Load Prediction Model of Gas Boiler Generator Set Based on CNN-LSTM*

Yan Xu[1,2,3], Min Wu[1,2,3], Jie Hu[1,2,3,*], Fu-Sheng Peng[1,2,3], Wen Zhang[1,2,3], Wen-Shuo Song[1,2,3], and Hui-Hang Li[1,2,3]

[1] School of Automation, China University of Geosciences Wuhan 430074, P. R. China
[2] Hubei Key Laboratory of Advanced Control and Intelligent Automation for Complex Systems Wuhan 430074, P. R. China
[3] Engineering Research Center of Intelligent Technology for Geo-Exploration, Ministry of Education Wuhan 430074, P. R. China

Abstract. With the proposal of carbon peak and carbon neutrality targets and the construction of a new power system, improving energy efficiency and reducing carbon emissions have become the key to the sustainable development of industrial boiler generator sets. However, the current load scheduling of gas boiler power plants is mainly based on manual experience, which limits achieving more efficient and stable operation. Accurate short-term load forecasting can help dispatchers to make reasonable production schedules. Therefore, this paper proposes a hybrid load forecasting model for generator sets based on convolutional neural networks combined with long short-term memory networks. First, the Isolation Forest algorithm eliminates anomalies in the historical data set. Then, variables related to unit load are selected by combining mechanistic analysis with the Spearman algorithm. Considering the multiple operating conditions and high complexity of unit loads, the CNN-LSTM network is used to fully extract the spatial and temporal characteristics to build the load prediction model. Finally, experiments are conducted using actual production data, showing that the proposed method is effectiveness.

Keywords: Load prediction · Long- and short-term memory networks · Convolutional networks · Hybrid models.

1 Introduction

As people's living standards continue to rise, so does energy consumption. However, boiler generator sets in power plants generally suffer from low operating efficiency and frequent load fluctuations. Research on load prediction for generator sets is essential for the rational scheduling of power plant operations and is fundamental to ensuring the safe and efficient operation of the units. With the development of artificial intelligence and big data technologies, more and more researchers are using actual industrial processes, analysing large amounts

* Corresponding author: Jie Hu (hujie@cug.edu.cn).

of historical data combined with process mechanisms, and employing intelligent methods to study load forecasting for boiler generator sets and explore their patterns.

The choice of load prediction model method is crucial for the model's accuracy. The gas boiler combustion and power generation process is a complex industrial process, and some scholars have approached it from a mechanistic perspective. In [1], by analysing the principles of the thermodynamic process of turbines, transfer functions were used to characterise the properties of high-pressure cylinders, reheaters and low-pressure crossover tubes. This study introduced the natural over-tuning coefficient of high-pressure cylinder performance and improved the data model of reheat condensing steam turbines. Wang et al. considered the influence of thermal storage effects on the turbine power generation process and established a mathematical model for drum boiler units [2]. However, these models require the determination of numerous transfer function coefficients and do not consider the coupling effects between variables in the combustion and power generation process or the complexity under different operating conditions. Therefore, these methods have little general applicability.

Based on the large amounts of data stored from actual industrial processes, data-driven modelling methods are increasingly used in unit load forecasting research. These methods are based on the data itself and do not require an in-depth understanding of the physical and chemical processes of the system. They can learn patterns, extract features from historical data, and handle extensive, complex, high-dimensional data for future predictions. Kong et al. improved ARIMA using wavelet transformation to build a short-term load forecasting model for thermal power plants [3]. Zhao et al. proposed using the SVM algorithm for unit load forecasting, which performs load forecasting under various operating conditions but shows poorer forecasting results when data fluctuations are significant [4]. Dong et al. investigated short-term load forecasting using LSTM and recurrent neural networks, respectively, and achieved good results during load fluctuation processes [5].

Single models have certain limitations in unit load forecasting. When faced with complex non-linear relationships and temporal features, they may need to more effectively capture the intricate characteristics of multi-dimensional data, thus compromising prediction accuracy. Combined models have become the mainstream trend in load forecasting [6-7]. For example, in [8], a combined CNN-GRU model was used to predict the electricity load in a specific region, achieving a high overall prediction accuracy and demonstrating the advantages of combined models.

These methods have achieved good results in load forecasting. However, in natural industrial environments, sensors are subjected to high-frequency use, high temperature and high noise, which can lead to anomalies in the collected data. These anomalies reduce the quality of the data and thus affect the accuracy of the constructed models.

The model uses the CNN for its ability to extract spatial features from data [9-10], and the LSTM model to capture the temporal relationships of these fea-

tures over time. This combination gives the hybrid model more powerful feature representation and prediction capabilities. First, the isolation forest algorithm is used to detect and eliminate anomalies in the actual production data. Then, correlation analysis methods are used to determine the inputs for the CNN-LSTM model based on an analysis of the process mechanism. Experimental results show that the proposed method is effective and matches well with the generator sets' actual production and operation processes.

2 Process Analysis and Prediction Strategy Design

This section describes the gas boiler combustion and power generation process, analyses the factors affecting the unit load, and finally designs a load forecasting scheme.

2.1 Process Description

Fig. 1 is a system diagram of a 150 MW ultra-high-temperature gas-fired subcritical boiler generator set in a power plant. The whole system can be divided into four main conversion modules: fuel-steam module, steam-pressure module, pressure-power module, and flue-gas-air module. It consists mainly of the blower, air preheater, steam drum, superheater, reheater, economizer, induced draft fan, furnace, stack, and unit.

For the gas boiler combustion and power generation process, it is essential first to understand the mechanism of high-temperature flue gas generation and then further analyse the circulation process of high-temperature flue gas. The generation of high-temperature flue gas requires fuel and oxygen, which are supplied by fans. These fans typically include primary and secondary air fans. In order to simplify the analysis, primary and secondary air will be discussed. The primary role of the primary air is to carry the fuel into the furnace and to provide the oxygen required for fuel combustion. The primary role of the secondary air is to provide additional oxygen to the furnace to ensure that the fuel burns entirely within the furnace, producing a large amount of high-temperature flue gas. This high-temperature flue gas passes through the superheater and reheater to heat the steam that drives the turbine, allowing the generator to produce electricity and meet the load requirements of the unit.

During this process, some high-temperature flue gas is discharged to the boiler exhaust with the steam. To avoid wasting this energy, the flue gas from the boiler exhaust is usually routed to the air preheater. The air preheater uses this high-temperature flue gas to heat the secondary air, and part of the primary air is injected into the furnace, thereby recovering the high-temperature flue gas and preheating the air.

2.2 Factors Affecting Unit Load

Load is a critical state parameter in the gas boiler power generation process. Accurate prediction of load variations is essential to ensure stable unit operation

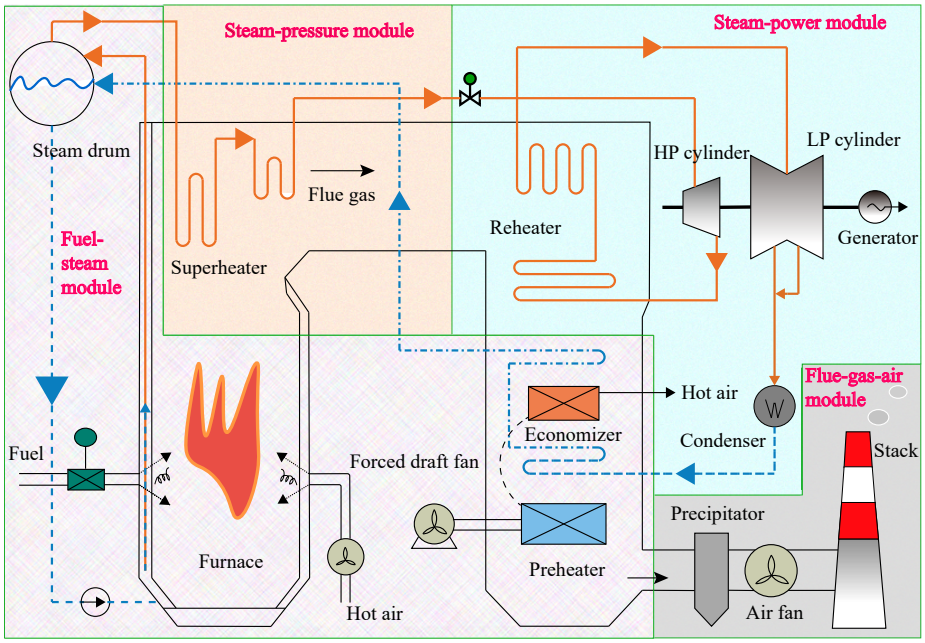

Fig. 1. Process flow diagram

while quickly meeting dispatch targets and improving energy efficiency. However, the power generation process of boiler units is complex and characterised by strong coupling, non-linearity, and multiple parameters. As a result of this complexity, many factors affect the load of gas boiler generator sets, the most important of which are environmental conditions and operating parameters.

When analysing the mechanisms and power generation characteristics, fuel quality and ambient temperature directly affect the unit load. Fuel quality determines the efficiency of combustion and the amount of power generated. High-quality fuel can provide a more stable energy output, thereby maintaining load stability. Ambient temperature affects the heat exchange efficiency of the gas boiler and the cooling system's performance. At higher temperatures, the heat generated inside the boiler is more easily lost to the external environment by radiation and convection, and the temperature of the cooling water also increases, resulting in poorer cooling performance and reduced overall thermal efficiency. In addition, changes in ambient temperature affect the power demand, which indirectly affects the unit load.

Operating parameters such as steam pressure, temperature, and feedwater flow rate directly impact the unit's operating status. Steam pressure and temperature must be maintained within optimum ranges to ensure efficient and stable power generation. In addition, the feedwater flow rate must be precisely controlled to maintain a stable steam supply. The stability of these parameters is critical to maintaining a stable load. Using artificial intelligence and data mining

techniques, it is necessary to develop load forecasting models based on analysing their process mechanisms and characteristics. These models can ensure a rapid response to changes in these parameters, thereby maintaining stable operation.

2.3 Unit Load Forecasting Strategy Framework

Considering that the unit's power generation process is influenced by multiple parameters and the load frequently fluctuates, this paper proposes a load forecasting strategy based on CNN-LSTM. The predictions obtained from this method provide decision support for load scheduling and adjustment. The architecture of the load forecasting strategy is shown in Fig. 2.

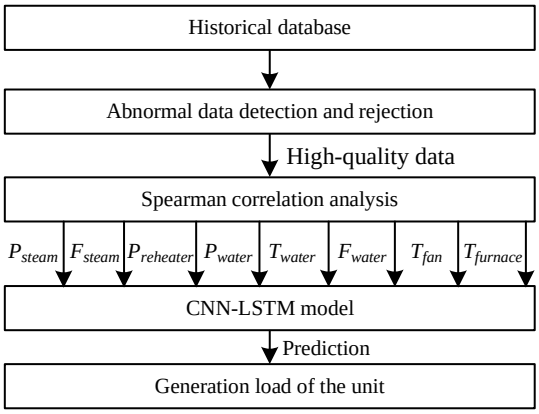

Fig. 2. Prediction strategy architecture

First, relevant parameter data is obtained from the historical database of the gas boiler unit's power generation process. Due to external environmental factors or sensor hardware issues, the collected data set may contain anomalies. Therefore, the isolation forest model detects and eliminates these anomalies.

Next, to address the issue of excess process parameters, the Spearman correlation analysis is used to calculate the correlation between process parameters and unit load. Variables most correlated with load changes are selected as inputs to the model, thereby reducing data redundancy and dimensionality. Considering that during actual unit operation, the load is influenced by factors such as fuel quality and production scheduling, resulting in frequent changes in load magnitude and frequency, a combined CNN-LSTM model is used for model. This approach can better extract temporal and spatial features from the data, improving the model's accuracy.

Finally, during the experimental testing phase, actual process data are fed into the trained model to predict the gas boiler unit's power generation load.

3 Load Prediction Model Based on CNN-LSTM

This section presents the load forecasting model based on CNN-LSTM. The isolation forest algorithm is used for data pre-processing to improve data quality. Next, the Spearman method performs correlation analysis on the parameters to determine the model inputs. Finally, the CNN-LSTM model is utilized to develop the load prediction model.

3.1 Abnormal Data Removal Based on Isolation Forest

Due to the high temperature, high noise, and poor equipment operating environment during the power generation process of gas boiler units, a small amount of abnormal data is likely to appear. These data cannot represent the normal operation process and will significantly impact subsequent modelling and prediction if not processed. Therefore, detecting and eliminating abnormal data through data pre-processing technology is necessary to obtain high-quality data.

The isolation forest algorithm is used to detect anomalous data in the power generation process. This algorithm is an unsupervised anomaly detection method that constructs multiple isolated binary trees in a randomised manner [11], allowing outliers to be split with only a few partitions, making it simple and efficient. Compared to the KNN method [12], the isolation forest is not affected by the failure of distance measurement in high-dimensional spaces, thus avoiding the curse of dimensionality. Similarly, the LOF method relies on density estimation [13], which can become inaccurate in high-dimensional data. The PCA-based method reduces the data dimensions to simplify the analysis [14], which can lead to the loss of important information. In contrast, the Isolation Forest does not require dimensionality reduction, providing greater adaptability and robustness.

In addition, the isolation forest does not need to assume any data distribution. In contrast, the GMM assumes that the data follows a Gaussian distribution [15], which may need to be revised for actual power generation process data. Therefore, the isolation forest algorithm is ideal for anomaly detection in complex data environments.

The structure of the isolation tree constructed by the isolation forest is shown in Fig. 3. Calculate the average path length $c(n)$ for all data points in the isolation forest, and then compute the anomaly score $s(x, n)$ using the formulas (1) to (3).

$$S(x, n) = 2^{\frac{-E[H(x)]}{c(n)}} \tag{1}$$

$$H(n) = \ln n + c \tag{2}$$

$$c(n) = \begin{cases} 2H(n-1) - \frac{2(n-1)}{n}, n > 2 \\ 1, \qquad\qquad\qquad\quad n = 2 \\ 0, \qquad\qquad\qquad\quad n < 2 \end{cases} \tag{3}$$

where $E[H(x)]$ is the expected value of $h(x)$, $h(n)$ is the harmonic function, c is the Euler constant, in $s(x, n)$, x is the number of data and n is the sample size, the S value range is [0,1], and the abnormal score of normal data is about 0.5.

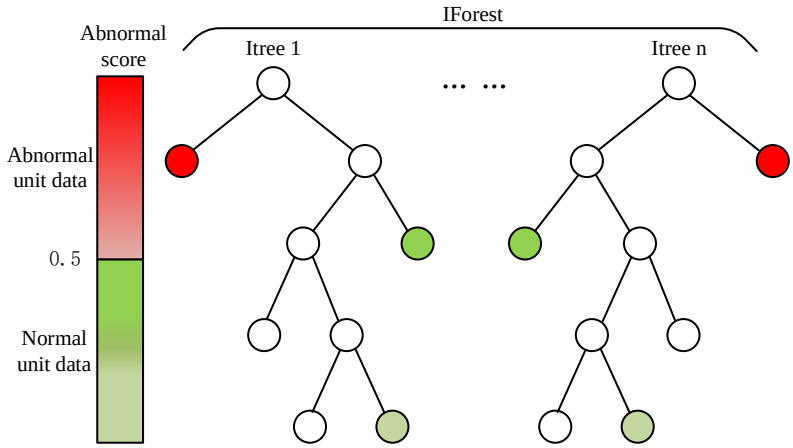

Fig. 3. Isolated forest structure

More than 40 parameters must be measured in the power generation process of gas boiler units, which are typical high-dimensional data, mainly including condition and operating parameters. The parameters that affect the unit load include: central steam pressure (P_{steam}), main steam flow (F_{steam}), reheater pressure ($P_{reheater}$), feed water temperature (T_{water}), feed water flow (F_{water}), furnace outlet flue gas temperature ($T_{furance}$), etc. Considering the parameters are detected simultaneously, the central steam pressure and reheater temperature are selected as representative for abnormal data detection and elimination, and a box plot is given.

3.2 Spearman Correlation Analysis

Based on the abnormal data detection and elimination method, a large amount of high-quality actual production data was obtained, including 47 process parameters such as central steam pressure, primary steam flow rate, main steam temperature, reheater pressure, reheater temperature, feed water pressure, feed water flow rate, feed water temperature, furnace vacuum, and so on.

Therefore, this paper selects the Spearman correlation analysis method to measure the correlation between process parameters and unit load. The absolute value of the calculated correlation coefficient is between 0 and 1.

The formula for calculating the Spearman correlation coefficient is as follows:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \tag{4}$$

where x and y represent the two columns of features for which the correlation coefficient is to be calculated, x_n and y_n represent the corresponding features in the nth sample, $\bar{x}$ and $\bar{y}$ represent the mean of the features.

3.3 Prediction Model Design

The power generation process of gas-fired boilers involves many parameters and complex data, and prediction accuracy needs to be improved. The neural network prediction model based on LSTM performs well in processing time series and has been applied in many fields.

Since LSTM has shortcomings in learning cross-features, it is not easy to effectively process high-dimensional data in unit load prediction. Considering that convolutional neural networks (CNN) can extract features efficiently, this paper adopts a CNN-LSTM combined model for unit load prediction to exploit both models' advantages fully. The structure of the CNN-LSTM prediction model is shown in Fig. 5.

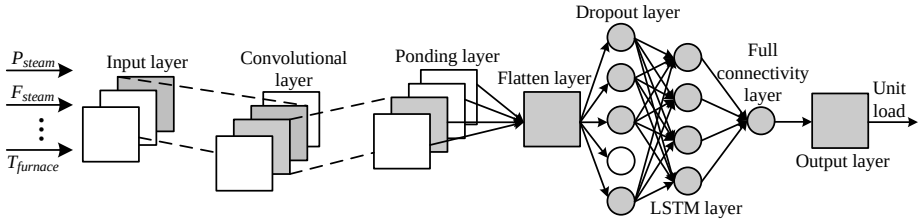

Fig. 4. Structure of CNN-LSTM forecasting model

4 Experimental Results and Discussion

This section introduces the evaluation indicators for evaluating the model's performance and conducts simulation experiments using actual gas boiler unit production data.

4.1 Evaluation Indices

In practical applications, the performance indicators of the calculation model must be used to evaluate the model's performance. This paper selects the root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and coefficient of determination (R^2) as the model performance evaluation indicators. The calculation formula is as follows.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{5}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i) \tag{6}$$

$$MAPE = \frac{100}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{\hat{y}_i}\right| \tag{7}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2} \tag{8}$$

where y_i is the true value of the unit load, N is the number of test sets, $\hat{y}_i$ is the predicted value, and $\bar{y}_i$ is the average value of the test sets.

RMSE, MAE and MAPE are the errors that represent the existence of predicted and actual values, and the smaller the value, the better the performance of the model built, while the closer the value of R^2 is to 1, the better the model's ability to explain the data.

4.2 Experiments Based on Actual Production Data

To verify the effectiveness of the proposed method, a total of 6500 data sets from the operation process of a power plant from February to October 2023 were used, of which 6000 sets were used as training sets, and 500 sets were used as test sets. The isolation forest algorithm was used to detect and remove abnormal data in the 6000 data sets, and a total of 501 data sets were removed. Fig. 6 shows two parameter box plots before and after the isolation forest algorithm, where the purple box is the data before removal, and the orange box is the data after removal. It can be seen from the box plot that the number of outliers has been significantly reduced, and the data quality has been effectively improved.

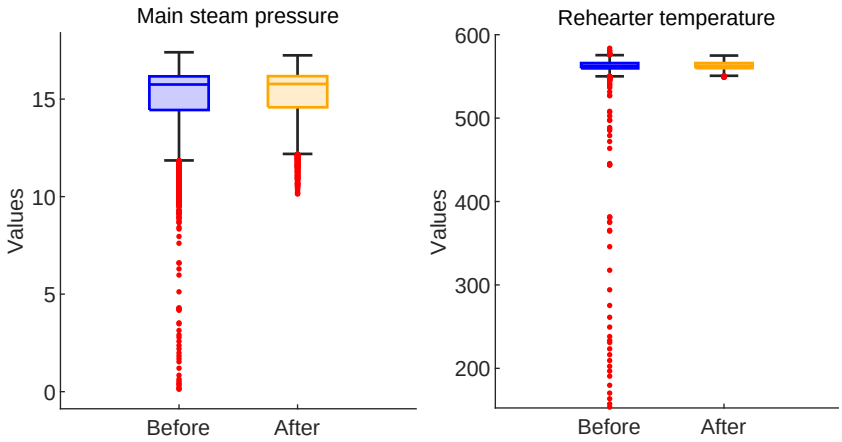

Fig. 5. Box plot comparing before and after removing outliers

Using the data processed by the isolation forest algorithm, the Spearman correlation coefficient between the process parameters and the unit load is calculated, as shown in Table 1. The variables with the highest correlation coefficient values are listed, among which the correlation coefficient values of central steam pressure, main steam flow, reheater pressure, feed water pressure and feed water flow with the unit load are all over 0.85, which proves that these five parameters have a strong correlation with the unit load, which is consistent with the process mechanism of gas boiler combustion power generation process. This paper selects variables with correlation coefficient values greater than 0.6 as the model's input, and the output is the unit load.

Table 1. Correlation coefficient value

	P_{steam}	F_{steam}	$P_{reheater}$	P_{water}
Unit load	0.8399	0.9157	0.9155	0.8655
	T_{water}	F_{water}	T_{fan}	$T_{furnace}$
Unit load	0.8292	0.9065	0.6341%	0.7476

From Table 2, it can be observed that the performance of the combined model surpasses that of the individual models. The proposed CNN-LSTM-based method achieves RMSE, MAE, MAPE, and $\mathbf{R^2}$ values of 5.1859, 4.7311, 3.2871%, and 0.8225, respectively. These metrics indicate that the prediction errors for the

Table 2. Comparative results for different models

Model	RMSE	MAE	MAPE	R^2
CNN	7.0496	6.0482	4.3188%	0.6720
LSTM	6.4884	5.2952	3.7414%	0.7222
CNN-LSTM	5.1859	4.7311	3.2871%	0.8225

load are relatively small, confirming the effectiveness of the proposed method. Thus, it is demonstrated that the model can largely meet the requirements for industrial on-site applications.

As can be seen from Fig. 7, the CNN-LSTM prediction curve fits better, and the prediction results are more reliable. In this paper, outliers are detected and removed from the training set using the isolated forest algorithm, and the data in the test set is not pre-processed because it is difficult to avoid the generation of easily abnormal data in the actual production process, and the results obtained are more in line with the actual industrial site.

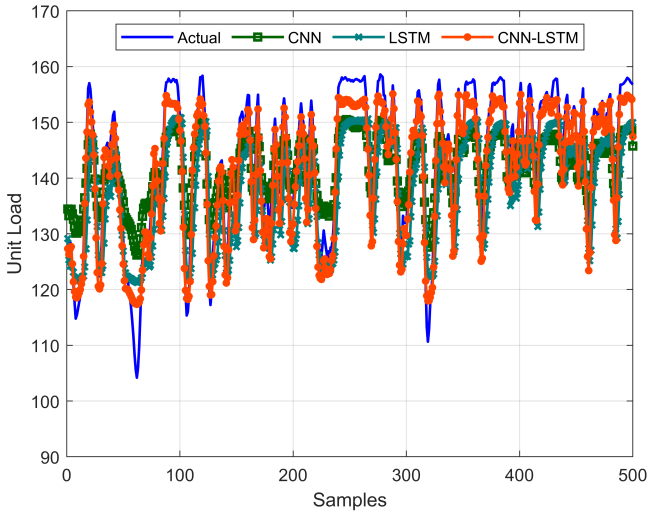

Fig. 6. Unit load prediction results based on different models

5 Conclusion

The power generation process of gas boiler units has the characteristics of solid coupling, multi-parameter and non-linearity. This paper develops a unit load prediction model based on CNN-LSTM, which cannot only describe the unit power generation process's characteristics but also solve the problem of modelling the unit power generation process. The experimental results show this model has better prediction accuracy than the individual LSTM and CNN models. The proposed strategy meets the site's needs and lays the foundation for the optimisation control of the unit power generation process. For future explorations, the use of hybrid modelling approaches to modelling can be investigated to improve the accuracy and stability of the model.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62303431, in part by the Hubei Provincial Natural Science Foundation of China under Grant 2024AFB589, in part by the 111 Project under Grant No. B17040, in part by the "CUG Scholar" Scientific Research Funds at China University of Geosciences (Wuhan) under Grant No.2021009, and in part by the Fundamental Research Funds for the Central Universities, China University of Geosciences under Grant CUG2106210.

References

1. P. Nikhil, V. Ashu, B. Terlochan Singh, "Automatic generation control of thermal power system under varying steam turbine dynamic model parameters based on generation schedules of the plants", *The Journal of Engineering*, vol. 2016, no. 8, pp. 302–314, 2016.
2. M. Plis, H. Rusinowski, "A mathematical model of an existing gas-steam combined heat and power plant for thermal diagnostic systems", *Energy*, vol. 156, pp. 606–619, 2018.
3. F. Kong and G. Song, "Middle-long power load forecasting based on dynamic Grey prediction and support vector machine", *International Journal of Advancements in Computing Technology*, vol. 4, no. 5, pp. 148-156, 2012.
4. E. M. Zhao, Z. N. Zhang, N. Bohlooli, "Cost and load forecasting by an integrated algorithm in intelligent electricity supply network", *Sustainable Cities and Society*, vol. 60, Art no. 102243, 2020.
5. M. Dong and L. Grumbach, "A Hybrid Distribution Feeder Long-Term Load Forecasting Method Based on Sequence Prediction", *IEEE Transactions on Smart Grid*, vol. 11, no. 1, pp. 470-482, 2020.
6. Y. Ren, P. N. Suganthan, N. Srikanth, Gehan Amaratunga, "Random vector functional link network for short-term electricity load demand forecasting", *Information Sciences*, vol. 367–368, pp. 1078-1093, 2016.
7. D. C. Yang, J. E. Guo, Y. Z. Li, S. L. Sun and S. Y. Wang, "Short-term load forecasting with an improved dynamic decomposition-reconstruction-ensemble approach", *Energy*, vol. 263, Art no. 125609, 2023.
8. D. X. Niu, M. Yu, L.J. Sun, T. Gao and K. K. Wang, "Short-term multi-energy load forecasting for integrated energy systems based on CNN-BiGRU optimized by attention mechanism", *Applied Energy*, vol. 313, Art no. 118801, 2022.
9. L. Xu, X. Zhou, Y. Tao, L. Liu, X. Yu and N. Kumar, "Intelligent Security Performance Prediction for IoT-Enabled Healthcare Networks Using an Improved CNN", *IEEE Transactions on Industrial Informatics*, Vol. 18, No. 3, pp. 2063-2074, 2022.
10. R. Hu and S. Xiang, "Reversible Data Hiding By Using CNN Prediction and Adaptive Embedding", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 10196-10208, 2022.
11. J. Hu, M. Wu, P. Zhang and W. Pedrycz, "Prediction Performance Improvement via Anomaly Detection and Correction of Actual Production Data in Iron Ore Sintering Process," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 12, pp. 7602-7612, 2020.
12. L. Chen, M. Li, W. Su, M. Wu, K. Hirota and W. Pedrycz, "Adaptive Feature Selection-Based AdaBoost-KNN With Direct Optimization for Dynamic Emotion Recognition in Human–Robot Interaction", *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 5, no. 2, pp. 205–213, 2021.
13. Q. Xie, G. Tao, C. Xie and Z. Wen, "Abnormal Data Detection Based on Adaptive Sliding Window and Weighted Multiscale Local Outlier Factor for Machinery Health Monitoring", *IEEE Transactions on Industrial Electronics*, vol. 70, no. 11, pp. 11725-11734, 2023.
14. H. Fan, X. Lai, S. Du, W. Yu, C. Lu and M. Wu, "Distributed Monitoring With Integrated Probability PCA and mRMR for Drilling Processes", *IEEE Transactions on Instrumentation and Measurement*, Vol. 71, Art no, pp. 1-13, 2022.
15. L. Li, W. Li, Q. Du and R. Tao, "Low-Rank and Sparse Decomposition With Mixture of Gaussian for Hyperspectral Anomaly Detection", *IEEE Transactions on Cybernetics*, vol. 51, no. 9, pp. 4363-4372, 2020.