Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 14, No. 3, 2024

A LSTM-based time series prediction model for China's power supply and its performance evaluation

Zhihao Zhang

School of Information and Electrical Engineering, Hebei University of Engineering, Handan City, Hebei 056038, China

Abstract: With the rapid development of China's economy and the acceleration of urbanisation, the stability and sustainability of power supply has become increasingly important. In order to cope with the continuous growth of power demand, accurate power supply prediction has become an essential research direction. In this paper, a time series prediction model for power supply is constructed based on the long short-term memory network (LSTM). Power supply data is organised using methods such as data pre-processing and correlation analysis, and the LSTM model is used to predict future power supply. The performance of the prediction results was evaluated, and the results showed that the model can effectively capture the changing trends of power supply, with high reliability and accuracy, providing important support for scientific management and policy formulation in the power industry.

Keywords: Long short-term memory network (LSTM); Power supply; Time series prediction; Data preprocessing; Performance evaluation.

1. Introduction

Globally, the stability and sustainability of the power supply is an important cornerstone for driving economic development and social progress. With China's rapid economic growth and accelerating urbanisation, power demand is showing an increasing trend. At the same time, the transformation of the energy structure, the need for environmental protection, and policy guidance make it particularly important to accurately predict the power supply. In this context, a time series prediction model based on long short-term memory networks (LSTM) has become the focus of research because it excels at processing time series data and can effectively capture the time-dependent and nonlinear characteristics of the data [1].

LSTM is a special type of recurrent neural network designed to solve the problems of gradient disappearance and gradient explosion that traditional neural networks may encounter when processing long sequences of data. By introducing memory units and multiple gating mechanisms, LSTM can effectively retain and forget information in time series data. This allows LSTM to identify key factors affecting power supply and generate more accurate predictions based on historical data in the time series prediction of power supply [2].

This study aims to construct an LSTM-based time series prediction model for power supply and evaluate the prediction performance of the model through in-depth analysis of China's power supply data. First, the research will organise and clean the power supply-related data through steps such as data pre-processing and correlation analysis to ensure its quality and applicability. Then, the LSTM model will be used to predict future power supply, and the model will be verified and optimised using a variety of performance evaluation indicators to ensure the reliability and accuracy of the prediction results. Through this series of work, the research not only provides data support for the scientific management of power supply, but also provides a basis for decision-making in the formulation and implementation of energy policies [3].

2. Related work

In recent years, China's power supply problems have received widespread attention, especially against the backdrop of rapid economic growth and energy restructuring. The long short-term memory (LSTM) network-based power supply time series prediction model is an emerging technical approach that has shown its potential for processing complex time series data. The LSTM model, through its unique structure, can effectively capture the time-series dependencies and nonlinear characteristics of power supply data, thereby providing accurate predictions for the power industry [4].

First, the LSTM model is designed to overcome the problems of vanishing gradients and exploding gradients that traditional recurrent neural networks often encounter when processing long sequences. By introducing memory cells and multiple gating mechanisms, LSTM can selectively remember or forget information at each time step, which provides strong support for the prediction of power supply. Specifically, the synergy of the input gate, forget gate and output gate enables LSTM to dynamically adjust its focus on historical data, thereby improving the accuracy of the prediction.

Data pre-processing is a crucial step in the model construction process. By using the box plot method, we can effectively detect and remove outliers in the data to ensure the accuracy of the analysis. At the same time, the Spearman correlation coefficient was used for correlation analysis, which helped us identify the strong correlation between power generation and various influencing factors, which provided a basis for subsequent modelling.

In the actual prediction, we used the constructed LSTM model to perform time series prediction of China's power supply. The results show that from 2024 to 2060, power supply will continue to grow, especially from 2024 to 2042, and the growth rate of power supply will gradually stabilize during the period from 2042 to 2060. This trend reflects China's efforts in sustainable development and energy transformation in the power industry [5].

For the evaluation of the model performance, we used multiple performance indicators, including regression plots, error histograms, etc., which show that the LSTM model fits well on the training and test sets, with small prediction errors and high reliability. Through this comprehensive performance evaluation, we can better understand the performance of the model in practical applications and provide a scientific basis for decision-making in the power industry.

In summary, the LSTM-based power supply time series prediction model not only provides an effective forecasting tool for the power industry, but also provides important data support and a decision-making basis for the formulation and implementation of future energy policies. With the global energy transition and environmental protection becoming increasingly important, accurate power supply forecasting is particularly critical [6]. In the future, we will continue to optimise the model to further improve the accuracy and practicality of the forecast and contribute to the sustainable development of the power industry.

3. Model Development

3.1. Data preprocessing based on box-and-line diagrams

During the data collection process we counted the data of different factors and found that there were large differences in the amount of data between different factors, in order to observe the differences between these factors more intuitively, we chose stacked line graphs to visually analyse the differences between them.

Stacked line graphs of the collected data used:

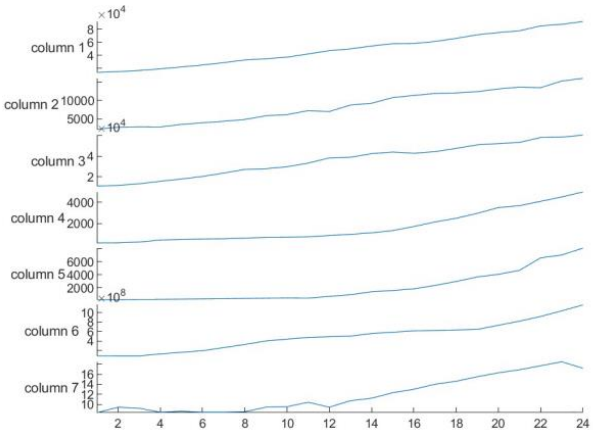

Fig.1 Stacked line graph of data

From the analysis of the above figure, we classify these factors to get four major categories of indicators: economy and industry, energy consumption and structure, population and society, and environment and emission, which is convenient for the subsequent treatment of the problem.

Using the box-and-line diagram method, when there are outliers in the data, especially when there are deviations from the larger deviation values, it will bring errors to the data analysis and model building [7]. Therefore outliers must be detected and removed. Since the data collected is not uniformly distributed and does not conform to normal distribution characteristics, box-and-line plot is used for numerical type characteristics for outlier detection.

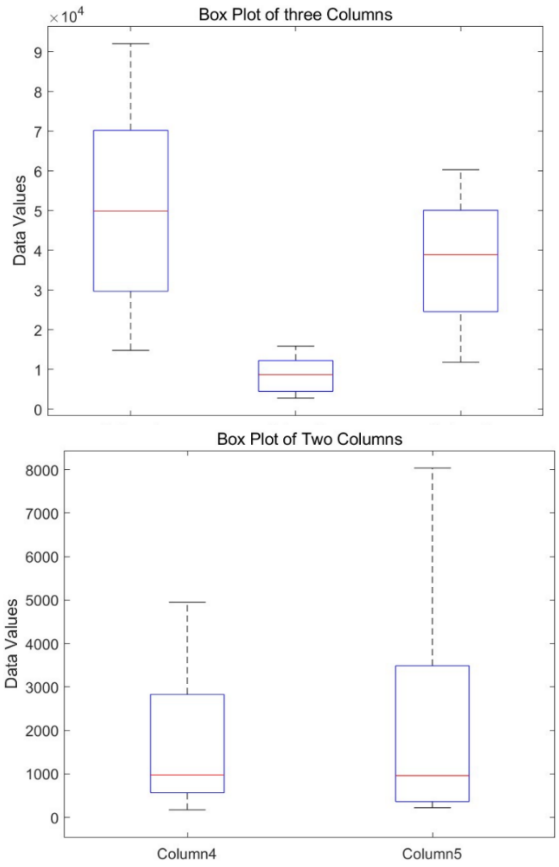

Fig.2 Box line diagram of outlier test

The principle of outlier detection through box-and-line plots is to identify outliers in the data by calculating the quartiles of the data as well as 1.5 times the interquartile distance, i.e., $Q1 - 1.5IQR$ and $\mathrm{Q3} + 1.5\mathrm{IQR}$. Boxplots can show the median, upper quartile, lower quartile, upper and lower edges, and potential outliers of the data. In this paper, the upper quartile is used to replace data greater than $\mathrm{Q3} + 1.5\mathrm{IQR}$, and the lower quartile is used to replace data less than $Q1 - 1.5IQR$, and the outliers are labelled in the box-and-line plot, as shown in Figure 2.

In the above figure, the middle line indicates the median, the upper and lower edges of the box indicate the upper and lower quartiles, respectively, the horizontal lines at the top and bottom of the figure indicate the upper and lower edges, and the uppermost and lowermost points are potential outliers [8]. As can be seen from the figure, the data for our selected indicators have fewer outliers, indicating that the quality of the data collected is relatively good.

3.2. Correlation analysis based on Spearman's coefficient

Firstly, the test for normal distribution of our selected data indicators was carried out as follows:

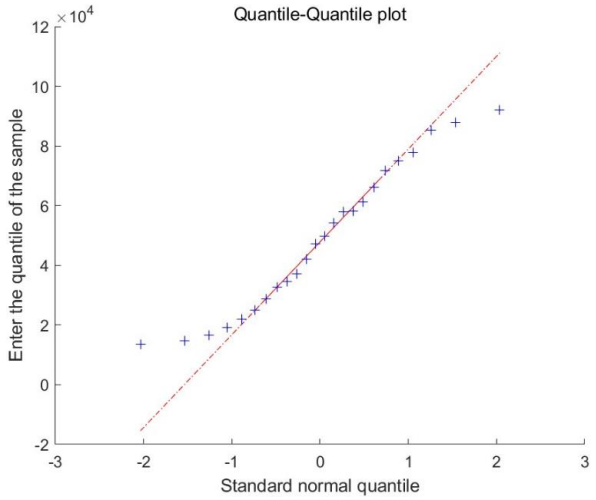

Fig.3 Normal distribution test

From this figure, it can be seen that the selected data does not conform to the normal distribution, so the Spearman correlation coefficient model can be chosen.

Spearman model development:

Spearman correlation analysis is the calculation of the degree of correlation between two or two data and measures the non-linear relationship between the variables, since the data we collected is non-normal, so this model can be used to calculate the data [9].

Spearman's correlation coefficient is calculated as follows: where R_i and S_i denote the rank of the values taken for observation i, $\bar{R}$ and $\bar{S}$ are the average ranks of variables x and y, respectively, and N is the total number of observations.

$$\rho_s = \frac{\sum_{i=1}^{N}(R_i - \bar{R})(S_i - \bar{S})}{[\sum_{i=1}^{N}(R_i - \bar{R})^2 \sum_{i=1}^{N}(S_i - \bar{S})^2]^{\frac{1}{2}}} \quad (1)$$

Data processing through metlab yielded a plot of correlation coefficients between the selected factors:

Fig.4 Plot of correlation coefficients

Spearman's correlation coefficient takes values between -1 and 1. When $\rho = -1$ it indicates that there is a perfect positive correlation between the ranks of the two variables, i.e., when one variable increases, the other also increases and follows some monotonic functional relationship. When $\rho = -1$, it means that there is a perfect negative correlation between the ranks of the two variables, i.e., when one variable increases, the other decreases and follows some monotonic functional relationship. When $\rho = 0$, it means that there is no linear or monotonic relationship between the ranks of the two variables, i.e., there is no correlation between them.

It can be concluded that there is a strong correlation between the amount of electricity produced (electricity supply) and the amount of electricity produced by hydropower, the amount of electricity produced by thermal power, the amount of electricity produced by nuclear power, the amount of electricity produced by wind power, the total input of the electricity production and supply industry, the share of primary electricity and other energy sources in the total amount of energy, the density of population, the emission of carbon dioxide, the area of arable land, the emission of methane, and the GDP per capita [10].

3.3. LSTM-based time series forecasting model

LSTM network is a recurrent neural network, and its biggest difference compared to other neural networks is the close connection between the hidden layer units. The hidden layer units do not exist independently, but are closely related to each other and the timing input of the previous node is closely related to it. This feature has important implications when dealing with time-series related data. LSTM neural networks can effectively remember long-term dependencies and overcome the problems of gradient explosion and gradient vanishing that can occur when dealing with long sequences of data.

The structural units in the LSTM model mainly consist of input gates, output gates, forgetting gates and self-connected memory unit state values. Their core function is to manage and deliver information to control what kind of information can be delivered to the current neuron and select how much information from the current neuron to the next neuron. Their structural units are shown in the following figure, and the values of these gates mainly depend on the values of the input x_t at the current time step, the hidden state h_{t-1} at the previous time step, and the memory unit state C_{t-1} at the previous time step [11].

Specifically, the input gate decides which input features are useful and need to be passed to the current neuron by using a Sigmoid activation function and a dot product operation. The output gate then calculates the information that needs to be output at the current time step based on the inputs of the current time step and the hidden state of the previous time step, and passes it on to the next time step. The forgetting gate then decides which information should be forgotten from the memory unit of the previous time step. The self-connected memory cell state values store past information and are updated based on the values of the input and forgetting gates.

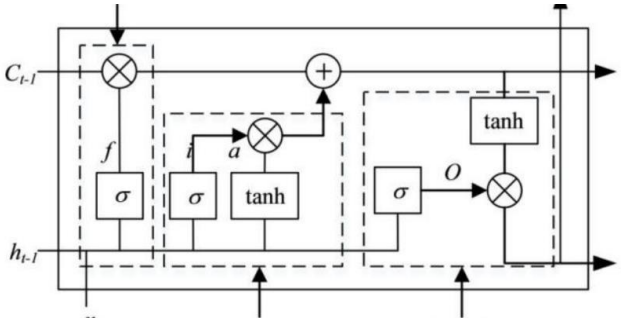

Fig.5 Structure of LSTM network

The forgetting gate relies on the output of the previous unit

as well as the input of the current unit, and thereafter again can rely on the value of the sigmoid function [0,1] to adjust the degree of forgetting of the previous unit's state, as detailed in the process described in the following equation.

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \quad (2)$$

The core function of the input gate is to identify the parameters that are required to be retained and those that are required to be updated, to determine the content of the demand update from a sigmoid, and thereafter to construct a new vector of candidate values using tanh.。

$$i_t = \sigma(W_i \cdot [h_{t-1}] + b_i) \quad (3)$$

$$\dot{C}_t = tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (4)$$

Update the old cell C_{t-1} to the new cell C_t, for which the following formula is applied to reach the process:

$$C_t = f_t^* C_{t-1} + i_t^* \dot{C}_t^* \quad (5)$$

Within the aforementioned formula $f_t * C_{t-1}$., the data needed is discarded and $i_t * \widetilde{C}_t$, the corresponding new candidate parameter, combined with the specific decision to carry out the change.

The output gate requires the application of the sigmoid function to confirm the output of the unit, after which the state is applied to the tanh function to carry out targeted processing, and multiply it with the output, and finally output the required data. The formula is:

$$O_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (6)$$

$$h_t = O_t^* \tanh(C_t) \quad (7)$$

Where: x_t is the current input vector; h_t is the current implicit layer unit output; h_{t-1} is the implicit layer output of the previous moment; C_t is the unit state of the current moment; C_{t-1} is the unit state of the previous moment; σ is the sigmoid activation function; tanh is the tangent function; f, i, O is the forgetting gate, the input gate, and the output gate; and w, b is the weight matrix and the deviation vector.

3.4. Time series prediction results for LSTM

Forecasted trend map of China's electricity supply from 2024 to 2060:

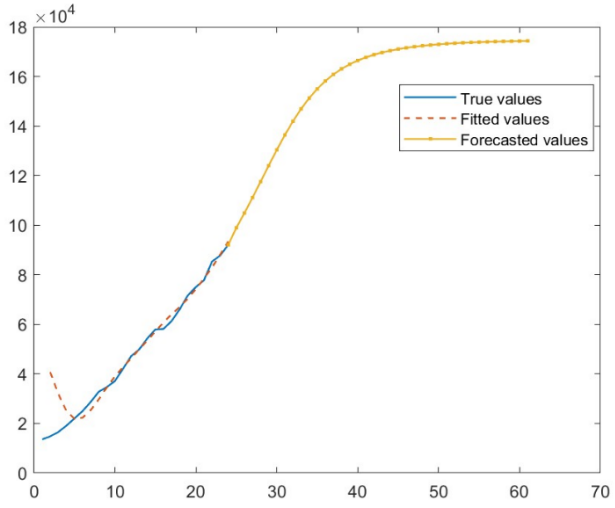

Fig.6 Map of forecast results

The data in the table 1 has been retained to one decimal place. From the forecast chart, it can be seen that China's electricity supply will increase year by year during the period of 2024-2042, and the change of China's electricity supply will be small and may tend to a stable value during the period of 2042-2060.

Table.1 Table of LSTM prediction results (Billion kilowatt hours)

vintages	electricity supply	vintages	electricity supply
2024	98943.4	2043	170404.1
2025	104875.8	2044	171027.9
2026	111123.3	2045	171553.8
2027	117564.9	2046	171998.6
2028	124044.2	2047	172375.9
2029	130380.4	2048	172697.0
2030	136394.6	2049	172971.1
2031	141937.2	2050	173205.6
2032	146907.8	2051	173406.9
2033	151262.0	2052	173579.9
2034	155004.8	2053	173729.2
2035	158176.7	2054	173858.0
2036	160838.5	2055	173969.6
2037	163058.7	2056	174066.4
2038	164904.7	2057	174150.5
2039	166438.3	2058	174223.7
2040	167713.1	2059	174287.5
2041	168774.9	2060	174343.1
2042	169661.5		

3.5. Performance evaluation of LSTM time series models

Regression chart:

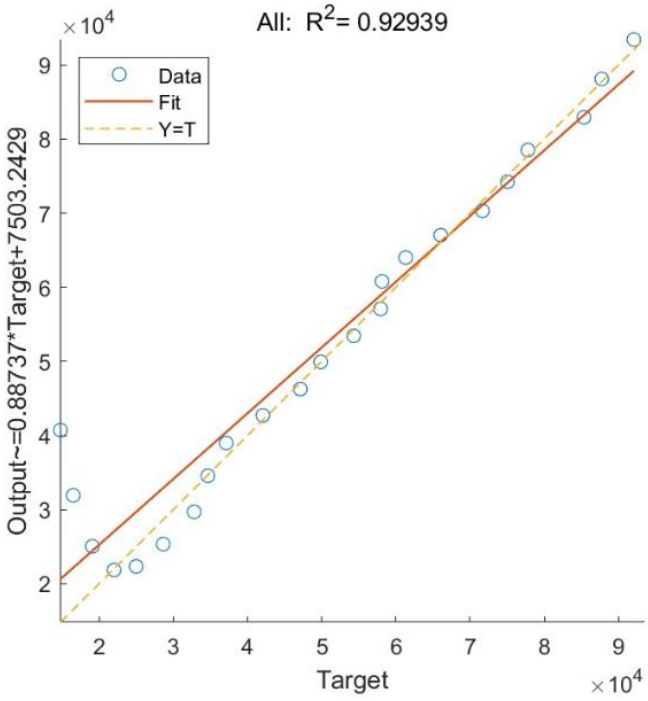

Fig.7 Fitting effect diagram

The model training process is shown below:

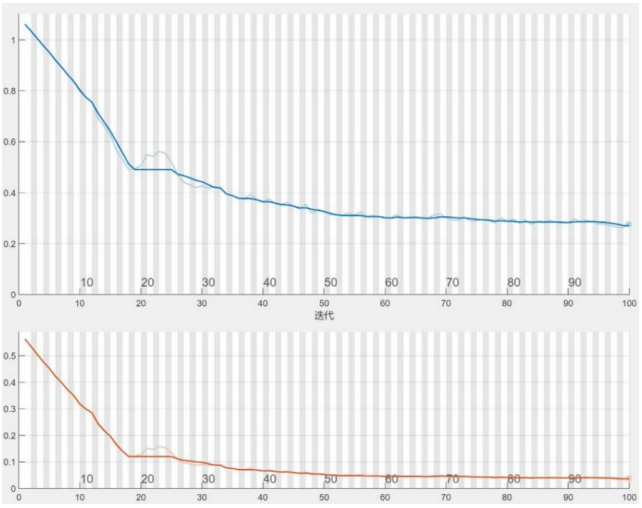

Fig.8 Diagram of the model training process

Histogram of the error of the LSTM model training test set:

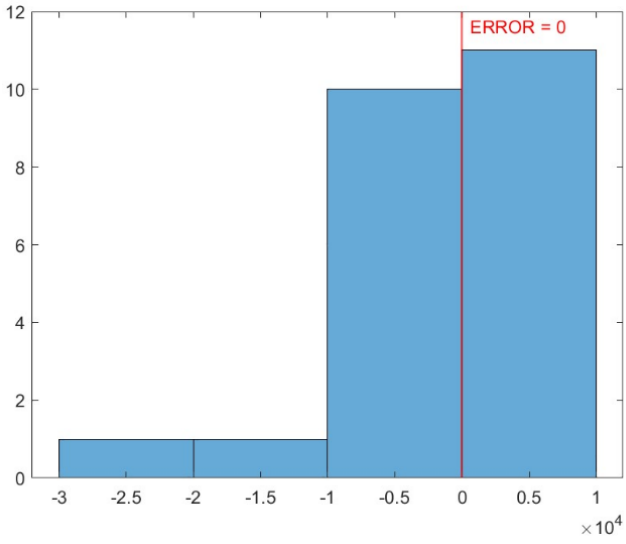

Fig.9 Error histogram

We then applied Spearman's correlation analysis to explore the correlation coefficients between the factors. Through this analysis, we were able to reveal the characteristics of the strong correlation between the factors in the data. Such an understanding helps us to gain a deeper understanding of the interactions between the factors and provides guidance for further modelling work.

Finally, we developed a time series forecasting model for electricity supply using the LSTM model with data up to 2060. We also analysed the accuracy of the prediction results. The error of the model training test set is very small, and the regression plots show that the model has a good fit [12]. So the model we chose has a high confidence in predicting future electricity supply. It is predicted that China's electricity supply will increase year by year during the period of 2024-2042, and the change of China's electricity supply during the period of 2042-2060 will be small and may converge to a stable value. The above steps enable a better understanding of the data, explore the potential information in the data, and provide a scientific basis for decision-making.

4. 4 Discussion

Against the backdrop of growing global energy demand, accurate forecasting of electricity supply has become an important issue for sustainable development. Time series forecasting models based on long short-term memory networks (LSTM) have gradually become the focus of researchers due to their unique advantages in dealing with nonlinear and time series data. The LSTM model effectively solves the gradient disappearance and gradient explosion problems of traditional neural networks when processing long sequence data through its structural design, so that it can better capture the complex relationships in time series data of electricity supply. Especially in a country like China with its rapid economic development and rising energy demand, an LSTM-based power supply prediction model not only provides a practical tool for the power industry, but also provides a scientific basis for the government's energy policy formulation.

In this study, we first laid a solid foundation for building the LSTM model through data preprocessing and correlation analysis. During data preprocessing, the box plot method was used to detect and remove outliers, ensuring the quality of the data. At the same time, Spearman's correlation coefficient analysis revealed the strong correlation between power supply and various factors. These steps provided a clear direction for the subsequent model construction. During the training and validation of the model, the LSTM's unique gating mechanism dynamically adjusts the focus on historical data, effectively improving the accuracy and reliability of the prediction.

After the model was trained, our prediction results show that China's power supply will continue to grow from 2024 to 2060, with a significant growth rate between 2024 and 2042. This trend not only reflects the efforts of China's power industry to meet the growing demand for electricity, but also reflects a positive response to the transition to renewable energy and environmental protection policies. However, after 2042, the growth rate of electricity supply will gradually level off, which suggests that when planning the future energy structure, we should pay more attention to the proportion of renewable energy and the stability of electricity supply.

In the evaluation of model performance, we used various indicators such as regression plots and error histograms. These results show that the LSTM model has a good fit with small errors on the training and test sets, and is highly reliable. Through this comprehensive evaluation, we not only verified the effectiveness of the model, but also provided important data support for scientific decision-making in the power industry. In future research, we will continue to optimise the LSTM model and explore more factors that affect power supply, with a view to improving the accuracy and practicality of the prediction, and thus contributing more to the sustainable development of China's power industry.

In summary, the LSTM-based power supply time series prediction model has broad prospects for application in the power industry. It not only provides new ideas for prediction, but also provides strong support for achieving China's energy transition and sustainable development goals. With the continuous advancement of data science and technology, future research is expected to achieve more significant results in the field of power supply forecasting, contributing to global energy security and environmental protection.

5. Conclusion

In this study, we constructed a long short-term memory network (LSTM)-based power supply time series prediction model to accurately predict the future trend of China's power

supply. Through in-depth data pre-processing and correlation analysis, we laid a solid foundation for the construction of the model, ensuring the high quality and applicability of the input data. The LSTM model, with its unique gating mechanism, can effectively deal with the gradient disappearance and gradient explosion problems faced by traditional neural networks when processing long sequence data, thereby capturing the deep-level temporal dependencies and nonlinear characteristics in complex power supply data.

The training and validation results of the model show that the LSTM model performs well in predicting power supply, especially during the period from 2024 to 2042, when the growth trend of power supply is significant. We predict that China's power supply will continue to grow until 2060, especially in the early stages, which reflects China's efforts to meet economic growth and power demand, as well as its positive response to renewable energy and environmental protection policies. At the same time, the growth rate of power supply will gradually slow down after 2042, which poses new challenges for the future restructuring of the energy sector and reminds us that more attention needs to be paid to the proportion of renewable energy and its impact on the stability of power supply when formulating energy policies.

Through a comprehensive evaluation of the model's performance, we have verified the high reliability and predictive accuracy of the LSTM model. These results not only provide data support for scientific decision-making in the power industry, but also provide an important reference for the government when formulating energy policies. In future research, we will further optimise the model and explore more potential factors that affect power supply, striving to improve the predictive accuracy and practicality while contributing more wisdom and strength to the sustainable development of China's power industry.

In summary, the LSTM-based power supply time series prediction model provides an effective tool for decision-making and planning in the power industry and shows great potential for application. Against the backdrop of global energy transition and sustainable development, accurate power supply forecasting will be key to achieving environmental protection and the rational use of resources. With the continuous development of data science and machine learning technology, we look forward to making greater breakthroughs in the field of power supply forecasting in the future and making a greater contribution to achieving global energy security.

References

[1] Kumar, I., Tripathi, B. K., & Singh, A. (2023). Attention-based LSTM network-assisted time series forecasting models for petroleum production. Engineering Applications of Artificial Intelligence, 123, 106440.

[2] Li, K., Huang, W., Hu, G., & Li, J. (2023). Ultra-short term power load forecasting based on CEEMDAN-SE and LSTM neural network. Energy and Buildings, 279, 112666.

[3] Guo, H., Chen, Q., Zheng, K., Xia, Q., & Kang, C. (2021). Forecast aggregated supply curves in power markets based on LSTM model. IEEE Transactions on power systems, 36(6), 5767-5779.

[4] Bulut, M. (2021). Hydroelectric generation forecasting with long short term memory (LSTM) based deep learning model for turkey. arXiv preprint arXiv:2109.09013.

[5] Han, H., Liu, H., Zuo, X., Shi, G., Sun, Y., Liu, Z., & Su, M. (2022). Optimal sizing considering power uncertainty and power supply reliability based on LSTM and MOPSO for SWPBMs. IEEE Systems Journal, 16(3), 4013-4023.

[6] Jailani, N. L. M., Dhanasegaran, J. K., Alkawsi, G., Alkahtani, A. A., Phing, C. C., Baashar, Y., ... & Tiong, S. K. (2023). Investigating the power of LSTM-based models in solar energy forecasting. Processes, 11(5), 1382.

[7] Wang, D., Gan, J., Mao, J., Chen, F., & Yu, L. (2023). Forecasting power demand in China with a CNN-LSTM model including multimodal information. Energy, 263, 126012.

[8] Chen, Y., Cui, S., Chen, P., Yuan, Q., Kang, P., & Zhu, L. (2021). An LSTM-based neural network method of particulate pollution forecast in China. Environmental Research Letters, 16(4), 044006.

[9] Branco, N. W., Cavalca, M. S. M., Stefenon, S. F., & Leithardt, V. R. Q. (2022). Wavelet LSTM for fault forecasting in electrical power grids. Sensors, 22(21), 8323.

[10] Li, Q., Yang, Y., Yang, L., & Wang, Y. (2023). Comparative analysis of water quality prediction performance based on LSTM in the Haihe River Basin, China. Environmental Science and Pollution Research, 30(3), 7498-7509.

[11] He, Y., & Tsang, K. F. (2021). Universities power energy management: A novel hybrid model based on iCEEMDAN and Bayesian optimized LSTM. Energy Reports, 7, 6473-6488.

[12] Li, Y., Tong, Z., Tong, S., & Westerdahl, D. (2022). A data-driven interval forecasting model for building energy prediction using attention-based LSTM and fuzzy information granulation. Sustainable Cities and Society, 76, 103481.