

SHORT-TERM POWER LOAD FORECASTING MODEL BASED ON VARIATIONAL MODE DECOMPOSITION AND SYMMETRICAL EXTENDED LONG SHORT-TERM MEMORY

JIANFAN ZHAI¹, LEI WU¹, HUAMING ZHANG^{1,*} AND TIE ZHONG²

¹CGN Wind Power Co., Ltd.

No. 188, South Fourth Ring Road West, Fengtai District, Beijing 100071, P. R. China
{zhaijianfan; wulei}@cgnpc.com.cn; *Corresponding author: zhanghuaming@cgnpc.com.cn

²Department of Communication Engineering

Northeast Electric Power University

No. 169, Changchun Road, Jilin 132012, P. R. China

zht@neepu.edu.cn

Received August 2025; revised December 2025

ABSTRACT. *The technical difficulty in short-term load forecasting in power systems stems from their non-stationary characteristics, random fluctuations, and the influence of complex factors. These characteristics often lead to a decline in the model's predictive performance. To address this issue, this paper proposes a hybrid forecasting framework that integrates time-series similarity analysis with an extended Long Short-Term Memory (LSTM) network (xLSTM). Specifically, we first use the Variational Mode Decomposition (VMD) algorithm to decouple the original non-stationary load curves into several stationary components. On this basis, xLSTM employs a symmetrical architecture to discern underlying features in each component for precise prediction. Meanwhile, we also construct a training dataset, which is combined with grey relational analysis and a historical-similar time-series screening mechanism to enhance performance and prediction accuracy further. To verify effectiveness, actual power loads in New Toronto (Canada) and a power plant in Northeast China are selected, and we also use Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) to evaluate prediction accuracy quantitatively. Moreover, some benchmark models are also selected as comparison methods. The results demonstrate that the proposed symmetrical framework achieves impressive performance, further verifying its effectiveness in short-term power load forecasting.*

Keywords: xLSTM, Symmetrical framework, Similar time-series, VMD, Stationary components

1. Introduction. The time-varying characteristics of power load and the complexity of user scale make accurate prediction a key technological requirement for maintaining the dynamic balance of the power system [1-3]. Among the current mainstream prediction methods, time series analysis is suitable for regular data with strong stationarity, but it is challenging to handle scenarios with drastic load fluctuations [4-6]. Machine learning and support vector machines have shown advantages in processing non-linear data, but their hyperparameter optimization highly depends on manual experience [7-11]. Research shows that single-prediction models have limitations in handling power load sequences with both high volatility and nonlinearity. In recent years, deep learning techniques have dominated the field of load forecasting. Particularly, attention-based mechanisms and Transformer models have been extensively applied due to their ability to capture global dependencies

[12,13]. For instance, Li and Cai [12] proposed a Transformer-based framework that effectively models long-range dependencies in time series via Crossformer. However, these Transformer variants often suffer from high computational complexity and may overlook local temporal mutations crucial for short-term load fluctuations. On the other hand, Recurrent Neural Networks (RNNs) and their variants remain highly effective for sequential data. Recent studies, such as Wang and Li [14], have explored hybrid LSTM models to enhance prediction stability. Despite these advancements, traditional LSTM units still face challenges with gradient vanishing when processing very long sequences, and their scalar memory structures limit the capacity to store complex feature interactions. Consequently, exploring the newly proposed xLSTM, which utilizes exponential gating and matrix memory to overcome these limitations, combined with advanced decomposition techniques, presents a promising yet underexplored direction for improving forecasting accuracy. Therefore, combined prediction models that integrate the advantages of multiple methods have gradually become the focus of research. Data quality optimization is a key strategy for improving prediction accuracy. Existing power load prediction methods are usually divided into two primary technical paths. Firstly, the method is based on influencing factors. It quantifies the coupling between meteorological factors (temperature/humidity) and load changes using grey relational analysis, selects exogenous variables such as social and economic activities, based on the Pearson correlation coefficient, constructs a multi-dimensional feature input set, and then builds a multi-dimensional feature data set [15]. Secondly, the prediction method is based on similar days. It uses clustering algorithms to identify the spatio-temporal similarity laws of historical load sequences, reduces data noise interference via an optimal similar-day matching mechanism, and enhances the model's ability to capture periodic characteristics [16-18]. These two types of methods optimize data quality from the perspectives of external factor correlation mining and internal factor law extraction, respectively [19-22].

Madrid and Antonio proposed a machine learning model to enhance the accuracy of 168-hour forecasting [23]. This model incorporates features from multiple sources, including historical load data, weather data, and holiday data. Taylor and Buizza modeled the unspecified nonlinear relationship between load and weather variables, exploring the application of weather-integrated forecasting in ANN for advanced load forecasting 1 to 10 days in advance [24]. Zhang et al. proposed a method for matching similar days based on an optimized clustering algorithm constructed using the human comfort index [25]. This method combines an LSTM neural network to predict hotel building loads, and its accuracy and efficiency are verified using actual measurement data. Kim et al. combined a hybrid convolutional neural network and an LSTM network with a similar data selection model, selecting data with similar dates as the input data. They verified the effectiveness of the proposed method by comparing the performance of convolutional neural networks and long short-term memory networks [26].

To address the problems of nonlinear coupling and random fluctuations in short-term power load forecasting, this paper constructs a hybrid forecasting framework (VMD Extended Long Short-Term Memory, VMD-xLSTM) that integrates similar sequence features, VMD technology, and an improved extended short-term memory network [27]. The implementation process of this model consists of three stages:

- 1) Construction of a multi-dimensional feature fusion dataset: The morphological similarity patterns of the load sequence are mined through a combined algorithm of curve trend degree and difference degree. Grey relational analysis is used to quantify the coupling strength between external variables such as temperature and humidity and load fluctuations, thereby forming a multi-dimensional feature input matrix. This effectively overcomes the problem of a single input dimension in traditional methods.

2) Modal decomposition-forecasting coupling mechanism: The VMD technology is used to decompose the original load sequence into multiple relatively smooth sub-modal sequences. An xLSTM forecasting model is constructed for each sub-modal sequence, and its gating mechanism is utilized to optimize the ability to capture time-series features.

3) Integrated output of forecasting results: The forecasting results of each sub-model are linearly superimposed to generate the final load forecasting value.

The experimental results on the power load dataset of the New Toronto area in Canada show that, compared with other forecasting models, the method proposed in this paper has higher forecasting accuracy, verifying the role of feature fusion and modal decomposition strategies in improving forecasting accuracy.

The remainder of this paper is organized as follows. Section 2 details the construction of the multi-dimensional load and feature fusion dataset, explaining the mechanisms of similar sequence screening and grey relational analysis. Section 3 elaborates on the proposed hybrid forecasting framework, providing a comprehensive theoretical analysis of the VMD algorithm and the architectural innovations of the symmetrical xLSTM network. Section 4 presents the experimental design, including data sources, evaluation metrics, and a comparative discussion of the forecasting results against benchmark models. Finally, Section 5 concludes the paper by summarizing the key findings and outlining directions for future research.

2. Construction of Load and Feature Fusion Dataset. This paper proposes a data construction method based on multi-dimensional feature fusion. By combining the optimal similar load sequence with multi-source influencing factor feature data, a high-information-density training dataset is formed, laying a data foundation for efficient feature mining by the VMD-xLSTM model. Figure 1 shows the construction process of this fused dataset. First, using the sliding window technique, the optimal similar load sequence is screened based on trend similarity and difference. Second, normalization and outlier correction are applied to the original load data, and high-influence features are selected using the grey relational analysis method. Finally, the screened similar load sequence and influencing factor features are tensor-spliced to construct a fused feature matrix. This fused

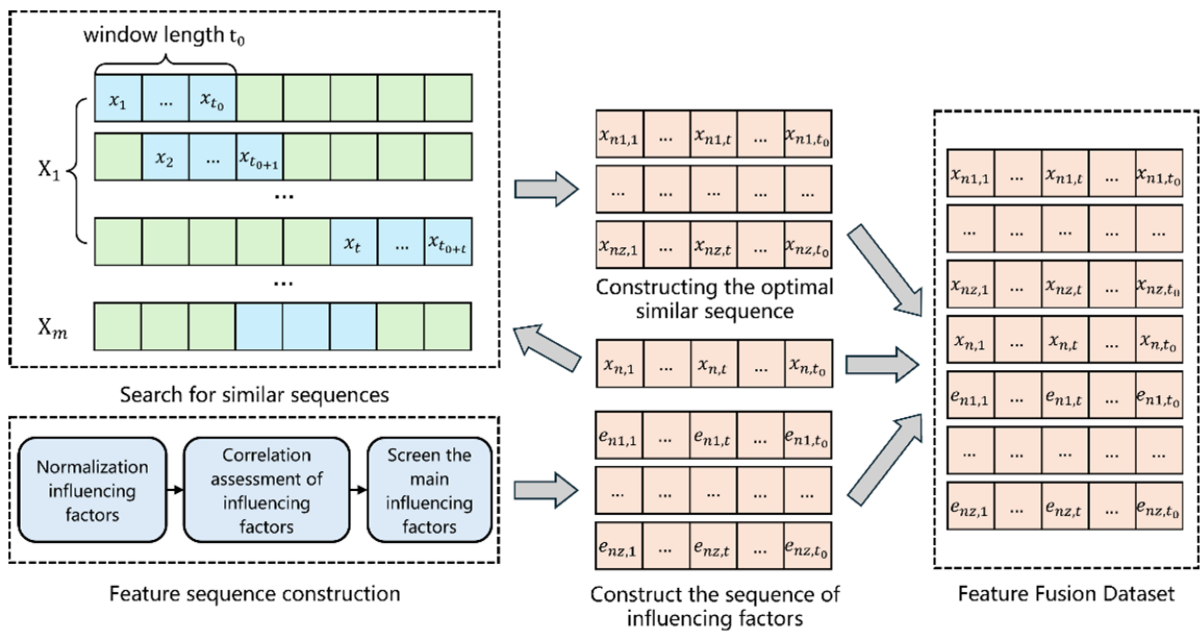


FIGURE 1. Flowchart for building a fusion dataset

dataset effectively addresses the single data type issue in traditional methods, enabling the VMD-xLSTM model to simultaneously capture the periodic evolution law of the load sequence and its response to sudden disturbances, and providing reliable data support for subsequent modal decomposition and prediction optimization.

2.1. Selection of optimal similar load sequence set. Short-term load forecasting systems often employ combined models to improve forecasting accuracy and adaptability to various scenarios. Among them, the accuracy of selecting similar-load sequences is a crucial factor affecting the forecasting results. The accuracy of traditional similar load forecasting methods largely depends on the selection of influencing factors based on manual experience, which is highly subjective. However, the load curves of the power systems already contain substantial potential information. By matching the load curves of the forecasting day with those of historical days for similar load sequences, we can not only extract the implicit features of the forecasting day but also avoid errors caused by manual selection of factors, thereby improving forecasting accuracy. Based on the above analysis, this paper proposes a screening mechanism based on similar load sequences: by combining quantitative analysis of curve trends and differences, the optimal set of similar load sequences is matched.

In the field of power load forecasting, the curve trend is defined as the vertical translation similarity between two curves (i.e., the maximum degree of coincidence when only vertical translation is allowed). This indicator quantifies the trend similarity between any two load curves through the following formula:

$$R_{mn}(m, t) = 2 \cdot \frac{E(X_{m,t+t_0}X_{n,t_0}) - E(X_{m,t+t_0})E(X_{n,t_0})}{D(X_{m,t+t_0}) + D(X_{n,t_0})} \quad (1)$$

Among them, $X_{m,t+t_0}$ represents the power load sequence of historical day m from time t to time t_0 , represented by a set of t_0 sampling points $x_{m,t}, \dots, x_{m,t+t_0}$ with equal probability distribution; X_{n,t_0} represents the power load sequence of day n before prediction time t_0 , represented by a set of t_0 sampling points $x_{n,1}, \dots, x_{n,t_0}$ with equal probability distribution; t_0 is the length of the sequence. $x_{m,t}$ and $x_{n,t}$ are the numerical values of the historical daily power load sequence $X_{m,t+t_0}$ and the predicted daily power load sequence X_{n,t_0} at the t -th point of their respective sequences; $E(\cdot)$ represents the mathematical expectation that $D(\cdot)$ is the variance. The greater the trend similarity, the more similar the shapes of two similar load curves are.

The curve trend-degree model can effectively characterize the morphological similarity of load curves. However, when the power load shows an overall upward or downward trend, even though the trend similarity is relatively high, the actual numerical deviation may still increase significantly. To overcome this limitation, this study introduces a dual-evaluation mechanism for curve difference degree based on trend-similarity screening.

The curve difference degree index quantifies the absolute numerical difference between load curves. The smaller its value, the smaller the error between the curves. By jointly applying the trend and difference degrees, truly similar sequences can be more accurately identified. Specifically, the algorithm first screens candidate curves whose trend similarity meets the standard, and then selects the optimal match with the slightest difference degree. The dual evaluation mechanism constructed in this way significantly improves the reliability of similar sequence screening. The calculation formula is as follows:

$$O_{mn}(m, t) = \sqrt{\sum_{k=t}^{t_0} |x_{m,k} - x_{n,k}|^2} \quad (2)$$

When used in combination, the difference degree needs to be normalized to unify its scale with that of the trend degree. Finally, by subtracting the normalized difference degree from the trend degree, we can obtain the final similarity coefficient V_{mn} for the power load sequences between time points t and $t + t_0$ on historical day m and t_0 time points prior to day n in the prediction period, as shown in Formula (3). The larger the similarity coefficient is, the more similar the two load curves are.

$$V_{mn}(m, t) = R_{mn}(m, t) - \frac{O_{mn}(m, t)}{\sum_{k=t}^{t_0} x_{n,k}} \quad (3)$$

The selection process of the optimal similar time period load sequence set is as follows.

1) Set a sliding time window with window length h , and slide through all historical load sequences in the historical load data according to Formula (3).

2) Calculate similarity coefficients for all historical sequences, sort them in descending order, and select the top z load curves to construct the optimal similar load sequence set $[X_{n1,t_0}; X_{n2,t_0}; \dots; X_{nz,t_0}]$ corresponding to time t_0 on the prediction day n .

2.2. Selection of main influencing factor feature sequences. Power load curves contain multidimensional characteristics of system operation. However, due to their non-linear fluctuation characteristics, traditional prediction models struggle to extract these features directly. In this study, we use the characteristic sequences of major influencing factors as model inputs, enabling the model to better learn differential changes in these features. Power loads are generally influenced by economic factors (e.g., electricity prices) and environmental factors (e.g., temperature, humidity, and wind speed), exhibiting regular fluctuation patterns. Therefore, we can use the Grey Relational Analysis (GRA) method to calculate grey correlation coefficients and perform correlation sorting, thereby quantifying and screening the factors. The specific steps of the GRA algorithm are as follows.

Firstly, all data undergo normalization processing. Subsequently, the grey correlation coefficients between the influencing factor sequence E_m^i and the daily load sequence X_m are calculated using Equation (4).

$$\xi_m^i = \sum_{t=1}^{t_0} \frac{\min_i \min_t |\Delta_m^i(t)| + \rho \max_i \max_t |\Delta_m^i(t)|}{|\Delta_m^i(i)| + \rho \max_i \max_t |\Delta_m^i(t)|} \quad (4)$$

Among them, $\Delta_m^i(t) = X_m(t) - E_m^d(t)$ and $\min_i \min_t |\Delta_m^i(t)|$ represents the second level minimum difference, $\max_i \max_t |\Delta_m^i(t)|$ represents the second level maximum difference, ρ is the resolution coefficient, and t_0 represents the sequence length of E_m^i .

Finally, the grey correlation coefficients of all historical days are averaged to obtain the correlation coefficients of each influencing factor:

$$r_i = \frac{1}{N} \sum_{m=1}^N \xi_i(m) \quad (5)$$

Among them, r_i is the correlation coefficient of the i -th influencing factor ($r_i \in (0, 1)$).

Based on the load dataset of the New Toronto Area, five typical influencing factors including temperature, humidity, wind speed, rainfall, and electricity prices were selected for correlation analysis. The correlation coefficients of each influencing factor were obtained through GRA, and the results are shown in Table 1.

From Table 1, it can be seen that among the above influencing factors, there is a strong correlation between temperature and electricity load ($\mu \in (0.8, 1.0)$), a strong correlation between electricity prices and humidity and electricity load ($\mu \in (0.65, 0.8)$), a moderate

TABLE 1. Correlation assessment of load influencing factors

Influence factor	Humidity (%)	Electricity price (yuan $\cdot$ kW $^{-1}$)	Wind speed (m $\cdot$ s $^{-1}$)	Rainfall (mm $\cdot$ h $^{-1}$)	Temperature ($^{\circ}$ C)
Correlation coefficient	0.712	0.677	0.231	0.469	0.913

correlation between electricity load and rainfall ($\mu \in (0.3, 0.65)$), and a weak correlation with wind speed ($\mu \in (0, 0.3)$). Based on the above analysis, humidity, temperature, and electricity prices are selected as the main influencing factors of the load. Collect feature data at time t_0 before the prediction date, construct the main influencing factor feature sequence set $[E_{n1,t_0}; E_{n2,t_0}; E_{n3,t_0}]$, and merge it with the feature data at time t_0 before the prediction date and the optimal similar load sequence set $[X_{n1}; X_{n2}; \dots; X_{nz}]$ to obtain the feature fusion dataset $\tilde{X}_{n,t_0}$, as shown in Equation (6).

$$\tilde{X}_{n,t_0} = \begin{bmatrix} X_{n1,t_0} \\ X_{n2,t_0} \\ \dots \\ X_{nz,t_0} \\ X_{n,t_0} \\ E_{n1,t_0} \\ E_{n2,t_0} \\ E_{n3,t_0} \end{bmatrix} = \begin{pmatrix} x_{n1,1} & \dots & x_{n1,t} & \dots & x_{n1,t_0} \\ x_{n2,1} & \dots & x_{n2,t} & \dots & x_{n2,t_0} \\ \dots & \dots & \dots & \dots & \dots \\ x_{nz,1} & \dots & x_{nz,t} & \dots & x_{nz,t_0} \\ x_{n,1} & \dots & x_{n,t} & \dots & x_{n,t_0} \\ e_{n1,1} & \dots & e_{n1,t} & \dots & e_{n1,t_0} \\ e_{n2,1} & \dots & e_{n2,t} & \dots & e_{n2,t_0} \\ e_{n3,1} & \dots & e_{n3,t} & \dots & e_{n3,t_0} \end{pmatrix} \quad (6)$$

Among them, $x_{nz,t}$ represents the numerical value of the z -th similar load sequence on prediction day n at its t -th point, and $e_{nz,t}$ represents the numerical value of the characteristic sequence of the z -th main influencing factor on prediction day n at its t -th point.

3. Optimization of VMD xLSTM Prediction Model Using Variational Mode Decomposition Algorithm. Given the nonlinear and nonstationary characteristics of power load time series data, this paper employs Variational Mode Decomposition (VMD) for signal preprocessing. The original load sequence is decomposed into multiple Intrinsic Mode Functions (IMFs), effectively eliminating noise interference while achieving data stationarization. Compared to traditional methods such as Empirical Mode Decomposition (EMD), VMD effectively suppresses mode mixing, providing high signal-to-noise-ratio input features for subsequent predictive models.

3.1. Variational mode decomposition algorithm.

3.1.1. Principle of variational mode decomposition. VMD constructs a constrained variational problem to decompose the original signal into multiple intrinsic mode components. Each intrinsic mode component fluctuates around a central frequency in the frequency domain and exhibits finite bandwidth, thereby enabling precise decoupling of multi-scale, non-stationary signals. The specific steps are as follows.

Step 1: Construct a variational problem. Decompose the raw sequence $f(i)$ to be processed into K variational modal components u_k with different center frequency bandwidths, while ensuring that the sum of estimated bandwidths for each modality is minimized. The specific variational constraint expression can be defined as

$$\begin{cases} \min_{\{u_k\}, \{\omega_k\}} \left\{ \sum_{k=1}^K \left\| \partial_t [B_k(i)] e^{-j\omega_k i} \right\|_2^2 \right\} \\ \text{s.t.} \quad \sum_{k=1}^K u_k(i) = f(i) \end{cases} \quad (7)$$

Among them, the original sequence $f(i)$ is the optimal similar daily load sequence set $[X_{n1}; X_{n2}; \dots; X_{nZ}]$, $\{u_k\}$ and $\{\omega_k\}$ represent the set of all sub-modes and their corresponding center frequencies, $B_k(i) = (\delta(i) + \frac{j}{\pi i}) \cdot u_k(i)$ represents the analytical signal of the k -th sub mode after Hilbert demodulation, $\delta(i)$ is the unit pulse signal, and $u_k(i)$ represents the modal function of the k -th sub mode.

Step 2: Lagrange transformation. Introduce Lagrange multiplier λ and penalty factor α to solve the optimal solution problem with variational constraints, and transform the constrained problem in Equation (7) into an unconstrained variational problem in Equation (8). Among them, α ensures the accuracy of signal reconstruction in high noise environments, while λ ensures the strictness of constraint conditions.

$$\begin{aligned} L(\{u_k\}, \{\omega_k\}, \lambda) = & \alpha \sum_k \left\| \partial_t [B_k(i)] e^{-j\omega_k i} \right\|_2^2 + \left\| f(i) - \sum_k u_k(i) \right\|_2^2 \\ & + \left\langle \lambda(i), f(i) - \sum_k u_k(i) \right\rangle \end{aligned} \quad (8)$$

Step 3: Alternating updates. Initialize $\{u_k^1\}$, $\{\omega_k^1\}$ and λ^1 . By using the alternating direction multiplier method to solve Equation (8), the update formulas for each main variable can be obtained as shown in Equation (9). Based on Equation (8), $\{u_k^1\}$, $\{\omega_k^1\}$ and λ^1 are iteratively updated.

$$\begin{cases} \hat{u}_k^{s+1}(\omega) = \frac{\hat{f}(\omega) - \sum_{\nu \neq k}^K \hat{u}_\nu^s(\omega) + \frac{\hat{\lambda}^s(\omega)}{2}}{1 + 2\alpha(\omega - \omega_k)^2} \\ \omega_k^{s+1} = \frac{\int_0^\infty \omega |\hat{u}_k(\omega)|^2 d\omega}{\int_0^\infty |\hat{u}_k(\omega)|^2 d\omega} \\ \hat{\lambda}^{s+1}(\omega) = \hat{\lambda}^s(\omega) + \tau \left(\hat{f}(\omega) \right) - \sum_k \hat{u}_k^{s+1}(\omega) \end{cases} \quad (9)$$

Among them, s is the number of iterations, ω represents the frequency, τ is the noise margin, $\hat{u}_k^{s+1}$ and ω_k^{s+1} represent the modal function and center frequency of the k -th sub mode at $s+1$ iterations, and $\hat{\lambda}^{s+1}$ is the Lagrange multiplier value of the VMD algorithm at $s+1$ iterations. The forms of $f(i)$, $u_k^{s+1}(i)$, $u_e^s(i)$, and $\lambda^s(i)$ after Fourier transform are represented by $\hat{f}(\omega)$, $\hat{u}_k^{s+1}(\omega)$, $\hat{u}_e^s(\omega)$, and $\hat{\lambda}^s(\omega)$, respectively.

$$\sum_k \frac{\left\| \hat{u}_k^{s+1} - \hat{u}_k^s \right\|_2^2}{\left\| \hat{u}_k^s \right\|_2^2} < \varepsilon \quad (10)$$

Among them, ε is the judgment accuracy ($\varepsilon > 0$).

3.1.2. Evaluation index of VMD. Define the reconstruction sequence $f'(i) = \sum_{k=1}^K u_k(i)$, which represents the sum of K sub modal signals after VMD decomposition. In an ideal scenario, the reconstructed sequence $f'(i)$ is identical to the original sequence $f(i)$, but there may be some decomposition loss during the actual decomposition process. This paper proposes a decomposition-quality evaluation index for power load forecasting, VMDLoss,

to address the problem of prediction accuracy degradation caused by mode mixing or over-decomposition in the Variational Mode Decomposition (VMD) process. Its essence is the mean of the sparsity of each sub modal signal after VMD decomposition, that is, the average envelope entropy. The specific calculation formula is as follows:

$$\begin{cases} \text{VMDLoss} = - \sum_{k=1}^K \sum_{i=1}^C \varepsilon_k(i) \lg \varepsilon_k(i) \\ \varepsilon_k(i) = \frac{b_k(i)}{\sum_{i=1}^C b_k(i)} \end{cases} \quad (11)$$

Among them, $b_k(i)$ represents the real part of the k -th sub modal analysis signal $B_k(i)$, $\varepsilon_k(i)$ represents the probability distribution sequence obtained by normalizing $b_k(i)$, and C is the number of sub modal signal samples. The smaller the VMDLoss, the smaller the decomposition loss, the more information the modality contains, and the more accurate the model.

3.2. xLSTM.

3.2.1. Structure of LSTM. Long Short-Term Memory (LSTM) is a special type of Recurrent Neural Network (RNN) designed to address the vanishing and exploding gradient problems during training with long sequences [28]. Its core structure consists of three gating mechanisms: the input gate (input gate), the forget gate (forget gate), and the output gate (output gate). These gating mechanisms regulate the flow of information using sigmoid functions and element-wise multiplication operations. The network architecture is illustrated in Figure 2.

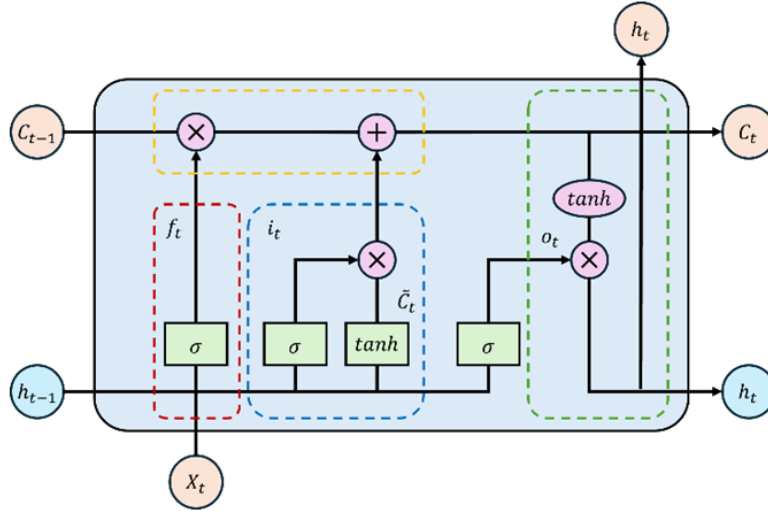


FIGURE 2. Network architecture diagram of LSTM

The forget gate is used to determine which information is discarded from the cellular state, and its formula is

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (12)$$

where σ is the sigmoid function, W_f and b_f are the weights and biases.

The input gate is used to determine which new information to store in the cell state, and its formula is

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (13)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (14)$$

Among them, $\tilde{C}_t$ is the candidate cell state. By combining the information from the forget gate and input gate, the cell state can be updated.

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \quad (15)$$

The output gate is used to determine which information to output, and its formula is

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (16)$$

$$h_t = o_t \cdot \tanh(C_t) \quad (17)$$

3.2.2. Structure of $xLSTM$. Traditional LSTM still has some limitations. It relies on sigmoid gating mechanisms, which restrict flexible information flow. Additionally, its memory cells store information as scalars, limiting their capacity to handle complex and large-scale data. Furthermore, due to the sequential dependency between time steps, LSTM struggles to achieve efficient parallel computation.

The core innovation of $xLSTM$ lies in its introduction of an exponential gating mechanism, which breaks through the output range restrictions of traditional gating functions [29]. This design enables higher flexibility and dynamic control over information flow. Unlike the conventional LSTM, which constrains gating values within $[0, 1]$ through sigmoid functions, $xLSTM$ reconstructs the gating system using exponential activation functions. This significantly expands the range of gating values and enhances the model's ability to regulate information flow, demonstrating unique advantages in modeling long-range dependencies and capturing high-complexity sequence patterns.

Moreover, $xLSTM$ introduces two novel memory units: $sLSTM$ and $mLSTM$. The $sLSTM$ preserves scalar storage while incorporating a novel multi-layer hybrid mechanism that enables interactions between memories across different layers. This improves both storage flexibility and information flow dynamics. On the other hand, $mLSTM$ employs matrix-based memory units, significantly increasing the model's storage capacity and effectively handling sparse information. The matrix memory units are updated via covariance update rules, significantly enhancing the model's performance in processing complex information. As shown in Figure 3, the architectures of $sLSTM$ and $mLSTM$ networks are illustrated.

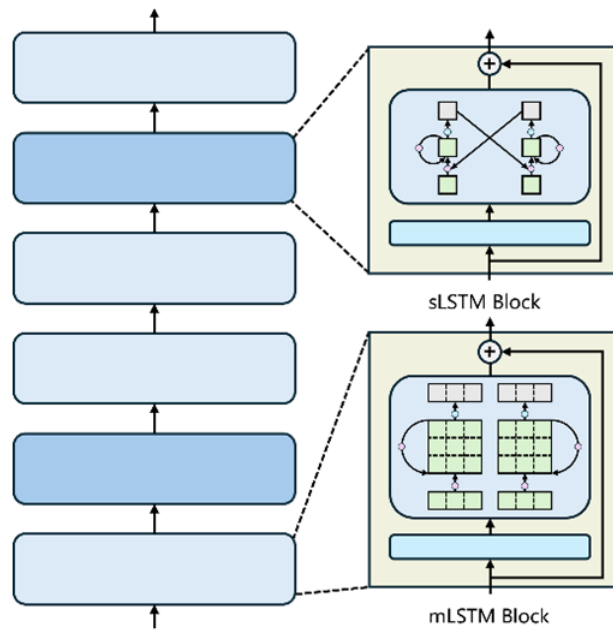


FIGURE 3. Network architecture diagram of $xLSTM$

xLSTM also introduces a residual stacking architecture similar to the Transformer, integrating sLSTM and mLSTM into residual blocks to form deep networks. This architecture not only enhances the model's scalability but also effectively prevents the vanishing gradient problem in deep networks, enabling the model to learn complex historical dependencies better. The residual stacking structure is illustrated in Figures 4 and 5.

Figure 3 depicts the overall architectural framework of the xLSTM network. It illustrates the hierarchical composition where sLSTM and mLSTM blocks are integrated into a deep residual structure. This macroscopic view shows how the input sequence is processed layer by layer, enabling the model to extract features at different levels of abstraction. The diagram highlights the residual connections that bypass the blocks, facilitating gradient flow and enabling the construction of deeper networks similar to Transformers.

Figure 4 illustrates the specific internal structure of the sLSTM module. It adopts a pre-LayerNorm residual structure to ensure training stability. The input features first pass through a normalization layer and are then processed by the sLSTM block, which retains the scalar memory mechanism but enhances it with exponential gating. A Group Norm

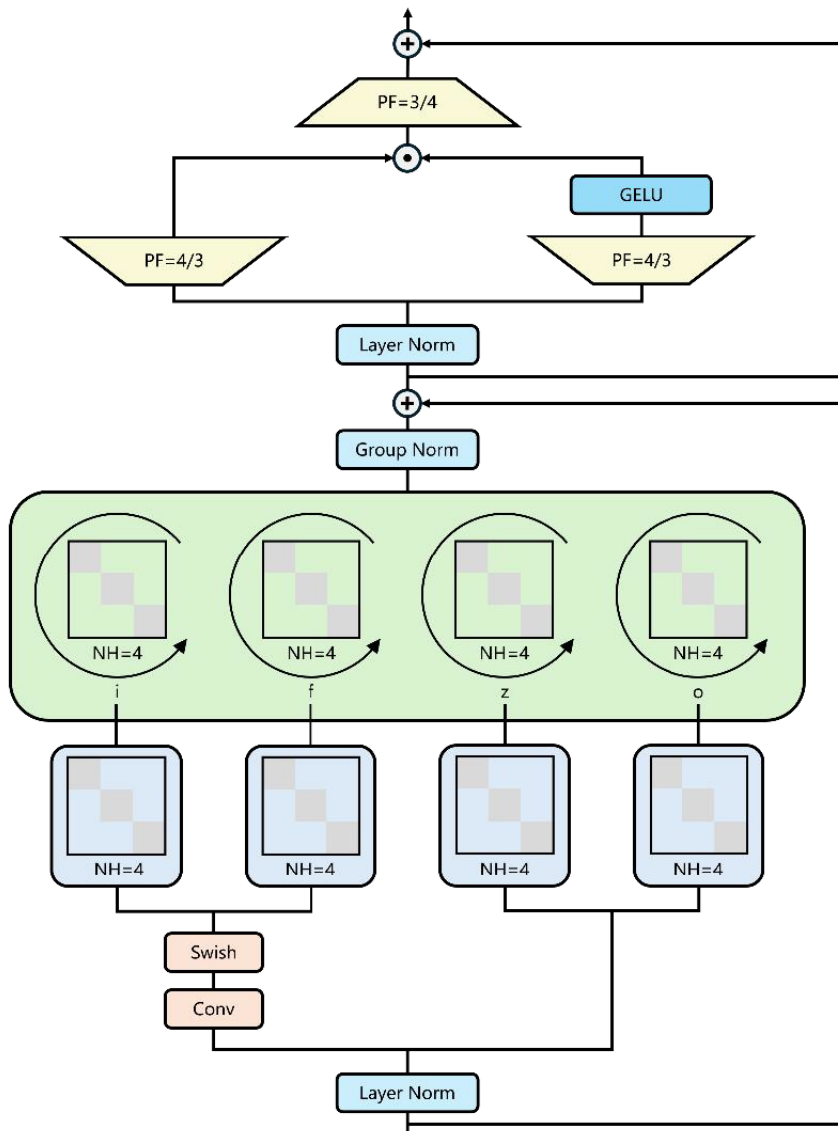


FIGURE 4. Structure diagram of sLSTM module

layer and a Swish activation function are subsequently applied to regulating the feature distribution, followed by a convolution operation to capture local dependencies. This design allows sLSTM to effectively manage scalar updates while maintaining gradient stability in deep networks.

The mLSTM module, as detailed in Figure 5, represents the matrix-based memory variant. Unlike the scalar memory, mLSTM projects the input into query (q), key (k), and value (v) vectors, similar to the attention mechanism. It uses a covariance update rule to store information in the matrix C_t , thereby significantly increasing the memory capacity for retrieving complex historical contexts. As shown in the diagram, the module also integrates Group Normalization and Swish activation within the residual block structure. This matrix memory architecture enables the model to parallelize computations more efficiently and capture long-range dependencies that are difficult for traditional scalar-based LSTMs to retain.

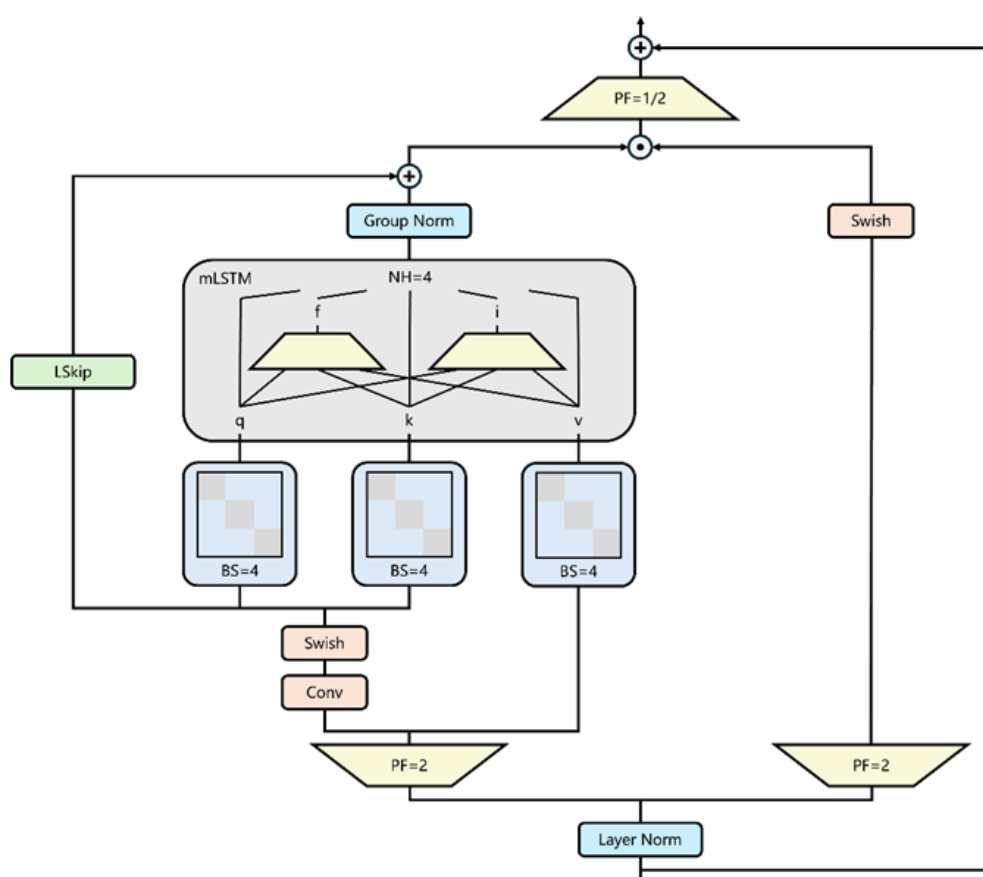


FIGURE 5. Structure diagram of mLSTM module

4. Case Analysis and Comparison of Experimental Results.

4.1. Data sources and evaluation indicators. To comprehensively evaluate the model’s performance and generalization ability, this study employs two distinct datasets. The primary dataset is from a specific region of New Toronto, data including six variables: load, temperature, humidity, rainfall, wind speed, and electricity price. The time period selected is from March 15, 2012, to June 8, 2013, with a sampling interval of 15 minutes, resulting in a total of 259,200 load data points. Additionally, to validate the model’s applicability to recent scenarios and different geographical environments, a secondary dataset

collected from a thermal power plant in Northeast China (covering the year 2023) is introduced. The specific analysis and validation based on this recent dataset are detailed in Section 4.3.3.

To eliminate the impact of dimensionality differences across factors, the input data for the prediction model are normalized to the range $[0, 1]$. The entire dataset, spanning from March 15, 2012, to June 8, 2013, is strictly divided into a training set and a testing set following a 6 : 1 ratio based on chronological order. This ensures temporal continuity and prevents information leakage. Specifically, the data from the earlier period constitutes the training set, used for model learning and parameter optimization, while the subsequent data forms the test set to evaluate the model's generalization performance.

Sample construction is performed using a sliding window approach across the whole timeline. For each prediction target, load data from the preceding 6 hours is extracted and combined with the fused feature data calculated using Equation (6). This process generates a comprehensive sequence of samples covering the entire study period, ensuring that the model is trained on diverse seasonal and fluctuation patterns.

4.2. Selection of load and feature fusion dataset. Set the window length to 6, and use a sliding time window to traverse the historical daily load sequence. For each time segment, calculate the similarity using a combined algorithm that combines curve trend degree and difference degree. Taking February 20, 2013, at 12:00 as an example, select the top 7 optimal similar sequences as load data for the prediction model, as shown in Table 2. Collect humidity, temperature, and electricity price data for February 20, 2013, using weather forecasts and other means. Combine these data with the load data from Table 2 to form a load-feature fusion dataset, which is then input into the prediction model.

TABLE 2. Selection of optimal similar day

	Time	Similar coefficient
1	2012-07-06, 12:15-18:15	0.982
2	2012-09-12, 12:00-18:00	0.979
3	2013-03-15, 10:30-16:30	0.973
4	2012-06-02, 13:00-19:00	0.971
5	2012-08-27, 11:45-17:45	0.965
6	2013-04-07, 12:30-18:30	0.962
7	2012-07-18, 13:15-19:15	0.958

4.3. Case study.

4.3.1. Comparative analysis of instances from different types of datasets. To validate the superiority of the load and feature-integrated dataset, this study conducted comparative experiments using two datasets. Type I dataset utilized historical load sequences from the six hours preceding the prediction time as input data. In contrast, Type II dataset used the load and feature-integrated dataset constructed in Section 4.2 as input. To ensure the experiments were rational, the VMD-xLSTM model was consistently used as the prediction model for both datasets. Comparing the experimental data horizontally, it was evident that Type II dataset demonstrated significant advantages in predictive performance. As shown in Figure 6's prediction curves and Table 3's evaluation metric comparison, Type II dataset achieved improvements across all three evaluation indicators compared to Type I baseline dataset. Specifically, MAPE decreased by 0.71 percentage points, while RMSE and MAE were optimized to 57.28 and 53.03, respectively, resulting

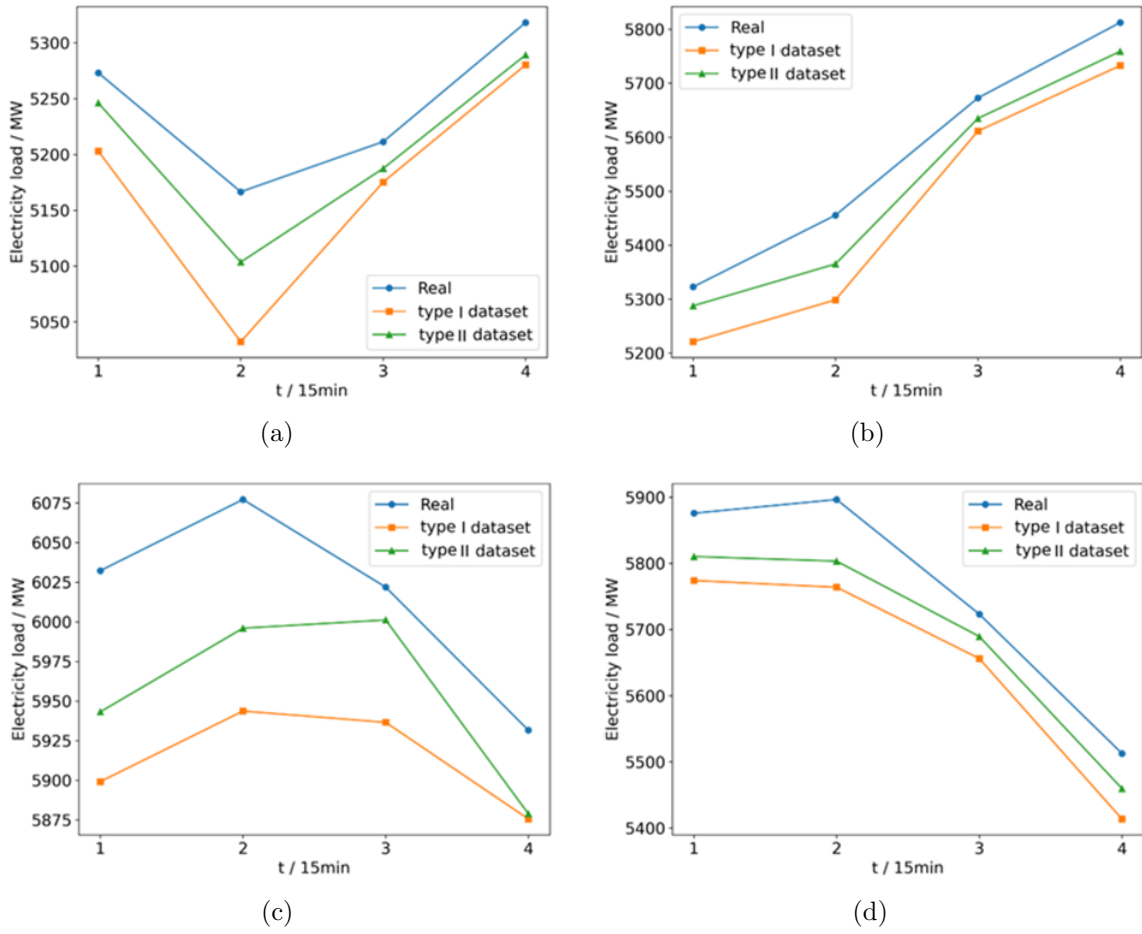


FIGURE 6. Prediction result curves for various types of datasets

TABLE 3. Comparison of prediction result of each data set

Dataset	RMSE	MAE	MAPE
Type I dataset	99.04	92.88	1.64%
Type II dataset	57.28	53.03	0.93%

in better prediction accuracy. The enhanced performance of Type II dataset can be attributed to their integration of multidimensional data features, which effectively strengthens the model's ability to uncover latent characteristics.

4.3.2. Comparative analysis of examples from different models. To comprehensively validate the performance of the VMD-xLSTM prediction model, this study uniformly employs a load-feature fused dataset as model input for comparison with BiLSTM, xLSTM, and VMD-xLSTM. In the comparative experiments, the configuration of the neural networks' hidden layers is maintained consistently, specifically with 32 hidden layer nodes, a maximum iteration limit of 120, an initial learning rate of 0.001, and the tanh function uniformly adopted as the activation function. The visual comparison of prediction results is shown in Figure 7, and the evaluation metric comparison is presented in Table 4.

As shown in Figure 7 and Table 4, the VMD-xLSTM model effectively decomposes the original nonlinear sequence into multiple simple modal components, thereby significantly reducing data complexity. Compared to both BiLSTM and xLSTM models, VMD-xLSTM demonstrates higher prediction accuracy, strongly validating its algorithmic effectiveness

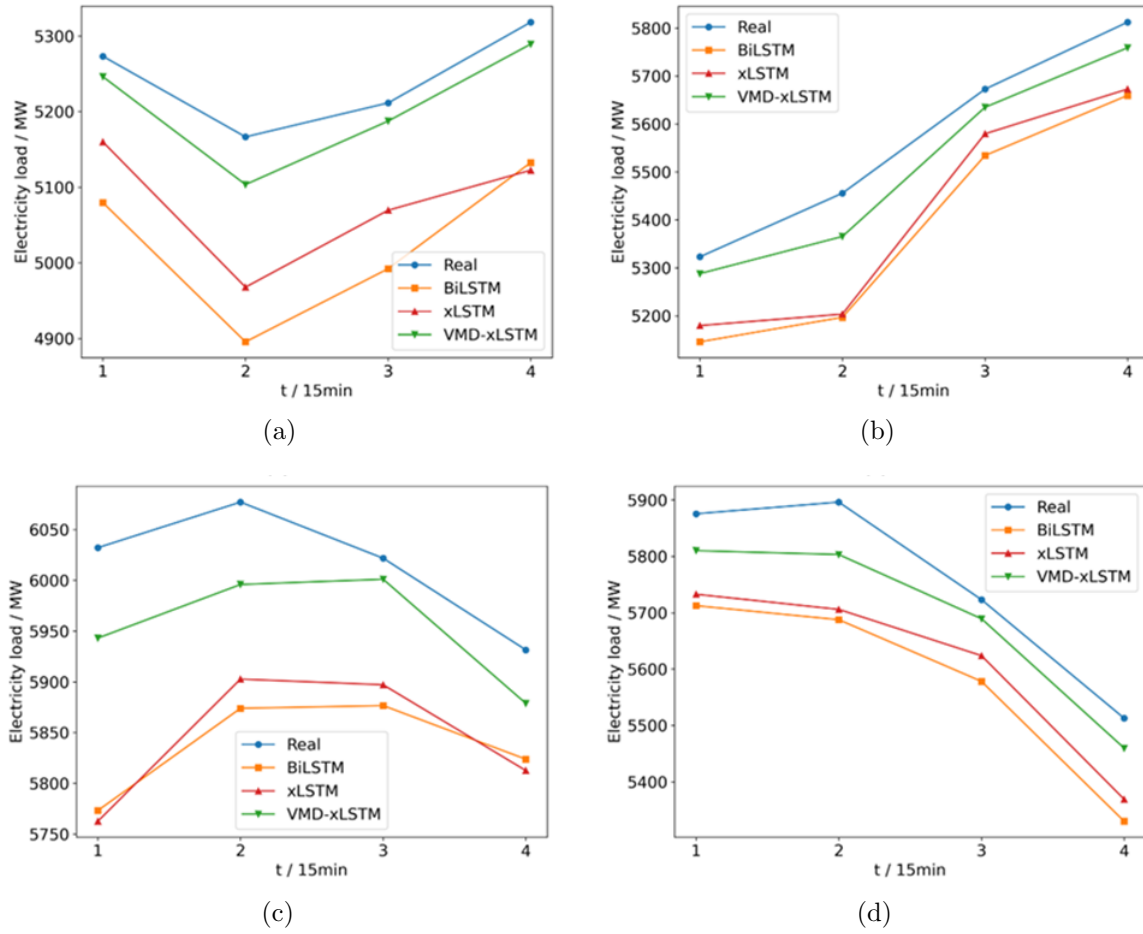


FIGURE 7. Prediction result curves of different methods

TABLE 4. Comparison of prediction result of each prediction model

Different models	RMSE	MAE	MAPE
BiLSTM	192.97	188.21	3.36%
xLSTM	165.95	158.86	2.82%
VMD-xLSTM	57.28	53.03	0.93%

and advancement. Furthermore, compared to the BiLSTM model, xLSTM achieves a 0.54% reduction in MAPE, with RMSE and MAE decreasing to 165.95 and 158.86, respectively, indicating that xLSTM possesses stronger learning capabilities and higher prediction accuracy than BiLSTM. When combined with VMD decomposition technology to form the VMD-xLSTM model, its performance is further enhanced: compared to the undecomposed xLSTM model, VMD-xLSTM sees RMSE and MAE drop significantly to 57.28 and 53.03, respectively, with MAPE also decreasing by 1.89%. This substantial performance improvement not only highlights the necessity and effectiveness of VMD decomposition technology but also further confirms the superiority of the VMD-xLSTM hybrid model in enhancing prediction accuracy.

4.3.3. Generalization verification using recent power plant data. To address the potential concern regarding data timeliness and to further verify the generalization ability of the proposed framework, this study introduces a second dataset for validation. The data is collected from a thermal power plant in Northeast China, covering the period from

January 1, 2024 to December 31, 2024. Compared to the New Toronto dataset, this dataset represents a more recent time series with distinct load characteristics influenced by different climatic and industrial factors. To ensure a rigorous evaluation, we conducted comparative experiments using BiLSTM, xLSTM, and the proposed VMD-xLSTM on this new dataset.

The prediction results of the different models are visualized in Figure 8. As observed, the prediction curves of the benchmark models BiLSTM and xLSTM exhibit noticeable deviations, in contrast, even with the change in data source and time span, the VMD-xLSTM model closely tracks the actual load curve, capturing local variations more accurately than the single models.

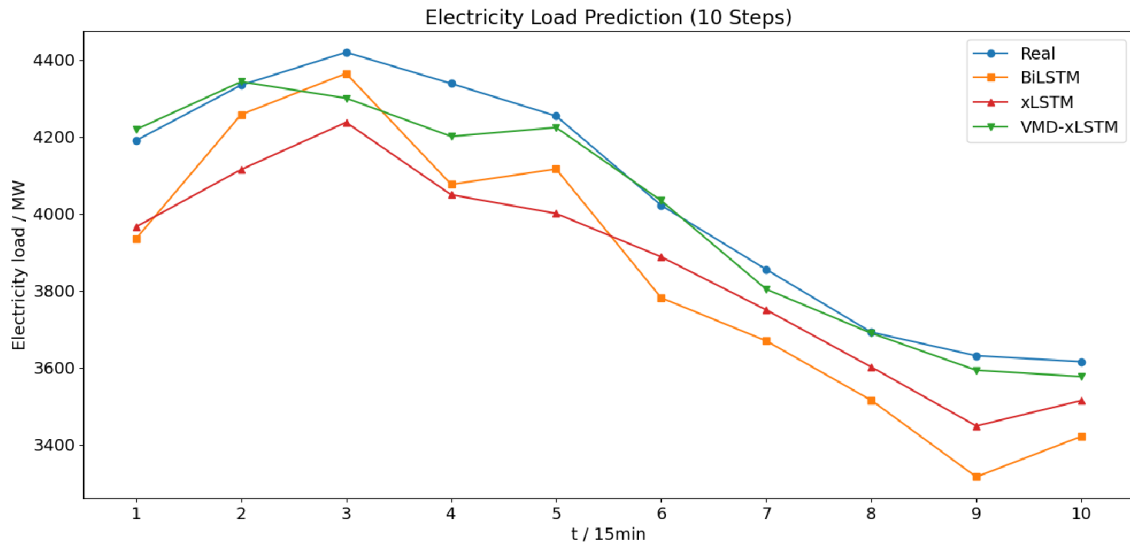


FIGURE 8. Prediction results of the power load in a power plant in North-east China

This indicates that the proposed symmetrical framework is not overfitting to a specific historical dataset.

Table 5 presents the quantitative evaluation metrics for the three models on the North-east China dataset. It is evident that the proposed VMD-xLSTM model maintains a significant advantage over the benchmark models, demonstrating superior robustness.

TABLE 5. Comparison of prediction performance on Northeast China dataset

Different models	RMSE	MAE	MAPE
BiLSTM	205.4	189.8	4.80%
xLSTM	189.7	178.1	4.35%
VMD-xLSTM	63.7	46.6	1.13%

Compared to the BiLSTM model: The VMD-xLSTM framework achieved a substantial reduction in error metrics. The MAPE decreased by 3.67 percentage points (from 4.80% to 1.13%), while the RMSE and MAE were reduced by 141.7 and 143.2, respectively. This indicates that the proposed model handles the complex volatility of the new dataset much better than the traditional bidirectional recurrent network. Compared to the single xLSTM model: It is noteworthy that the standalone xLSTM model already outperformed the BiLSTM (with an MAPE of 4.35% versus 4.80%), proving its effectiveness in capturing temporal dependencies. However, the hybrid VMD-xLSTM framework further elevated

the prediction accuracy. The integration of VMD decomposition reduced the MAPE by another 3.22 percentage points compared to the single xLSTM. This strongly validates the necessity of the decomposition strategy: even with a powerful base model like xLSTM, decomposing the non-stationary load into stationary components is essential for achieving state-of-the-art forecasting precision.

These results demonstrate that the superiority of the VMD-xLSTM framework is not limited to a specific historical dataset. Even when applied to a recent, geographically distinct dataset, the proposed method consistently outperforms traditional and single-structure deep learning models, proving its strong robustness and practical value for modern power grid forecasting.

5. Conclusions. This paper focuses on short-term power load forecasting by integrating similar sequence selection with the xLSTM model. Through comprehensive example calculations and error analysis, the following key conclusions are drawn.

1) Given the significant regularity characteristics of power load data, this paper optimizes the strategy for selecting similar sequences by combining curve trend degree and difference degree algorithms. This approach significantly improves the quality and applicability of similar sequences.

2) To fully capture the temporal and feature information in power data, a load-feature fused dataset is constructed as the input for the prediction model. This effectively avoids information loss caused by single data type inputs, providing the model with richer and more comprehensive data.

3) Considering the inherent nonlinear temporal characteristics of power load data, this paper designs and implements the VMD-xLSTM hybrid prediction model. By decomposing complex power load sequences into multiple simpler sub-modal sequences using VMD and predicting each, this strategy not only reduces prediction difficulty but also effectively minimizes model bias, thereby significantly improving prediction accuracy.

4) Taking the load dataset from New Toronto and a distinct power plant in Northeast China as examples, this paper proposes a prediction method based on the combination of similar sequences and the xLSTM model. The comparative experiments on both historical and recent datasets demonstrate higher prediction accuracy, fully validating the effectiveness and practicality of the proposed method. These empirical results provide new research perspectives and solutions for the field of power load forecasting.

5) Limitations and future work. Although the proposed VMD-xLSTM model demonstrates superior performance, there are still areas for improvement. Currently, the model's hyperparameter optimization relies partially on manual tuning, which may not ensure the global optimum in varying scenarios. Future research will focus on integrating intelligent optimization algorithms (such as Particle Swarm Optimization or the Whale Optimization Algorithm) to select optimal VMD and xLSTM parameters automatically. Additionally, we plan to incorporate external factors, such as holidays, social events, and extreme weather alerts, to enhance further the model's robustness and adaptability to complex real-world power grids.

Acknowledgement. This paper is supported by CGN technology project (020-GN-B-2023-C45-P.0.99-01291).

REFERENCES

- [1] Y. Eren and İ. Küçükdemiral, A comprehensive review on deep learning approaches for short-term load forecasting, *Renewable and Sustainable Energy Reviews*, vol.189, 114031, 2024.
- [2] M. G. Pinheiro et al., Short-term electricity load forecasting – A systematic approach from system level to secondary substations, *Applied Energy*, vol.332, 120493, 2023.

- [3] Q. Dong, R. Huang, C. Cui, D. Towey, L. Zhou, J. Tian and J. Wang, Short-term electricity-load forecasting by deep learning: A comprehensive survey, *Engineering Applications of Artificial Intelligence*, vol.154, 110980, 2025.
- [4] K. Lang, M. Zhang, Y. Yuan and X. Yue, Short-term load forecasting based on multivariate time series prediction and weighted neural network with random weights and kernels, *Cluster Computing*, vol.22, no.5, pp.12589-12597, 2019.
- [5] S. Motepe, A. N. Hasan and R. Stopforth, Improving load forecasting process for a power distribution network using hybrid AI and deep learning algorithms, *IEEE Access*, vol.7, pp.82584-82598, 2019.
- [6] L. Alhmoud and Q. Nawafleh, Short-term load forecasting for Jordan power system based on NARX-ELMAN neural network and ARMA model, *IEEE Canadian Journal of Electrical and Computer Engineering*, vol.44, no.3, pp.356-363, 2021.
- [7] B. Bischl et al., Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges, *WIREs Data Mining and Knowledge Discovery*, vol.13, no.2, e1484, 2023.
- [8] H. Hewamalage et al., Forecast evaluation for data scientists: Common pitfalls and best practices, *Data Mining and Knowledge Discovery*, vol.37, pp.578-616, 2023.
- [9] K. Sharifani and M. Amini, Machine learning and deep learning: A review of methods and applications, *World Information Technology and Engineering Journal*, vol.10, no.7, pp.3897-3904, 2023.
- [10] A. Roy and S. Chakraborty, Support vector machine in structural reliability analysis: A review, *Reliability Engineering & System Safety*, vol.233, 2023.
- [11] Q. Wang, Support vector machine algorithm in machine learning, *2022 IEEE International Conference on Artificial Intelligence and Computer Applications*, pp.750-756, 2022.
- [12] S. Li and H. Cai, Short-term power load forecasting using a VMD-Crossformer model, *Energies*, vol.17, no.11, 2773, 2024.
- [13] W. Lu and X. Chen, Short-term load forecasting for power systems with high-penetration renewables based on multivariate data slicing transformer neural network, *Frontiers in Energy Research*, vol.12, 1355222, 2024.
- [14] N. Wang and Z. Li, Short-term power load forecasting based on BES-VMD and CNN-Bi-LSTM method with error correction, *Frontiers in Energy Research*, vol.10, 1064972, 2023.
- [15] M.-S. Tsai and F.-Y. Hsu, Application of grey correlation analysis in evolutionary programming for distribution system feeder reconfiguration, *IEEE Transactions on Power Systems*, vol.25, no.2, pp.1126-1133, 2010.
- [16] A. Jahani, K. Zare and L. M. Khanli, Short-term load forecasting for microgrid energy management system using hybrid SPM-LSTM, *Sustainable Cities and Society*, vol.98, 104775, 2023.
- [17] J. E. Pesantez et al., A comparison study of predictive models for electricity demand in a diverse urban environment, *Energy*, vol.283, 129142, 2023.
- [18] Y. Wang, Y. Hao, B. Zhang and N. Zhang, Short-term power load forecasting using SSA-CNN-LSTM method, *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, vol.46, no.1, 2343297, 2024.
- [19] T. A. Alghamdi and N. Javaid, A survey of preprocessing methods used for analysis of big data originated from smart grids, *IEEE Access*, vol.10, pp.29149-29171, 2022.
- [20] Z. Jiang, L. Zhang and T. Ji, NSDAR: A neural network-based model for similar day screening and electric load forecasting, *Applied Energy*, vol.349, 2023.
- [21] X. Liao, N. Cao, M. Li et al., Research on short-term load forecasting using XGBoost based on similar days, *2019 International Conference on Intelligent Transportation, Big Data & Smart City*, pp.675-678, 2019.
- [22] S. Arora and J. W. Taylor, Short-term forecasting of anomalous load using rule-based triple seasonal methods, *IEEE Transactions on Power Systems*, vol.28, no.3, pp.3235-3242, 2013.
- [23] E. A. Madrid and N. Antonio, Short-term electricity load forecasting with machine learning, *Information*, vol.12, 2021.
- [24] J. W. Taylor and R. Buizza, Neural network load forecasting with weather ensemble predictions, *IEEE Transactions on Power Systems*, vol.17, no.3, pp.626-632, 2002.
- [25] Y. Zhang, X. Yang, C. Fang et al., Short-term building load forecasting based on similar day selection and LSTM network, *2018 2nd IEEE Conference on Energy Internet and Energy System Integration*, pp.1-5, 2018.
- [26] D. Kim, D. Lee, H. Nam et al., Short-term load forecasting for commercial building using convolutional neural network (CNN) and long short-term memory (LSTM) network with similar day selection model, *Journal of Electrical Engineering & Technology*, vol.18, pp.4001-4009, 2023.

- [27] K. Dragomiretskiy and D. Zosso, Variational mode decomposition, *IEEE Transactions on Signal Processing*, vol.62, no.3, pp.531-544, 2014.
- [28] S. Hochreiter and J. Schmidhuber, Long short-term memory, *Neural Computation*, vol.9, no.8, pp.1735-1780, 1997.
- [29] M. Beck, K. Poppel, M. Spanring et al., xLSTM: Extended long short-term memory, *Advances in Neural Information Processing Systems*, vol.37, pp.107547-107603, 2025.

Author Biography



Jianfan Zhai received the B.S. degree in Electrical Engineering and Automation from Zhengzhou University, China, in 2007. He is currently a Senior Engineer. He works in CGN Wind Power Co., Ltd., China. His research interests include power system relay protection.



Lei Wu received the B.S. degree in Electrical Engineering and Automation from Tianjin University, China, in 2011. He works in CGN Wind Power Co., Ltd., China. His research interests include lean operation and maintenance of renewable energy power stations, and transmission and distribution relay protection.



Huaming Zhang received the B.S. degree in Mechanical Engineering and Automation plus Software Engineering from Dalian Jiaotong University, China, in 2011, and the M.S. degree in Mechanical Engineering from Beijing University of Technology, China, in 2019. He is currently a Ph.D. candidate at Beijing University of Technology and serves as the Technical Support Business Manager in the Operation and Maintenance Division of CGN Wind Power Co., Ltd., China. His research interests include intelligent operation and maintenance technology for wind turbines, structural health monitoring methods, and the coupling mechanism between wind resource characteristics and equipment safety.



Tie Zhong received the B.S., M.S., and Ph.D. degrees in Communication Engineering from Jilin University, China, in 2006, 2008 and 2016, respectively. He is currently an Associate Professor in the Department of Communication Engineering, Northeast Electric Power University, China. His research interests include electric signal processing techniques, deep learning, and intelligent information processing.