hybrid modeling, oxygen content of flue gas, carbon content in fly ash.

I. INTRODUCTION

Coal-fired power plants play a significant role in the global electricity supply, but they face considerable environmental and energy challenges. Despite the worldwide push for green energy policies, coal-fired units must adapt to system load fluctuations during start-ups, shutdowns, and load deep peak regulation to ensure power grid stability.

However, these operational changes can lead to increased coal consumption, reduced energy efficiency, but also offer significant energy-saving potential. In the era of energy conservation and emission reduction, combustion optimization has become a primary means of regulation [1]. Yet, the intricate nonlinear, time-varying, and coupled nature of combustion processes poses a formidable challenge for predicting and controlling boiler combustion efficiency [2], [3], [4]. Traditional boiler combustion efficiency prediction methods struggle to capture nonlinear relationships, thus inadequately representing the complexity of combustion processes.

Economic adjustments in boiler combustion encompass reference indicators such as the net calorific value of coal, oxygen content of flue gas, and carbon content in fly ash. Constructing a boiler efficiency model based on these indicators and performing real-time optimization can significantly enhance combustion efficiency [5], [6], [7]. However, in real coal-fired power plants, accurate measurement of boiler parameters is difficult due to the absence of real-time net calorific value of coal measurement devices, inaccurate and costly carbon content in fly ash measurement, as well as the high cost of oxygen content sensors. Obtaining accurate real-time data for boiler efficiency is particularly challenging during unit start-ups, shutdowns, and load deep peak regulation.

The pursuit of an efficient and accurate method for predicting combustion efficiency has become a focal point in research. In recent years, the integration of machine learning and data mining techniques into the energy sector has led to new perspectives and tools for predicting boiler combustion efficiency. Various methods, such as hybrid Least Squares Support Vector Machine (LS-SVM) with factor analysis [8], [9], distributed extreme learning machines [10],

The associate editor coordinating the review of this manuscript and approving it for publication was Oussama Habachi.

Deep Bidirectional Learning Machine [11], and neural networks [12], [13], treat boiler efficiency as the direct output of machine learning models. However, these models primarily rely on hardware measurement values, such as oxygen content of flue gas and carbon content in fly ash, as input variables. The accuracy of these initial input variables is often insufficient, which limits predictive accuracy. To address this issue, a preliminary step could involve using hybrid modeling techniques to tackle the challenges posed by hard-to-measure parameters, followed by integrating them as input variables into the model.

Support Vector Machines (SVM) are a powerful tool for classification and regression in machine learning. LS-SVM, a variant of SVM, trains models by minimizing the square sum of the loss function and has exceptional nonlinear fitting capability. Its ability to handle small-sample data and high-dimensional features makes it outstanding in predictive problems. For example, Wu et al. [14] used LS-SVM to construct a combustion model for coal-fired power plants, capturing the complex nonlinear features of combustion processes through historical data to predict carbon content in fly ash and boiler efficiency. However, due to exclusive reliance on LS-SVM for prediction, limitations arise in handling large-sample prediction, hindering real-time precise online prediction based on abundant data from power plants.

To address this issue, multi-model methods, especially those considering the boiler combustion process's multivariable and severe nonlinearity, can be introduced. Affinity propagation (AP) algorithm is a clustering method suitable for multi-model modeling and has been proven to be a superior algorithm compared to others [15], [16], [17]. It autonomously determines the sample center [18], which is critical in boiler combustion modeling. In this paper, we select a 600MW supercritical boiler with opposed wall swirling burners and direct-fired system as the research object. We improve the LS-SVM method by employing a hybrid modeling approach to address hard-to-accurately-measure parameters, using them as model input variables. Simultaneously, by combining AP with LS-SVM, we enhance the classification performance, generalization performance, and prediction accuracy of the boiler combustion model. This research makes a noteworthy contribution by presenting a pioneering methodology that amalgamates the improved LS-SVM technique with Affinity Propagation, effectively tackling the challenges linked to hard-to-measure parameters.

II. MATHEMATICAL MODEL

In this paper, An AP clustering method is employed to enable multi-model modeling of boiler combustion. The AP algorithm clusters the test sample set into multiple categories, each resulting in a sub-model tailored to a specific combustion characteristic, such as the oxygen content of flue gas and the carbon content in fly ash. In constructing each sub-model, we employ an improved LS-SVM method to address the intricacies posed by challenging-to-measure parameters, using techniques including mechanism analysis and factor analysis. These parameters are then integrated as input variables, enhancing predictive accuracy and generalization ability.

A. AP CLUSTERING

The multi-modeling algorithm based on AP Clustering is illustrated in Figure 1. Initially, the sample data undergoes clustering using the AP Clustering method, resulting in several subclasses (denoted as m classes in Figure 1). Subsequently, algorithms such as LS-SVM are employed to independently train models for each subclass, generating distinct sub-models. For testing samples, a classification is initially performed using a similarity measure, followed by individual predictions and outputs from the corresponding sub-models.

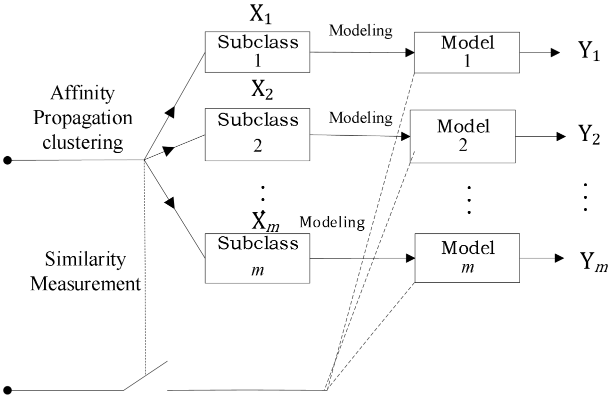

FIGURE 1. Example of an AP clustering algorithm.

The AP clustering method is known for its efficiency and speed. It starts by identifying potential cluster centers, followed by iterative similarity calculations to refine these centers and improve the accuracy of multi-model clustering.

The underlying principle of AP involves defining the similarity between any two sample points x_i and x_j in a sample space containing N sample points, with the function $S(i,j)$ ($S_{i,j} \in \boldsymbol{S}$), which can be measured by negative Euclidean distance $S_{i,j} = -\text{dcor}(x_i - x_j)$ or any similarity/distance function, as our goal is to maximize the correlation among group members. The results will be stored in the N × N similarity matrix $\boldsymbol{S}$ by using an expression like Eq. (1). For the AP algorithm, it is not necessary to define the number of clusters. However, the preference value $S_{k,k}$ for each of the diagonal elements in $\boldsymbol{S}$ must be provided. The diagonal elements of the matrix represent the bias parameter p, indicating the degree to which x_i becomes a clustering center, with an initial value set as the median of $\boldsymbol{S}$. Adjusting p can change the clustering results, and when p is less than a certain threshold, it will cause a change in the number of classifications. Preferences should have the same value, i.e., the median of the similarities, to give all the points the probability of being a prototype. The larger $p(S_{k,k})$, the higher probability of being chosen as a prototype and may result in more clusters.

$$\begin{bmatrix} x_1 \cdots x_i \cdots x_j \cdots x\mathrm{N} \end{bmatrix}$$
$$\Rightarrow \begin{bmatrix} S(1,1) & \cdots & \cdots & \cdots & S(1,\ \mathrm{N}) \\ \cdots & \cdots & S(i,j) & \cdots & \cdots \\ \cdots & S(j,i) & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ S(\mathrm{N},1) & \cdots & \cdots & \cdots & S(\mathrm{N},\ \mathrm{N}) \end{bmatrix} \tag{1}$$

The previous message passing procedure relies on iterative updates of the responsibility matrix **R** and the availability matrix **A**. To find a suitable clustering center x_k, the AP algorithm continually gathers evidence for **R** and **A** from data samples. $R(i,k)(R_{i,k} \in \mathbf{R})$ accumulates evidence of how suitable is x_k to be the representative (prototype) for x_i if both would be in the same cluste (i.e. the extent to which sample point x_k suits being the clustering center for x_i). R is calculated by using Eq. (2). $A(i,k)$ ($A_{i,k} \in \mathbf{A}$) represent the availability of x_k to be chosen as the representative for point x_i (i.e. how well x_i chooses x_k as its clustering center). Non-diagonal elements of matrix **A** are calculated by using (3) and diagonal elements are shaped by (4). The iterative formulas are as follows:

$$R_{i,k} = S_{i,k} - \max_{j \neq k} \{A_{i,j} + S_{i,j}\} \tag{2}$$

$$A_{i,k} = \min \{0,\ R_{k,k} + \sum_{i' \notin \{i,k\}} \max\{0,\ R_{i',k}\}\} \tag{3}$$

$$A_{k,k} = \sum_{i' \neq k} \max\{0,\ R_{i',k}\} \tag{4}$$

where i, j, and k are integers in the 1-to-N closed range.

The AP clustering is a process of constantly iterating and updating the evidence according to formula (2), (3) and (4). The iteration termination condition is as follows: for any sample x_i, if sample x_k makes **R** +**A** the maximum value in $R(i,\ j) + A(i,\ j),\ j = 1,2,\ldots,\mathrm{N}$, then the sample x_k is the clustering center of x_i.By means of iterative competition, AP clustering can obtain the optimal clustering center and the category of each sample [18].

To assess clustering quality, the k-index is introduced, calculated as follows:

$$S_{il}(t) = \frac{\min\{d\ (t,\ C_i)\} - a(t)}{\max\{a(t), \min\{d\ (t,\ C_i)\}\}} \tag{5}$$

In Formula (5), C_i ($i = 1,2,\ldots,k$) denotes the i-th category, d(t, C_i) represents the average distance between sample t in C_i and all samples in another category C_j, while $a(t)$ signifies the average distance between sample t in C_i and other samples within C_i. For a dataset of all sample points, the average value of the k-index is:

$$S_{il\_av} = mean\left[\sum_{t=1}^{n} S_{il}(t)\right] \tag{6}$$

S_{il_av} ranges between [0, 1], displaying a positive correlation. To achieve clear separation between different sub-classes, S_{il_av} should not be less than 0.5. In cases where separation between subclasses is difficult, its value usually doesn't exceed 0.2.

B. LEAST SQUARES SUPPORT VECTOR MACHINE

SVM is suitable for small and nonlinear training samples with kernel approaches for classification and regression. In order to calculate simplified, Suykens and Vandewalle propose a modification algorithm of standard SVM algorithm as LS-SVM at 1999 [19]. The most important difference is that LS-SVM uses a set of linear equations for training while SVM uses a quadratic optimization problem. In this paper, LS-SVM is used to get regression model.

The depiction of LS-SVM for regression problems is delineated as follows:

Consider a training dataset $\{(x_i, y_i) \mid i = 1,2,\ldots,\mathrm{N}\}$, where N is the number of training samples, $x_i \in \mathbf{R}^{\mathrm{N}}$ is the input for the i-th sample, and $y_i \in \mathbf{R}$ is the corresponding output. The input space $\mathbf{R}^{\mathrm{N}}$ is mapped to a high-dimensional feature space $\mathbf{Z}$ through a nonlinear function $\varphi(x_i)$. In $\mathbf{Z}$, an unknown nonlinear function is estimated using an expression like (7), where w and b are undetermined parameters.

$$y(x) = w^T\varphi(x) + b,\ w \in \mathbf{Z}, \quad b \in \mathbf{R} \tag{7}$$

The optimization problem for LS-SVM is defined as:

$$\min_{w,e} J(w,\ e) = \frac{1}{2}w^T w + \frac{\gamma}{2}\sum_{i=1}^{\mathrm{N}} e_i^2, \quad \gamma > 0 \tag{8}$$

Subject to the equality constraint:

$$y_i = w^T\varphi\ (x_i) + b + e_i, \quad i = 1, 2, \cdots, \mathrm{N} \tag{9}$$

Here, the first term of the objective function corresponds to the model's generalization capability, and the second term represents the model's accuracy. γ serves as a compromise parameter between model generalization and accuracy, subject to human tuning. e_i represents the error between the actual and predicted output for the i-th data.

Solving the Lagrange function for the optimization problem yields the optimal solution [20]:

$$\begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} 0 & I^T \\ l & \Omega + \gamma^{-1}l \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ Y \end{bmatrix} \tag{10}$$

Here, the vectors are denoted as $Y = [y_1,\ y_2,\ \ldots,\ y_{\mathrm{N}}\ ]^T$, $\mathit{1} = [1,\ 1,\ \ldots,\ 1]^T$, and $a = [a_1,\ a_2,\ \ldots,\ a_{\mathrm{N}}\ ]^T$. Ω is an N × N symmetric matrix, where $\Omega_{i,j} = \varphi(x_i)^{\mathrm{T}}\varphi(x_j) = K(x_i,\ x_j)(i,j = 1,2,\ldots,\mathrm{N})$, and $K(\bullet,\ \bullet)$ is the kernel function. The final expression for the LS-SVM model is:

$$y(x) = \sum_{\mathrm{i}=1}^{\mathrm{N}} a_i K(x,\ x_i) + b \tag{11}$$

where $a_i \in \mathbf{R}(i = 1,2,\ldots,\mathrm{N})$, and $K(x, x_i)(i = 1,2,\ldots,\mathrm{N})$ are any kernel functions satisfying the Mercer condition [19].

After establishing these LS-SVM sub-models, testing samples can be assessed based on the degree of similarity,

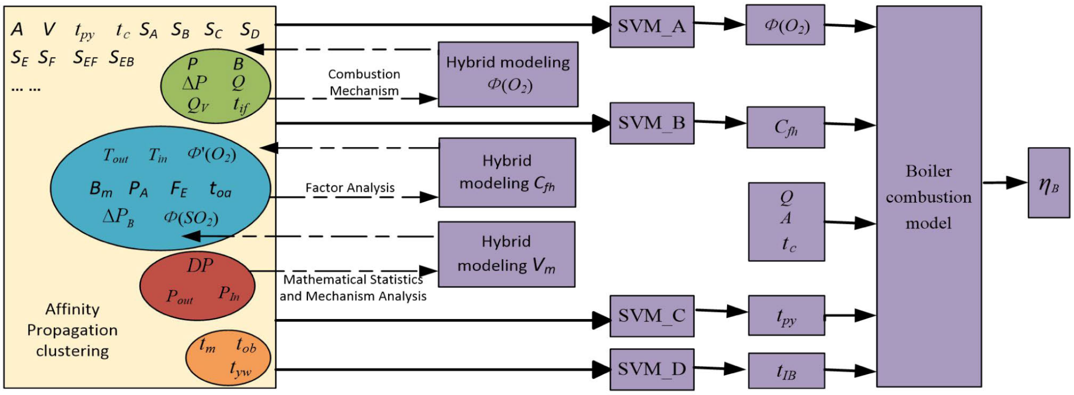
A V t_{py} t_c S_A S_B S_C S_D
S_E S_F S_{EF} S_{EB}
… …
P B ΔP Q Q_V t_{if}
T_{out} T_{in} $\Phi'(O_2)$
B_m P_A F_E t_{oa}
ΔP_B $\Phi(SO_2)$
DP P_{out} P_{In}
t_m t_{ob} t_{yw}
Affinity Propagation clustering
Combustion Mechanism
Factor Analysis
Mathematical Statistics and Mechanism Analysis
Hybrid modeling $\Phi(O_2)$
Hybrid modeling C_{fh}
Hybrid modeling V_m
SVM_A
SVM_B
SVM_C
SVM_D
$\Phi(O_2)$
C_{fh}
Q A t_c
t_{py}
t_{IB}
Boiler combustion model
η_B

FIGURE 2. Structure of the boiler combustion model.

selecting different sub-models for testing output, and utilizing them as the ultimate output.

In cases where separation hybrid modeling approach, we establish soft-sensing models for two subclasses: The oxygen content of flue gas and the carbon content in fly ash. These soft-sensing values are then integrated back into the sample database. Following this, utilizing the LS-SVM technique, we develop models for various parameters, including oxygen content of flue gas (SVM_A), carbon content in fly ash (SVM_B), exhaust gas temperature (SVM_C), and temperature (SVM_D). After training these models, a simplified algorithm is employed to calculate boiler average furnace efficiency [21], [22], [23].

III. NUMERICAL FORMULATION

A. STRUCTURE OF BOILER COMBUSTION MODEL

The structure of our boiler combustion model, as illustrated in Figure 2, entails an initial application of the AP algorithm to cluster data from the power plant's sample database. This clustering process results in the creation of multiple sub-models. Subsequently, through a hybrid modeling approach, we establish soft-sensing models for two subclasses: The oxygen content of flue gas and the carbon content in fly ash. These soft-sensing values are then integrated back into the sample database. Following this, utilizing the LS-SVM technique, we develop models for various parameters, including oxygen content of flue gas (SVM_A), carbon content in fly ash (SVM_B), exhaust gas temperature (SVM_C), and average furnace temperature (SVM_D). After training these models, a simplified algorithm is employed to calculate boiler efficiency.

NOTATIONS

A	Received the ash content of the fuel.
V	Volatile fraction of fuel.
t_{py}	Exhaust gas temperature.
t_c	Reference temperature.
S_A,S_B,S_C,S_D,S_E,S_F	Secondary air damper openings of layers A, B, C, D, E, F.
S_{EF}, S_{EB}	Front and rear wall burnt-out air damper openings.
P	Unit load.
B	Total fuel consumption.
ΔP	Front and rear air preheater differential pressure.
Q	Net calorific value of coal.
Q_V	Air flow content.
t_{if}	Average inlet air temperature of forced draft fans A and B.
T_{out}, T_{in}	Pulverizers outlet and inlet temperature.
B_m	Pulverizers load.
P_A	Primary air pressure.
F_E	Frequency of pulverizers separator.
t_{oa}	Average temperature at the outlet of air preheaters A and B.
$\Delta \mathrm{P}_B$	Difference between the secondary air total pressure and the furnace pressure.
$\Phi(SO_2)$	SO_2 concentration at the boiler outlet.
DP	Inlet and outlet differential pressure of pulverizers.
P_{out}, P_{in}	Pulverizers outlet and inlet pressure.
V_m	Mixed air flow rate entering pulverizers.
t_m	Temperature at the center of the combustion flame.
t_{ob}	Outlet temperature of the furnace.
t_{yw}	Flue gas temperature when the smoke temperature probe is inserted.
t_{IB}	Average temperature inside the furnace.
$\Phi(O_2)$	Oxygen content of the flue gas at the boiler tail.
C_{fy}	Carbon content in the fly ash;.
η_B	Represents the boiler efficiency.

As in Figure 2, the symbols are interpreted as follows:

To construct the sub-models, DCS soft-sensing technology is initially utilized for real-time online hybrid modeling. This process involves various techniques such as mechanism analysis, mathematical statistics, factor analysis, and experimental modeling, either individually or in combination. These methods contribute to the creation of sub-models for the economic indicators of flue gas oxygen content and fly ash carbon content, ultimately enhancing the accuracy of crucial input variables within the boiler efficiency model. These sub-models provide optimal setpoints for various operational variables, serving as vital references for real-time combustion economy adjustments.

B. SUB-MODEL FOR OXYGEN CONTENT OF FLUE GAS

The calculation of flue gas oxygen content necessitates consideration of the net calorific value of coal. For subcritical drum boilers, the direct energy balance (DEB) method is employed to estimate the net calorific value of coal. Conversely, for supercritical boilers, we employ a specific coal consumption coefficient method [24]. This coefficient, denoted as K, encapsulates the principle of energy conservation, as expressed in Formula (12), which enables us to estimate the net calorific value of coal:

$$K = \frac{B \div P}{B_0 \div P_0} \tag{12}$$

In Formula (12), K is usually confined to a range of 0.7 to 1.3. B represents the actual total fuel consumption entering the furnace (primarily the total coal consumption, including oil consumption converted to coal consumption during oil gun use). P represents actual power generation load, P_0 corresponds to various load points under the design coal type, and B_0 corresponds to the fuel quantity at each load point under the design coal type. Special scenarios like Run Back (RB) and high-pressure heater exit handling are also considered.

For direct-fired system, an approach simulating several minutes of coal feeding is used to calculate actual total fuel consumption B and actual load P. The total coal feed signal undergoes transmission through a series of third-order and first-order inertia links to derive the corresponding boiler steam generation capacity. This method offers greater accuracy compared to algorithms utilizing averaged actual loads and fuel consumption values over a few minutes, as it better reflects the characteristics of the controlled object. The time constant t of this inertia link is obtained through on-site simulation experiments, and its transfer function is as shown in Formula (13):

$$t = G(s) = \frac{0.8/(s+1)^3}{5.1/(100s+1)} \tag{13}$$

Utilizing the transfer function above, we derive net calorific value of coal Q:

$$Q = \frac{Q_s}{K(t)} \tag{14}$$

In Formula (14), Q_s represents net calorific value of the design coal, and $K(t)$ signifies the coal consumption ratio coefficient value after a time constant t.

Subsequently, guided by the basic principles of boiler combustion, the coal combustion inside the furnace occurs at high temperatures. Under these conditions, the combustion of combustible substances and oxygen from the air takes place via high-temperature exothermic chemical reactions. The formula for calculating oxygen content in the flue gas at the tail end of the boiler is given [25]:

$$\Phi(O_2) = \frac{Q_v - Q_r B}{Q_v + (Q_{FT} - Q_r)\,B} \tag{15}$$

In Formula (15), Q_v represents the air flow content entering the furnace (adjusted using the DCS-calculated value of secondary air density and considering leakage coefficient, corrected using the air preheater's front and rear differential pressure values under corresponding conditions); B represents the total fuel consumption entering the furnace. Q_r and Q_{FT} respectively represent the theoretical air required for complete combustion of 1 kg of entering coal and the theoretical flue gas volume (adjusted using the DCS-calculated flue gas density), with units of m^3/kg.

The formula for calculating air density is as follows:

$$\rho = \frac{\rho_0 T_0 p x}{p_0 T} \tag{16}$$

In Formula (16), ρ, p_x, T respectively represent the density, pressure, and thermodynamic temperature of dry air under other states, while ρ_0, p_0, T_0 represent the density, pressure, and thermodynamic temperature of dry air under standard conditions. They are measured in units of kg/m^3, kPa, and K.

For standard conditions ($T_0 = 273$ K and $p_0 = 101.3$ kPa), represents the density of dry air under normal composition ($\rho_0 = 1.293$ kg/m^3). Here, p_x represents atmospheric pressure, which is transmitted to DCS via transmitters installed on-site to measure atmospheric pressure. T represents the average air temperature at the inlet of forced draft fans A and B, i.e., t_{if}.

The calculation formulas for Q_r and Q_{FT} are given by:

$$Q_r = 0.0889\left(C_{ny} + 0.375\,S_{ny}\right) + 0.265\,H_{ny} - 0.0333\,Q_{ny} \tag{17}$$

$$Q_{FT} = \frac{1.866\,C_{ny} + 0.7\,S_{ny} + 11.1\,H_{ny} + 1.24\,W_{ny} + 0.8\,N_{ny}}{100} + 0.79\,Q_r \tag{18}$$

Through Formula (17) and Formula (18), it is evident that despite the relatively simple combustion mechanism of the boiler, real-time calculations are challenging due to the necessity of promptly knowing the reference content of various elements. However, as ultimate analysis of coal is intricate and entails lag, power plants generally perform proximate analysis, and DCS lacks real-time measurement data for various elements. Through extensive analysis of coal properties, we discern that coal primarily consists of combustible elements such as carbon, hydrogen, and sulfur. Under complete combustion conditions, these elements generate CO_2, SO_2, and H_2O. According to relevant literature, the air required for the combustion of 1 kg of coal per unit mass shows an approximately linear relationship with net calorific value of coal. Thus, in the absence of fuel element analysis data, net calorific value of coal can be used as an approximation to calculate the theoretical air quantity [26]:

$$Q_r = 0.251Q/1000 + 0.278 \tag{19}$$

$$Q_{FT} = 0.248Q/1000 + 0.77 \tag{20}$$

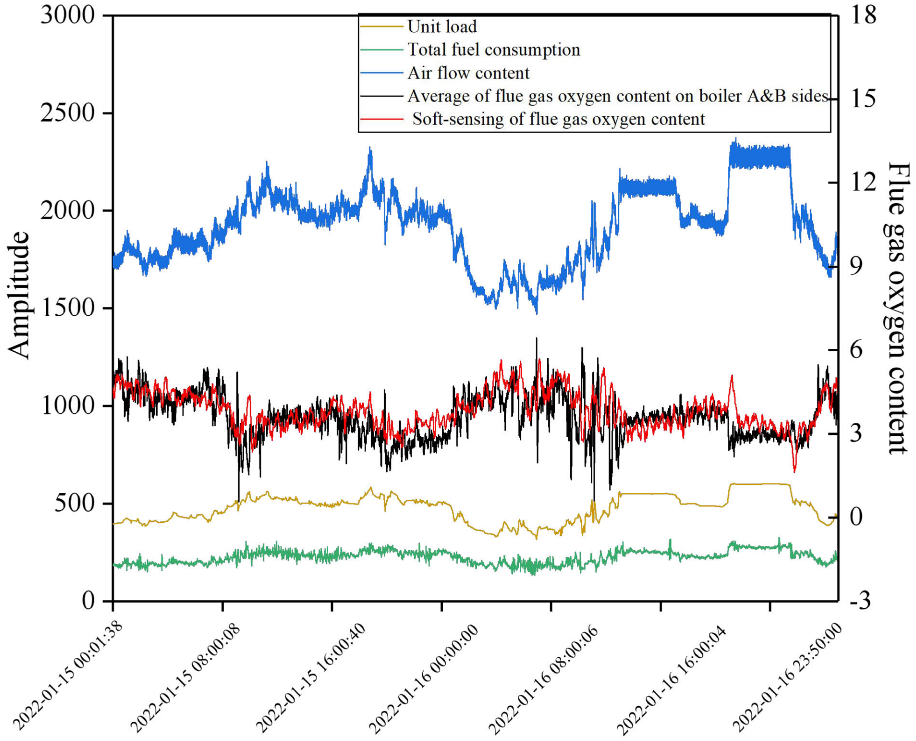

FIGURE 3. Trend of soft-sensing values of flue gas oxygen content.

In Formula (19) and Formula (20), Q represents net calorific value of coal, measured in kJ/kg.

Finally, leveraging the real-time net calorific value of coal obtained through Formula (14) in the DCS system, combined with the results of Formula (19) and Formula (20), we calculate the theoretical air quantity Q_r and the theoretical flue gas volume Q_{FT} required for the complete combustion of 1 kg of entering coal and subsequently obtain the oxygen content in the boiler flue gas through Formula (15).

To validate the soft-sensing of flue gas oxygen content, we utilized actual operational data for verification, Figure 3 shows three aspects were examined: First, the calculated results from the established soft-sensing model were highly accurate and could replace physical sensors. Second, the soft-sensing model demonstrated remarkable accuracy improvement in measuring oxygen content, especially at low loads, proving more reliable than the measurements from physical sensors. Third, the soft-sensing results provided effective redundant information for physical sensors in cases of hardware failures or malfunctions, thereby enhancing measurement reliability across various operational conditions.

C. SUB-MODEL FOR CARBON CONTENT IN FLY ASH

Carbon content in fly ash is influenced by various factors, making it challenging to model through mechanism analysis. As a result, factor analysis is employed in this paper to select key indicators or factors from numerous variables, forming the basis of the fly ash carbon content model. Subsequently, the model is refined using experimental data.

Initially, factor analysis is applied to obtain the maximum variance of the factor component matrix through orthogonal rotation, as per Formula (15), establishing a factor model for carbon content in fly ash C_{fh}:

$$C_{fh} = -0.25F_1 + 0.87F_2 + 0.24F_3 - 0.082F_4 \tag{21}$$

In Formula (21), F_1 represents the factor of hot air temperature, F_2 stands for the factor of burnout, F_3 corresponds to the factor of primary air pressure, and F_4 represents the coal quality factor. Therefore, the main influencing factors on fly ash carbon content can be decomposed into: the hot air temperature factor (primary air temperature T_1 and secondary air temperature T_2), the burnout factor (excess air coefficient and coal fineness R_{90}), the primary air pressure factor, and the coal quality factor. In this paper, T_{out} represents the outlet temperature of pulverizers (T_1), t_{oa} signifies the average temperature at the outlet of air preheaters A and B (T_2), the excess air coefficient can be obtained from the calculation formula of the oxygen content $\Phi(O_2)$ in the flue gas of the boiler tail (soft-sensing value) and the oxygen content $\Phi'(O_2)$ at the inlet of the air preheater, and R_{90} represents the particle size distribution of coal particles, which can be modeled through soft-sensing using primary air pressure P_A and pulverizers separator frequency F_E.

Subsequently, the coal quality factor is primarily determined by the contents of the coal's total moisture M, inherent ash A, volatile matter V, and total sulfur S. The influence weights of these parameters on fly ash carbon content, as well as the variations in fly ash carbon content due to changes in coal quality under specific conditions, can be calculated using boiler design residence time [27]. This vital data requires online data to ensure its real-time availability. For instance, A can be derived through a linear relationship function corresponding to the net calorific value of coal calculated earlier (Q); S can be derived through a linear relationship function corresponding to the boiler outlet SO_2 measurement value; M can be obtained by considering energy balance principles and soft-sensing modeling based on variables such as pulverizers air flow rate V_m, pulverizers load B_m, pulverizers inlet temperature T_{in}, pulverizers outlet temperature T_{out}, and coal fineness R_{90} [28]; V can be input manually based on the coal type or proximate analysis.

Finally, for refining the factor model, individual variable experiments were conducted, collecting actual test data (including offline assay data for fly ash carbon content). From these data, curves depicting the numerical variations of fly ash carbon content due to individual variable influences were derived. These curves were then corrected using the least squares method. Through multiple rounds of experimental data correction, a relatively accurate factor model for fly ash carbon content was established.

D. IMPROVED AP-BASED LS-SVM MULTI-MODEL BOILER EFFICIENCY MODELING

After obtaining the soft measurement values of boiler combustion characteristic parameters, such as oxygen content of flue gas and carbon content in fly ash, this paper incorporates these values into the sample library, thereby significantly enhancing the real-time and accuracy of the sample library. Through this improvement, we utilize the LS-SVM algorithm to establish sub-models for oxygen content of flue gas, carbon content in fly ash, exhaust gas temperature, and average furnace temperature. In the multi-modeling phase, a weighted sum method is employed to combine the LS-SVM models and the functional relationship models, enabling the prediction of boiler efficiency. The following steps outline the process:

Preprocessing of training sample data: This involves eliminating outliers, identifying steady state operating conditions, and performing normalization. Training of sample AP clustering: For the normalized training sample data, the parameters λ and p are initialized for the AP clustering. Here, the range of λ is set between 0.5 and 0.9, and p is set as $p = \text{median}(\boldsymbol{S})$. The clustering result corresponding to $\max\{S_{il\_av}\}$ is selected, and the parameter p is adjusted based on the S_{il_av} index, with the down step defined as shown in Formula (22).

$$p_{step} = \frac{0.01p_m}{0.1\sqrt{k+50}} \tag{22}$$

In Formula (22), k denotes the number of cluster centers.

Training of sub-models: Based on the clustering results, the LS-SVM regression algorithm is applied to train the sub-models. The RBF kernel function $K(x, x_i) = \exp\{\text{-}||x\text{-}x_i||^2/\sigma^2\}$ is used, and the regularization parameter γ and the kernel parameter σ of the LS-SVM are selected through cross-validation. The predictive ability index of the model is expressed as shown in Formula (23).

TABLE 1. Details of data dimensions and sample numbers for LS-SVM sub-models.

Sub-Model	Data Dimension	Sample Number (Training Set)	Sample Number (Testing Set)
SVM_A	$m \times n_A = 300 \times 8$	$k_A = 210$	$k_A' = 90$
SVM_B	$m \times n_B = 300 \times 22$	$k_B = 210$	$k_B' = 90$
SVM_C	$m \times n_C = 300 \times 4$	$k_C = 210$	$k_C' = 90$
SVM_D	$m \times n_D = 300 \times 3$	$k_D = 210$	$k_D' = 90$

$$press = \sum_i (y_i - y_{-i})^2 \tag{23}$$

In Formula (23), y_i represents the actual value of sample i, and y_{-i} denotes the model's predicted regression value for y_i when the i-th sample is excluded from the training data, a smaller "*press*" value indicates higher regression accuracy, signifying optimal parameter selection.

Weighted sum prediction of multi-models: For a new test sample x', its similarity to each sub-cluster center is calculated. It is then assigned to the class with the highest similarity and predicted using the corresponding sub-model. By summing the outputs of multiple sub-models with weights, the prediction of boiler efficiency can be obtained, as shown in Formula (24).

$$y' = \sum_{i=1}^{n} (W_i y_i') \tag{24}$$

In Formula(24), y$'$ signifies the predicted boile efficiency, n represents the number of sub-models, W_i is the weight of the i-th sub-model, and $\mathrm{y_i}'$ is the prediction of the i-th sub-model.

The weights of the weighted sum are influenced by the prediction errors of the sub-models, with smaller errors resulting in larger weights. To prevent overly large weights due to excessively small prediction errors, we introduce the parameter α, resulting in the weight calculation formula as presented in Formula (25):

$$Wi = \frac{1}{\varepsilon i + \alpha} \tag{25}$$

In Formula (25), W_i represents the weight of the i-th sub-model, ε_i denotes the prediction error of the i-th sub-model, and α is a parameter controlling the upper limit of weights, typically set to 0.1.

IV. NUMERICAL EXPERIMENTS

A. DATA COLLECTION AND PREPROCESSING

In accordance with the model input data as described in Section III-A, data collection was conducted on Boiler No. 2. After identifying steady state operating conditions using system discrimination, data was collected at 10-minute intervals. The real-time calculations for net calorific value of coal (Q), total moisture (M), inherent ash (A), total sulfur (S), and volatile matter (V) were derived from DCS data.

TABLE 2. Comparison results between predicted oxygen content and actual measurements from the oxygen grid method sensors.

Number / Load	A1(%)	A2(%)	A3(%)	B1(%)	B2(%)	B3(%)	AVER(%)	$\Phi(O_2)$(%)
40%P	5.9	6.0	5.8	5.6	5.6	5.5	5.73	5.84
60%P	5.4	5.5	5.4	5.2	5.1	5.1	5.28	5.22
80%P	4.7	4.8	4.7	4.5	4.4	4.5	4.60	4.81
100%P	3.4	3.4	3.3	3.3	3.1	3.2	3.28	3.43

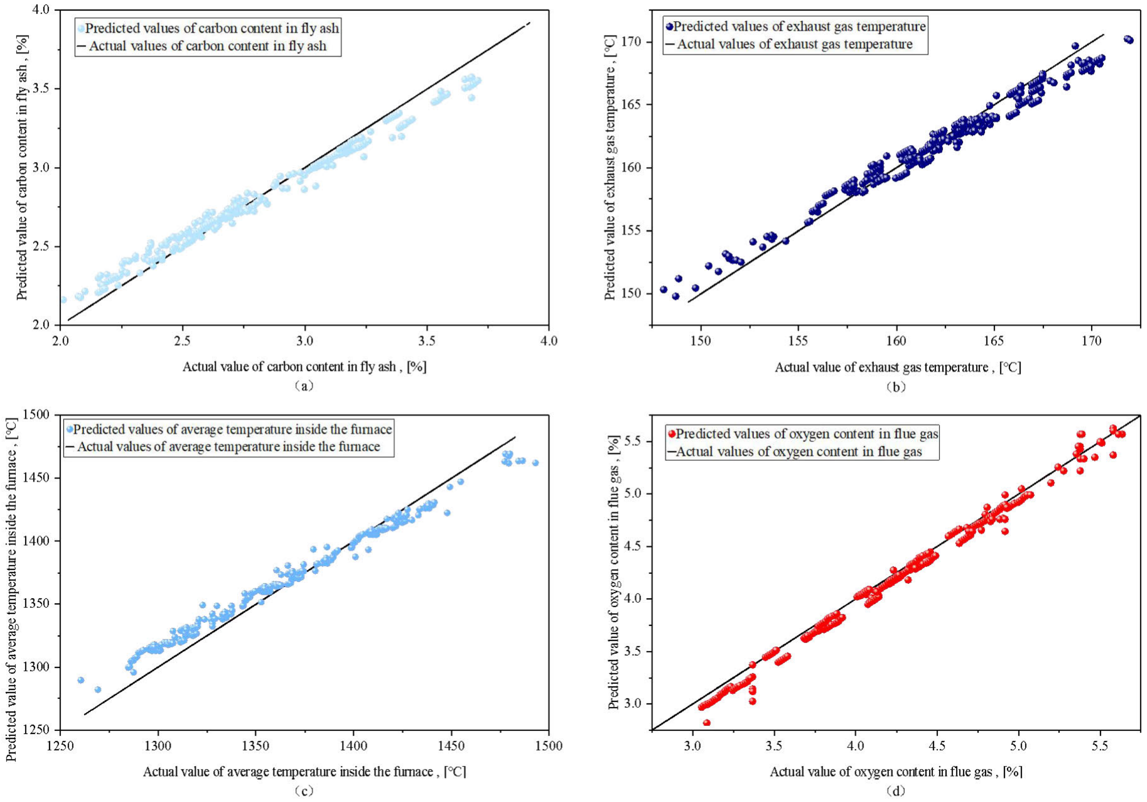

FIGURE 4. Validation results of carbon content in fly ash (C_{fh}), exhaust gas temperature (t_{py}), average temperature inside the furnace (t_{iB}) and oxygen content in flue gas ($\Phi\,(O_2)$.

These real-time measurements were supplemented by manual input from operators based on proximate analysis. If manual inputs were provided, they would override the soft sensing values, though if no new inputs were provided within four hour, the system would revert to using the soft sensing values. V can be input manually based on the coal type or proximate analysis.

Furthermore, to ensure data quality, the 3σ principle was employed to eliminate outliers. Data normalization was performed using Formula (26), which limited input and output values within the range of [−1, 1].

$$x' = [x - \frac{\max(x) + \min(x)}{2}]/[\frac{\max(x) - \min(x)}{2}] \quad (26)$$

In Formula (26), x and x' represent the original and normalized values, respectively.

B. TESTING AND ANALYSIS OF SUB-MODELS

During the training of sub-models, the data was divided into training and testing sets at a ratio of 7:3. Using cross-validation, the parameters γ and σ of the sub-models were determined. The improved LS-SVM model was then constructed using these optimized parameters. Subsequently, predictive experiments were conducted for various sub-models, including oxygen content of flue gas (SVM_A), carbon content in fly ash (SVM_B), exhaust gas temperature (SVM_C), and average temperature in the furnace chamber (SVM_D). The precise details of data dimensions and sample numbers used in these experiments are outlined in Table 1.

In the table, m represents the total number of data points, n_A, n_B, n_C, n_D represent the respective data dimensions for each sub-model, and k_A, k_B, k_C, k_D denote the sample

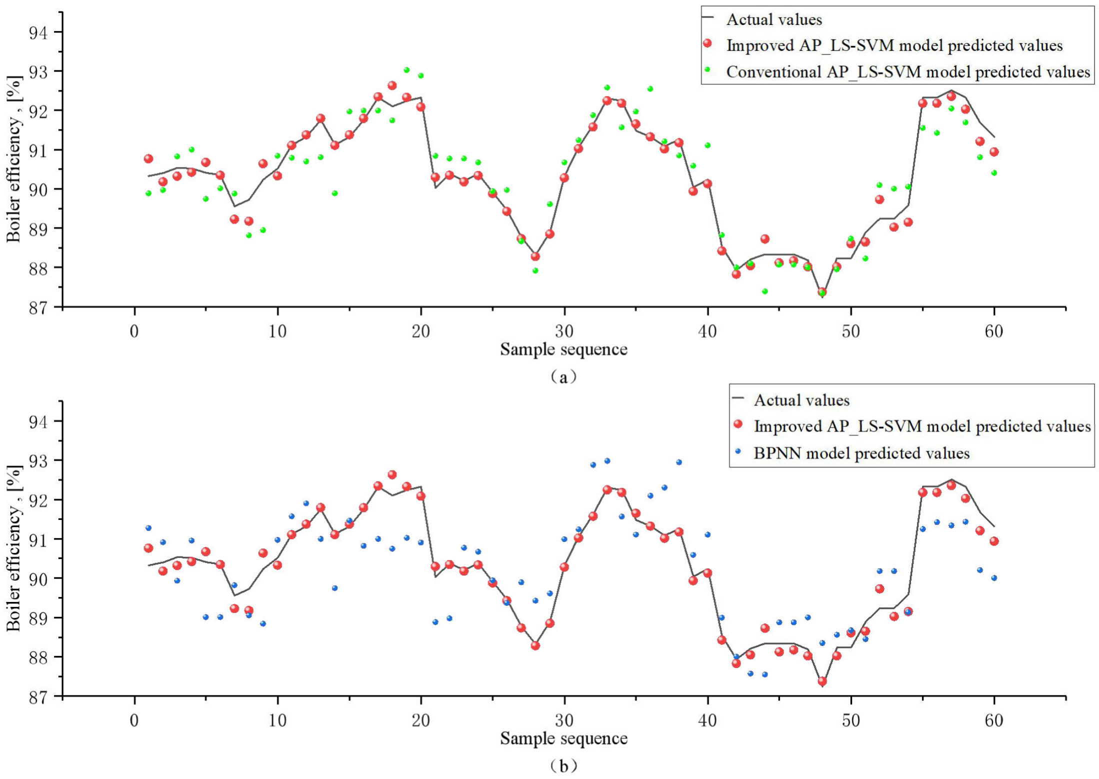

FIGURE 5. Comparison of predictive efficiency among the improved AP_LS-SVM, unimproved AP_LS-SVM, and BPNN models.

numbers in the training set. Similarly, k_A', k_B', k_C', k_D' represent the sample numbers in the testing set.

Regarding to SVM_A, in order to accurately capture the oxygen content across the entire flue, we employed an oxygen grid method. In this method, three high-precision zirconia oxygen sensors were strategically positioned on each side of the flue, denoted as A1, A2, and A3 for the A side, and B1, B2, and B3 for the B side. By collecting data from these sensors at different positions and varying insertion depths, followed by computing the average of all recorded values, we obtained the precise actual oxygen content within the flue gas.

Subsequently, we conducted predictive experiments using the SVM_A model under different load conditions. The predicted oxygen content values were then compared with the corresponding actual measurements, and the outcomes of this comparative analysis are presented in Table 2.

Regarding to SVM_B, we conducted model training using sample data collected at different loads, specifically focusing on the data points surrounding offline assay sampling moments. Regarding to SVM_C and SVM_D, data for model training were collected during Boiler No. 2 efficiency tests, including parameters such as flue gas temperature (t_{py}), combustion flame center temperature (t_m), and furnace outlet temperature (t_{ob}).

Figure 4 illustrates the comparison between the actual measured t_{py} and the predicted t_{py} by the improved LS-SVM, and the actual $\Phi(O_2)$, C_{fh}, t_{IB} and predicted $\Phi(O_2)$, C_{fh}, t_{IB} by the improved LS-SVM on the test data set, respectively. It is easy to see that almost all of the data have been fallen in a diagonal distribution along with the perfect line, where predicted values are equal to actual values. It means that all $\Phi(O_2)$, C_{fh}, t_{py}, and t_{IB} could be predicted with good accuracy by the improved LS-SVM for all testing data sets.

In addition, Table 3 summarizes the precision calculation outcomes. The $\Phi(O_2)$, C_{fh}, t_{py}, and t_{IB} models, established based on the improved LS-SVM, demonstrate a notably high level of predictive capability.

TABLE 3. Accuracy of LS-SVM models.

Parameters	$\Phi(O_2)$/%	C_{fh}/%	t_{py}/°C	t_{IB}/°C
MAXE	0.2100	0.1900	1.9200	17.8400
RMSE	0.1530	0.1303	0.6005	9.1713

C. TESTING AND ANALYSIS OF BOILER EFFICIENCY MODEL

In this paper an improved AP-based LS-SVM multi-model approach was employed for boiler efficiency modeling experiments. During the Boiler No.2 efficiency test, sample data was collected and randomly divided into training set (45 groups) and testing set (15 groups).

Firstly, AP clustering was performed on the training samples. By adjusting parameter p, the k index (the number of cluster centers) reached its maximum value when samples were divided into four clusters. Subsequently, corresponding LS-SVM sub-models were established. Cross-validation was employed to determine parameters γ and σ for the sub-models, minimizing the "*press*" value within each sub-model. The optimal parameter combinations are presented in Table 4.

TABLE 4. Parameters of LS-SVM models.

Parameters	SVM_A	SVM_B	SVM_C	SVM_D
γ	70	60	50	90
σ	0.45	0.4	0.6	0.6

Next, the Euclidean distances between test samples and cluster centers of each class were calculated. Test samples were classified accordingly, and corresponding sub-models were utilized for prediction. Subsequently, the efficiency was predicted using the boiler efficiency function relationship and weighted summation.

Lastly, the efficiency calculated from the efficiency test was defined as the actual boiler efficiency value. To assess the performance of the aforementioned models in predicting boiler efficiency, a comparison was made between these predicted values, the unimproved AP_LS-SVM model, and the BP neural network model. Figure 5 shows, the predictive performance of the three models, and performance parameters are presented in Table 5.

TABLE 5. Comparison between improved the AP_LS-SVM, unimproved AP_LS-SVM, and BPNN models.

Model	Training time, (s)	MAXE	RMSE
Improved AP_LS-SVM	0.4195	0.5505	0.2262
Conventional AP_LS-SVM	0.6372	1.2808	0.5981
BPNN	2.2134	1.6883	0.9048

Clearly, the improved AP-based LS-SVM model exhibits a distinct advantage in predicting boiler efficiency. This model not only offers a shorter training time but also demonstrates higher levels of prediction accuracy and generalization ability.

V. CONCLUSION

In conclusion, this paper introduced an improved AP-based LS-SVM method into the modeling of boiler combustion efficiency in coal-fired power plants. This method effectively divided training samples and established four sub-models based on the improved LS-SVM approach. The approach exhibited advantages in terms of training time, prediction accuracy, and generalization capability, holding great promise for widespread application in the field of thermal power generation.

ACKNOWLEDGMENT

The authors declare no conflict of interest.

REFERENCES

[1] M. Mollo, A. Kolesnikov, and S. Makgato, "Simultaneous reduction of NOx emission and SOx emission aided by improved efficiency of a once-through benson type coal boiler," *Energy*, vol. 248, Jun. 2022, Art. no. 123551.

[2] O. Mohamed, A. Khalil, and J. Wang, "Modeling and control of supercritical and ultra-supercritical power plants: A review," *Energies*, vol. 13, no. 11, p. 2935, Jun. 2020.

[3] T. Chen, Y.-J. Zhang, M.-R. Liao, and W.-Z. Wang, "Coupled modeling of combustion and hydrodynamics for a coal-fired supercritical boiler," *Fuel*, vol. 240, pp. 49–56, Mar. 2019.

[4] K. Sun and Y. Li, "Modeling method of boiler combustion system based on empirical mode decomposition," in *Proc. China Autom. Congr. (CAC)*, Fujian, China, Nov. 2022, pp. 5198–5203.

[5] A. A. M. Rahat, C. Wang, R. M. Everson, and J. E. Fieldsend, "Data-driven multi-objective optimisation of coal-fired boiler combustion systems," *Appl. Energy*, vol. 229, pp. 446–458, Nov. 2018.

[6] W. Xu, Y. Huang, S. Song, B. Chen, and X. Qi, "A novel online combustion optimization method for boiler combining dynamic modeling, multi-objective optimization and improved case-based reasoning," *Fuel*, vol. 337, Apr. 2023, Art. no. 126854.

[7] C. Wang, Y. Liu, S. Zheng, and A. Jiang, "Optimizing combustion of coal fired boilers for reducing NOx emission using Gaussian process," *Energy*, vol. 153, pp. 149–158, Jun. 2018.

[8] S. Li, Z. Wu, and Y. Ge, "Mixed modeling and optimization of NOx emission and efficiency for power plant boiler," *J. Therm. Sci. Tech.*, pp. 30–35, 2007.

[9] J. Cai, X. Ma, and Q. Li, "On-line monitoring the performance of coal-fired power unit: A method based on support vector machine," *Appl. Thermal Eng.*, vol. 29, nos. 11–12, pp. 2308–2319, Aug. 2009.

[10] X. Xu, Q. Chen, M. Ren, L. Cheng, and J. Xie, "Combustion optimization for coal fired power plant boilers based on improved distributed ELM and distributed PSO," *Energies*, vol. 12, no. 6, p. 1036, Mar. 2019.

[11] G.-Q. Li, X.-B. Qi, K. C. C. Chan, and B. Chen, "Deep bidirectional learning machine for predicting NOx emissions and boiler efficiency from a coal-fired boiler," *Energy Fuels*, vol. 31, no. 10, pp. 11471–11480, Oct. 2017.

[12] S. B. Savargave and M. J. Lengare, "Modeling and optimizing boiler design using neural network and firefly algorithm," *J. Intell. Syst.*, vol. 27, no. 3, pp. 393–412, Jul. 2018.

[13] X. Hu, P. Niu, J. Wang, and X. Zhang, "Multi-objective prediction of coal-fired boiler with a deep hybrid neural networks," *Atmos. Pollut. Res.*, vol. 11, no. 7, pp. 1084–1090, Jul. 2020.

[14] X. Wu, Z. Tang, and S. Cao, "A hybrid least square support vector machine for boiler efficiency prediction," in *Proc. IEEE 3rd Inf. Technol. Mechatronics Eng. Conf. (ITOEC)*, Chongqing, China, Oct. 2017, pp. 1202–1205.

[15] J.-J. Li, F. Alzami, Y.-J. Gong, and Z. Yu, "A multi-label learning method using affinity propagation and support vector machine," *IEEE Access*, vol. 5, pp. 2955–2966, 2017.

[16] Q. Gao, Y. Wang, X. Cheng, J. Yu, X. Chen, and T. Jing, "Identification of vulnerable lines in smart grid systems based on affinity propagation clustering," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 5163–5171, Jun. 2019.

[17] L. Yan-jun and M. Qian, "AP-LSSVM modeling for water quality prediction," in *Proc. 31st Chin. Control Conf.*, Hefei, China, Jul. 2012, pp. 6928–6932.

[18] B. Liu, F. Zhang, X. Li, H. Wu, and L. Xu, "Power load identification based on long-and-short-term memory network and affinity propagation clustering algorithm," *Energy Rep.*, vol. 8, pp. 1137–1144, Jul. 2022.

[19] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Process. Lett.*, vol. 9, no. 1, pp. 293–300, 1999.

[20] L. Li, H. Su, and J. Chu, "Modeling of isomerization of C_8 aromatics by online least squares support vector machine," *Chin. J. Chem. Eng.*, vol. 17, no. 3, pp. 437–444, Jun. 2009.

[21] J. Xu, "Comparison and simplification of boiler efficiency calculation model," *Energy Res. Utili.*, vol. 1, pp. 16–18, 2021.

[22] C. Mi, J. Guo, X. Li, G. Wei, and G. An, "Flue gas determination based soft-sensing model for coal quality monitoring for utility boilers," *Therm. Power Gen.*, vol. 44, no. 7, pp. 62–65, 2015.

[23] N. Qin and R. J. Li, "Online simplified model and experimental comparison of CFB boiler thermal efficiency," *Appl. Thermal Eng.*, vol. 171, May 2020, Art. no. 115021.

[24] H. Li and P. Yang, "Research and application of real-time measurement for lower calorific value of burning coal in coal-fired power plants," *Elec. Power Const.*, vol. 34, no. 10, pp. 60–64, 2013.

[25] Z. Zhao, D. Zeng, L. Tian, and J. Liu, "Research on soft-sensing of oxygen content based on datafusion," in *Proc. CSEE.*, 2005, vol. 25, no. 7, pp. 7–12.

[26] J. Cai, X. Q. Ma, and Y. Liao, "Study on boiler's operation performance and optimization of oxygen content in flue gas," *Therm. Power Gen.*, vol. 7, pp. 28–30, 2006.

[27] J. Fu and J. Feng, "Resident time of combustion products and its effect on carbon content in fly ash in pulverized coal fired boilers," *Boiler Tech.*, vol. 36, no. 6, pp. 21–24, 2005.

[28] Z. Zhao, X. Chen, and Y. Wang, "Research on soft-sensing of coal moisture," *Elec. Power Sci. Eng.*, vol. 25, no. 11, pp. 67–69, 2009.

MING YAN was born in Anshun, China, in 1986. He received the master's degree in control engineering from Zhejiang University, China. He is currently a Senior Engineer. He has been extensively involved in the fields of data mining and intelligent control technology in thermal power plants.

LIANG WANG was born in Sichuan, China, in 1993. He received the master's degree in engineering from Southwest University, China, in 2020. His research interest includes electrical and automation.

MEILING ZHANG was born in Hebei, China, in 1988. She is currently pursuing the master's degree in engineering management with Chongqing University, China. Previously, she was extensively involved in chemical analysis and characteristic studies of coal quality and flue gas components in thermal power plants.

PAN SHI was born in Dezhou, China, in 1986. He received the master's degree in energy and power engineering from Chongqing University, China. He is currently a Senior Engineer. He has been extensively involved in the fields of in boiler combustion and intelligent control technology in thermal power plants.