Textile & Leather Review

ISSN 2623-6281| www.tlr-journal.com | 10.31881/TLR

Short-Term Power Load Forecasting Algorithm Based on Hybrid Transformer-LSTM and Attention Mechanism

Minjing Yang, Binghong Su, Qinwei Duan, Yashan Zhong, Jiaxin Zhuo, Xuanli Lan

How to cite: Yang M, Su B, Duan Q, Zhong Y, Zhuo J, Lan X. Short-Term Power Load Forecasting Algorithm Based on Hybrid Transformer-LSTM and Attention Mechanism. Textile & Leather Review. 2026; 9:6068-6094. https://doi.org/10.31881/TLR.2026.6068

How to link: https://doi.org/10.31881/TLR.2026.6068

Published: 15 June 2026

Short-Term Power Load Forecasting Algorithm Based on Hybrid Transformer-LSTM and Attention Mechanism

Minjing Yang[1], Binghong Su[1], Qinwei Duan[1], Yashan Zhong[1], Jiaxin Zhuo[2*], Xuanli Lan[2]

[1]Power Dispatch and Control Center, Guangdong Power Grid Co., Ltd, China Southern Power Grid, Guangzhou 243000, Guangdong, China

[2]Beijing TsIntergy Technology Co., Ltd, Beijing 100084, China

*zhuojx960408@126.com

Article

https://doi.org/10.31881/TLR.2026.6068

Published 15 June 2026

ABSTRACT

For short-term forecasting in the textile industry, a single model is inherently limited because it struggles to balance local temporal dynamics (such as rapidly changing seasonal styles) with global dependencies (such as fluctuations in raw material costs), and often fails to capture the time-varying characteristics of feature importance. This paper proposes a hybrid Transformer-LSTM architecture incorporating a multi-head temporal-feature attention mechanism. The input sequence is segmented into sliding windows, with features and variables uniformly encoded via an embedding layer. The Transformer's multi-head self-attention models global temporal dependencies and long-term patterns. Its output feeds into the LSTM, which uses a gating structure to capture local dynamic changes and enhance sequence evolution modeling. Finally, an attention mechanism on the bidirectional LSTM's hidden state adaptively weights key time steps to generate a context-aware feature vector. Finally, the attention output and Transformer features are concatenated, and a fully connected layer performs regression to predict the load value, achieving multi-scale feature optimization. Experiments confirm this method's significant advantage in short-term forecasting, showing an MAE of (0.75±0.04) MW and RMSE of (0.90±0.06) MW. In sudden change scenarios, the mean MAPE is ≤5.12% and mean R2 is ≥0.928, effectively capturing dynamic correlations under social temporal changes. This study tackles the complexity of the textile industry by explicitly accounting for both local temporal dynamics, like fast-fashion cycles, and global dependencies, such as supply chain risks, alongside the dynamic changes in feature importance.

KEYWORDS

short-term power load forecasting, hybrid architecture, long short-term memory, attention mechanism, textile industry

INTRODUCTION

Accurate short-term demand forecasting is the core technology for ensuring the efficient operation of the textile industry's supply chain and optimizing production scheduling [1,2]. While not altering the focus on power systems, the modern textile industry needs to draw parallels, as these systems face the dual challenges of high-proportion renewable energy access and increased fluctuations in electricity demand, placing higher demands on the accuracy and timeliness of forecasting models [3]. A crucial technical path to promote the intelligent development of power systems remains building a forecasting framework with strong time series analysis capabilities and context-aware characteristics; this is analogous to the need for advanced systems in the textile industry to manage complex supply chains.

Current power load forecasting models still face a deep modeling bottleneck in handling the interaction between multi-scale temporal structures and dynamic features. Load sequences contain both long-term, daily, and weekly cycles and sudden local fluctuations. A single network structure struggles to achieve a balanced representation of global dependencies and fine-grained dynamics within a unified framework [4,5]. While Transformer-type models can effectively capture long-range temporal correlations, their sensitivity to local temporal variations is limited, and they are particularly susceptible to response lags in scenarios with sudden load increases or decreases [6,7]. Conversely, sequence models such as LSTM excel at tracking instantaneous evolutionary trends, but due to the inherent limitations of their recursive mechanisms, they are prone to information decay in long sequences, weakening their ability to robustly extract periodic patterns [8,9]. The overall framework design lacks a collaborative optimization path between the feature extraction, time series modeling, and context selection modules, resulting in a linear flow of information and suppressing the potential for higher-level interactions. These inherent limitations restrict the model's ability to achieve further breakthroughs in high-precision, robust prediction tasks.

This paper focuses on building a short-term power load forecasting model that can collaboratively capture global dependencies and local dynamics while also enabling adaptive adjustment of feature importance. The core goal is to overcome the scalability limitations of single-structure models in time series modeling, proposing a hierarchical and progressive hybrid neural network framework. This framework uses a front-end Transformer module to model long-range temporal dependencies, fully exploiting the periodic and trend patterns in load series. Its output serves as the input to a bidirectional LSTM, which utilizes gated units to

finely characterize local variations and transient responses, enhancing its adaptability to nonstationary fluctuations. Building on this foundation, it innovatively designs a multi-head joint time-feature attention mechanism that operates on the LSTM hidden state sequence, simultaneously identifying key time steps and assigning dynamic feature weights to generate a context-sensitive, comprehensive representation. This mechanism concurrently learns time-feature interactions across multiple subspaces, significantly improving the model's ability to discern complex influencing factors. Ultimately, by concatenating the attention outputs with high-level Transformer semantic features and performing load regression prediction via a fully connected layer, it achieves deep fusion of multi-scale information. This paper contributes in three key areas: First, it constructs a cascaded Transformer and bidirectional LSTM architecture to achieve hierarchical extraction of global and local temporal features; second, it proposes a multi-head time-feature joint attention module, breaking through the modeling bottleneck of traditional single-dimensional attention; and third, it verifies the effectiveness of the dynamic feature perception mechanism on real-world load data, providing a scalable technical path for high-precision forecasting. This method remains robust under sudden load changes and extreme weather conditions, demonstrating strong practical potential, a critical quality that also translates to managing supply chain disruptions and volatile demand within the textile industry.

RELATED WORK

In terms of load forecasting modeling, researchers have explored a variety of neural network architectures to improve forecasting performance. Previous studies extensively utilized recurrent neural networks and their variants (e.g., long short-term memory networks, LSTM) for load forecasting [10,11], which use gating mechanisms to capture short-term dynamic changes in sequences and show strong capabilities in processing local temporal dependencies. Xia M proposed a method based on an improved stacked gated recurrent unit (GRU) recurrent neural network. The method selects input variables through correlation analysis and uses adaptive gradient and adjustable momentum to optimize the training process, effectively improving the accuracy and robustness of renewable energy generation and power load forecasting in single-variable and multi-variable scenarios [12]. For multi-scale time series modeling and dynamic feature selection, the combination of hybrid architectures and attention mechanisms has shown potential advantages. Previous studies have shown that cascading a global modeling module with a local sequence processor can enhance

feature representation. Some studies have used CNNs to extract periodic features and then connected them to LSTMs for sequence evolution modeling, initially achieving the fusion of coarse-grained and fine-grained information [13,14]. Notably, while pure Transformer models can enhance local modeling capabilities through stacking layers or increasing attention heads, their self-attention mechanism remains limited in capturing high-frequency transient changes due to its tendency to average globally within fixed windows. Although deep-tuned LSTM/GRU networks can capture local dynamics, they struggle to effectively model cross-day or even cross-week periodic dependencies. The hybrid architecture proposed in this paper avoids simple stacking by implementing structural specialization: Transformers are dedicated to extracting cross-periodic semantic patterns, while LSTMs focus on refining local non-stationary fluctuations. This complementary information flow between the two components achieves substantial improvements in multi-scale modeling capabilities without significantly increasing computational complexity.

CONSTRUCTION OF THE HYBRID TRANSFORMER-LSTM MODEL

Figure 1 shows the architecture of a short-term power load forecasting model based on a hybrid Transformer-LSTM and attention mechanism. After the input sequence is uniformly encoded through the embedding layer, it is first input to the Transformer module to model long-term global dependencies; its output is then used as the input to the bidirectional LSTM network to capture local temporal dynamics. The LSTM hidden state generates a context vector through a multi-head time-feature joint attention mechanism. This vector is then concatenated and fused with the high-level features to achieve multi-scale representation synergy. Finally, the predicted value is regressed and output through a fully connected layer. The overall structure emphasizes the parallelism of global and local modeling, the hierarchical integration of multi-source information, and the adaptive learning of dynamic weights. It demonstrates the ability to deeply analyze the complex spatiotemporal characteristics of load, and possesses high accuracy and strong robustness.

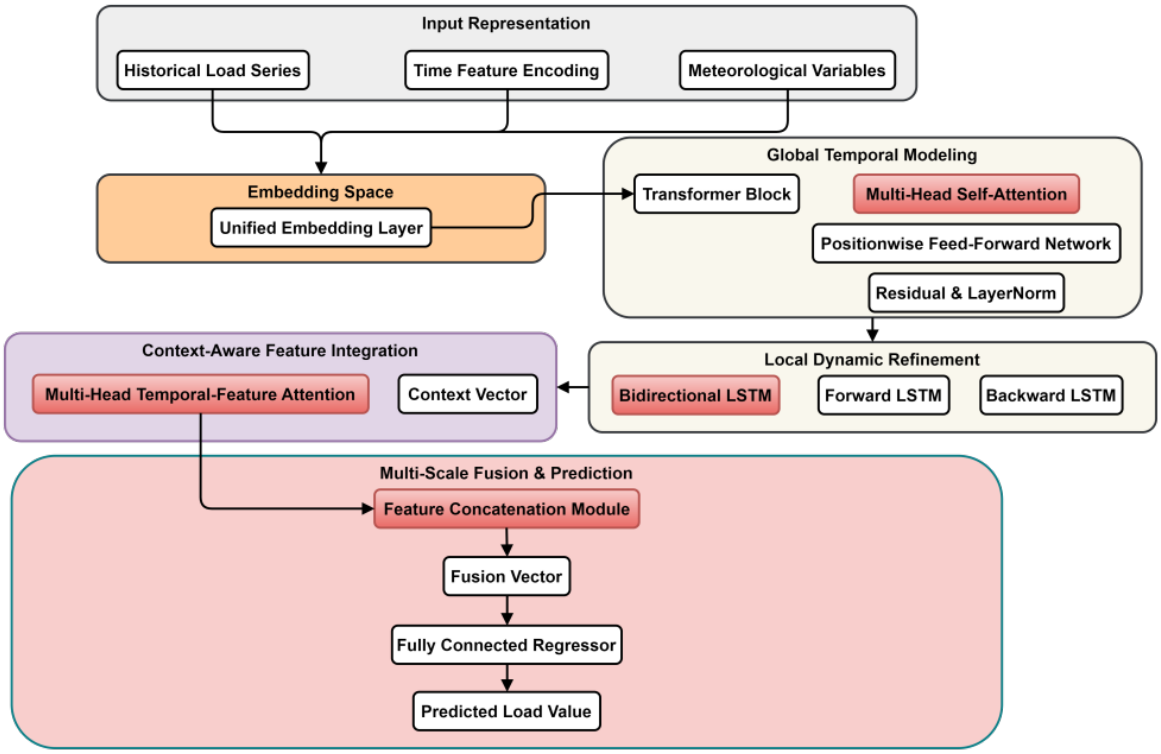

Figure 1. Short-term power load forecasting model architecture

Multi-Scale Feature Extraction of Input Sequences

Construction of Multidimensional Input Sequence and Sliding Window Organization

To fully explore the temporal evolution patterns and dynamic influences of external driving factors in short-term power load series, the raw load data are first reconstructed into a structured multidimensional input sequence. Based on the continuity characteristics of time series, a fixed-length sliding window is used to sample the historical load series in segments. The window length is set based on the typical cyclical characteristics of load changes, covering both intraday and interday rhythms. The data within each sliding window constitutes an independent sample, containing hourly (or half-hourly) historical load values within that period, which serves as the basic input for the model to perceive the sequence dynamics.

Embedding-Driven Unified Vector Coding Mechanism

To achieve effective fusion and semantic alignment of heterogeneous variables, a learnable embedding layer is designed to map multidimensional inputs to a unified dimension. The embedding layer contains a learnable

linear projection and a fixed sine position code to preserve the sequential information of time steps. Because the original input contains continuous variables (such as load and temperature) and discrete categorical variables (such as hours and weeks), direct concatenation can lead to semantic space misalignment and unbalanced gradient propagation. After all variables are transformed by their respective paths, they are concatenated along the feature dimension and nonlinearly fused through a shared fully connected layer to generate the final input embedding sequence. This process can be formally expressed as:

$$\mathbf{E}_t = \mathrm{FC}\left(\mathrm{Concat}\left(\mathbf{e}_{\mathrm{load},t}, \mathbf{e}_{\mathrm{time},t}, \mathbf{e}_{\mathrm{weather},t}\right)\right) \quad (1)$$

$\mathbf{E}_t \in \mathbb{R}^{d_{\mathrm{model}}}$ represents the joint embedding vector at time t; $\mathbf{e}_{\mathrm{load},t}$, $\mathbf{e}_{\mathrm{time},t}$, and $\mathbf{e}_{\mathrm{weather},t}$ are the preliminary encodings of load, time, and meteorological variables, respectively; FC represents the fully connected transformation operation; and d_{model} is the unified latent dimension set by the model. This embedding mechanism not only achieves semantic alignment of multi-source information but also, through parameterized learning, endows each variable with adjustable representation capabilities in vector space, making the model sensitive to key features from the outset. The embedded sequence serves as input to subsequent modules, ensuring that global and local modeling processes unfold within a unified semantic space and laying the structural foundation for multi-scale feature extraction.

Global Dependency Modeling in Transformer Modules

Long-Range Dependency Modeling with Multi-Head Self-Attention

To accurately capture the global cross-time correlation characteristics of power load sequences, a multi-head self-attention structure based on scaled dot products is used to perform deep relationship mining on the input embedding sequence. The input sequence is linearly projected to generate three matrices: query (Q), key (K), and value (V), representing the information retrieval pattern, matching criteria, and content carrier, respectively. Each matrix is independently similarized across multiple parallel attention heads. Dot product operations are used to measure the dynamic correlation strength between any two time steps, identifying long-range time series segments with periodic responses or trend consistency. To prevent the gradient saturation of the dot product resulting in high-dimensional space, a scaling factor is applied to normalize the

inner product result to ensure the stability of the attention weight distribution. After the attention weight is normalized by the Softmax function, it is applied to the value matrix to achieve weighted aggregation and generate a context-enhanced representation for each time step. This process is executed in parallel in multiple subspaces, enabling the model to simultaneously capture multi-modal dependencies in load fluctuations from different representation subspaces, such as complex interactions, such as similar load patterns on weekdays and lagged temperature response paths. The outputs of the multi-head mechanism are concatenated and restored to the original dimension through linear transformation to form the final self-attention output. This process can be expressed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \tag{2}$$

T is the sequence length, and d_k is the dimension of the key vector for each attention head. This mechanism overcomes the attenuation bottleneck of traditional recursive structures in modeling long-distance dependencies, enabling the model to directly establish non-local connections between the current moment and key historical cycle points, significantly improving the ability to identify long-term patterns in load sequences, such as those spanning days and weeks.

Intra-Layer Structure Optimization and Stable Transfer of Semantic Features

After completing the multi-head self-attention operation, a feedforward neural network and structured normalization components are applied to form a complete Transformer encoding unit to further enhance the nonlinear representation of features and ensure the training convergence of deep networks. The feedforward network consists of two fully connected layers, with the intermediate activation function using a Gaussian error linear unit (GELU). The residual structure is followed by layer normalization, which standardizes the mean and variance of all feature dimensions at each time step, eliminates internal covariate shifts, and improves the stability of the training process. Its mathematical expression is:

$$\mathbf{Z} = \text{LayerNorm}\big(X + \text{Attention}(X)\big) \tag{3}$$

$$\mathbf{O} = \text{LayerNorm}\big(\mathbf{Z} + \text{FFN}(\mathbf{Z})\big) \tag{4}$$

$\mathbf{Z}$ is the output of the attention sublayer; $\mathbf{O}$ is the final output; and FFN represents the feedforward network mapping. Through the synergistic effect of this hierarchical structure, the Transformer module achieves efficient modeling of complex global dependencies in load sequences while maintaining parallel computing efficiency, providing high-level semantic support for subsequent detailed temporal evolution analysis.

Figure 2(a) shows the autocorrelation function of the normalized power load series, with lag hours (0–12) on the horizontal axis and the autocorrelation coefficient on the vertical axis. Positive correlation is observed for short lags (0–5 hours), indicating continuity and short-term dependence within adjacent time periods. This is closely related to the persistent pattern of peaks and troughs in daily electricity demand. Negative correlation is observed for lags 6–12 hours, reflecting the cyclical fluctuations in load over the course of the day, with an inverse relationship between morning and evening peaks. Figure 2(b) shows a stacked area plot of the three-head and multi-head attention weights, with the horizontal axis representing the 24-hour period and the vertical axis representing the hourly contribution of each attention head. Different heads exhibit complementary attentional distributions within the day: the morning peak favors Head 1, the afternoon peak favors Head 2, and the nighttime trough is dominated by Head 3.

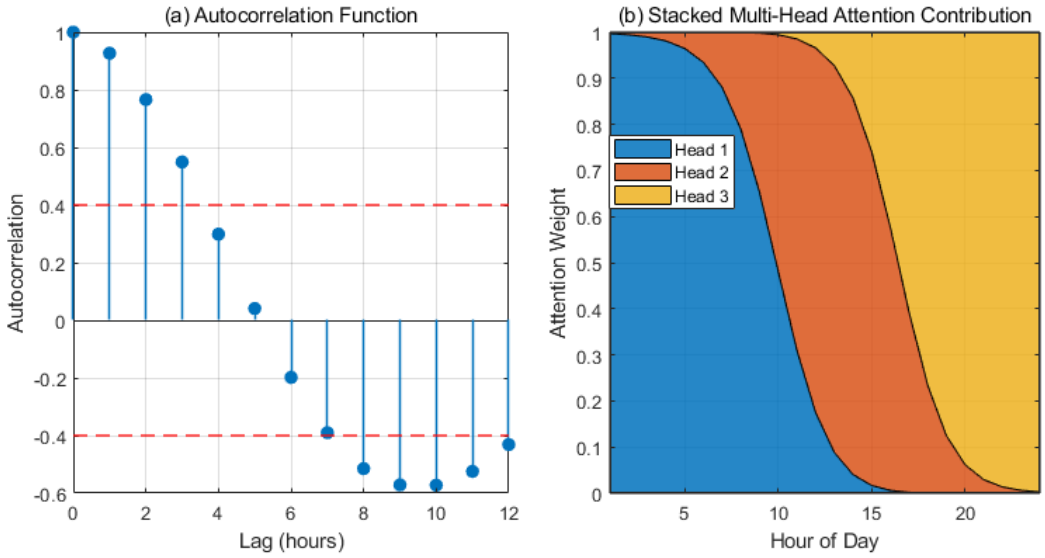

Figure 2. Autocorrelation function and attention weight

Capturing Local Temporal Dynamics in LSTM Modules

Modeling Local Dynamic Evolution Based on Gating Mechanisms

To precisely characterize the short-term non-stationary fluctuations and transient changes in power load sequences, a long short-term memory (LSTM) network is employed to analyze the time-step-by-time evolution of the high-level semantic representations output by the Transformer. The LSTM unit dynamically regulates information flow through its inherent gating structure, effectively balancing the retention of historical memory with responsiveness to current input. At each time step, the model receives the hidden state $\mathbf{h}_{t-1}$ and cell state $\mathbf{c}_{t-1}$ from the previous moment, as well as the current input vector $\mathbf{x}_t$ (i.e., the context-enhanced representation of the Transformer module output). Three learnable gating functions coordinate the decision-making information update path. The forget gate $\mathbf{f}_t$ determines which historical states need to be attenuated or discarded, the input gate $\mathbf{i}_t$ controls the acceptance of new information, and the output gate $\mathbf{o}_t$ determines the final representation of the current hidden state. As a long-term memory carrier, the cell state accumulates key trend information under an additive update mechanism to prevent the gradient from disappearing due to multiplicative decay during backpropagation. This process is achieved through the following nonlinear transformation:

$$\mathbf{f}_t = \sigma\big(W_f[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_f\big) \tag{5}$$

$$\mathbf{i}_t = \sigma(W_i[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_i) \tag{6}$$

$$\tilde{c}_t = \tanh(W_c[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_c) \tag{7}$$

$$c_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{c}}_t \tag{8}$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t) \tag{9}$$

W_* and $\mathbf{b}_*$ represent the weight matrix and bias term of each gate, respectively; σ represents the Sigmoid activation function; and $\odot$ represents an element-by-element multiplication operation. This gating mechanism endows the model with high sensitivity to local dynamics, such as sudden load changes and peak demand shifts. It can further capture fine-grained temporal evolution trajectories based on global patterns, demonstrating particularly strong responsiveness in critical regions where the load curve experiences sharp

rises or falls. Through recursive computation time-step by time step, the LSTM generates a sequence of time-aligned hidden states that fully captures the dynamic evolution of the input sequence at the local scale.

Bidirectional Structure-Driven Context-Aware Enhancement

To further improve the model's modeling completeness of local temporal dependencies, a bidirectional architecture is used to expand the information perception range of LSTM, allowing it to simultaneously fuse contextual information from the past and future directions under the constraints of a sliding window. Specifically, two independent LSTM sub-networks are constructed: the forward link processes the input sequence in chronological order, gradually accumulating historical dependencies from the starting point to the current moment; the backward link traverses the sequence in reverse order, extracting the potential impact of subsequent time periods on the current state. For the same time step, the forward hidden state $\vec{\mathbf{h}}_t$ encodes the cumulative dynamics from t=1 to t, while the backward hidden state $\overleftarrow{\mathbf{h}}_t$ reflects the reverse evolution trend from the end of the sequence to t. The two are spliced in the feature dimension to form a joint representation that integrates the bidirectional context:

$$\mathbf{h}_t^{\mathrm{bi}} = \left[\vec{\mathbf{h}}_t; \overleftarrow{\mathbf{h}}_t\right] \tag{10}$$

$\mathbf{h}_t^{\mathrm{bi}}$ represents a bidirectional hidden state. This design overcomes the unidirectional information acquisition limitations of unidirectional recurrent networks, enabling the model to comprehensively consider the fluctuation patterns of the preceding and following periods when processing the current load. The entire LSTM module achieves high-fidelity reconstruction of local non-stationary characteristics while maintaining the continuity of sequence modeling.

Table 1 presents the core parameter configuration of the bidirectional LSTM module in the short-term power load forecasting architecture, covering network structure, optimization strategy, and regularization mechanism.

Table 1 Core parameter configuration of bidirectional LSTM module

Parameter Name	Value	Description
LSTM Hidden Dimension (unidirectional)	128	Hidden state size of single-direction LSTM
Bidirectional LSTM Output Dimension	256	Concatenated forward and backward output size
Number of LSTM Layers	2	Number of stacked bidirectional LSTM layers
Sequence Length (Time Steps)	96	Number of time steps in input sliding window (corresponding to 24 hours, 15 minutes/step)
Initial Learning Rate	0.001	Initial learning rate for Adam optimizer
Batch Size	64	Number of samples processed per training batch
Dropout Rate	0.3	Dropout ratio applied between LSTM layers
Gradient Clipping Threshold	1	Maximum norm threshold to clip gradients
Weight Regularization (L2)	1e-5	L2 penalty coefficient for model parameters

Figure 3 schematically illustrates the simulated 24-hour normalized power load input sequence (blue dashed line) and the resulting state transitions within the LSTM unit. The solid black line represents the LSTM cell state, showing a smoother trajectory than the original load input, effectively filtering out short-term noise and demonstrating the LSTM's ability as a "memory unit" to integrate and retain long-term trend information. The dynamics of the green dashed line (input gate) and the red dashed line (forget gate) reveal the workings of the gating mechanism: at time points 6-10 (morning load ramp-up) and 18-20 (early evening peak), the input gate value increases significantly, indicating that the cell is actively accepting new input information to update its state. Simultaneously, the forget gate value decreases, indicating that the cell is selectively discarding some old, no longer relevant memories. The synergistic antagonism of these two gates enables the LSTM to finely regulate the information flow, thereby sensitively capturing local load increases and peaks while maintaining a stable memory of plateaus.

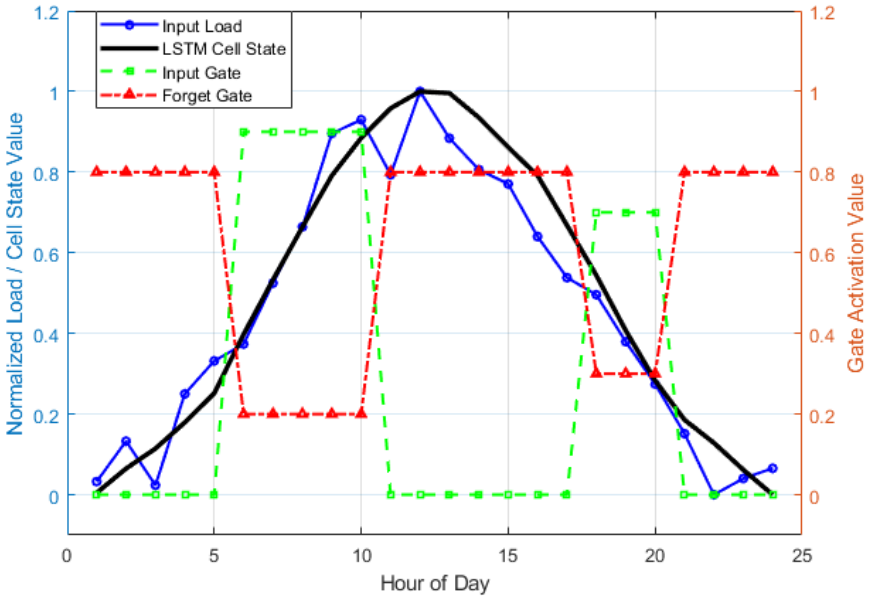

Figure 3. Input sequence and LSTM internal state dynamics

Figure 4 visualizes the activation intensity of all 64 hidden states of the bidirectional LSTM over a 24-hour period in heatmap format. Colors represent the activation value of each feature dimension at the corresponding time point. The heatmap shows some localized highlights, indicating that different feature neurons respond specifically to patterns in different time periods. Around time point 15, there is an exceptionally prominent highlight (marked by a white dashed box) in the 25th to 40th dimension.

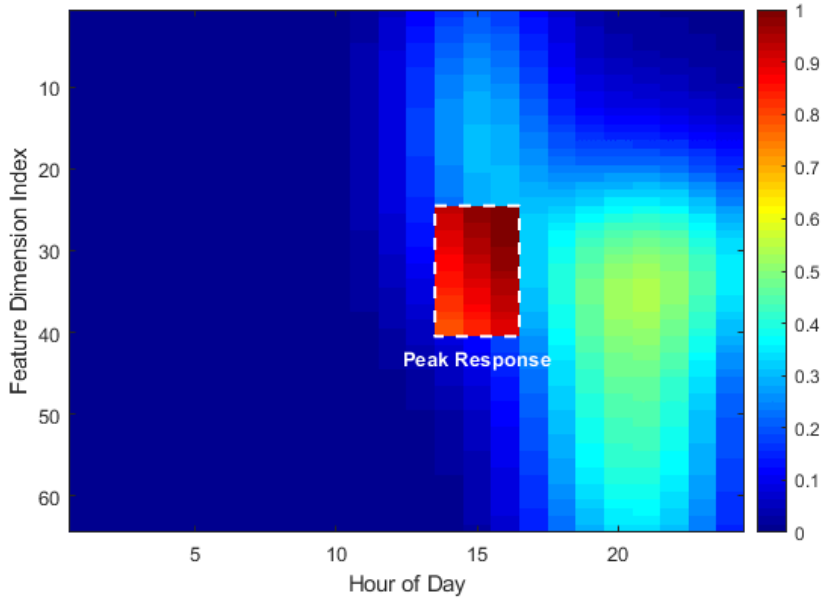

Figure 4. Bidirectional LSTM hidden state activation

Weighted Feature Integration in the Attention Fusion Layer

Construction of a Multi-Head Temporal-Feature Joint Attention Mechanism

To achieve the collaborative identification of key historical moments and dominant influencing factors, a learnable multi-head joint time-feature attention structure is designed to act on the hidden state sequence output by a bidirectional LSTM. The attention mechanism operates on the hidden state sequence from bidirectional LSTM outputs. Through multi-head parallel computation, it jointly weights both temporal steps and feature channels. Each head generates temporal step weights to focus on critical moments and dynamically adjusts dimensional contributions within the feature subspace. The final output is a contextual vector. This mechanism overcomes the limitations of traditional attention in a single dimension and simultaneously models the importance differences between time steps and the dynamic contributions between feature channels through parallel multi-subspace mapping. Each attention head performs a scaled dot product operation in an independent subspace, calculating the strength of the association between the current aggregate state and the hidden state of each historical moment, thereby identifying the time segment that is most indicative of the current prediction. The outputs of all heads are concatenated and linearly transformed to generate the final attention-weighted representation. This process is implemented through the following nonlinear mapping:

$$\mathbf{a}_t = \sum_{i=1}^{T} \alpha_{ti}\, \mathbf{h}_i^{\mathrm{bi}} \tag{11}$$

$$\alpha_{ti} = \frac{\exp\big(\mathrm{score}(\mathbf{q}_t, \mathbf{k}_i)\big)}{\sum_{j=1}^{T} \exp\Big(\mathrm{score}\big(\mathbf{q}_t, \mathbf{k}_j\big)\Big)} \tag{12}$$

$\mathbf{a}_t$ represents the context vector generated at time t; α_{ti} is the normalized attention weight; $\mathbf{q}_t$ and $\mathbf{k}_i$ are the query vector derived from the current state and the key vector corresponding to the i-th step, respectively. score uses the dot product function to measure the degree of match. The aforementioned calculations are executed in parallel across multiple attention heads. The outputs from each head are concatenated and then subjected to a linear transformation to generate the final context vector, enabling

joint modeling of time-feature interactions across multiple subspace dimensions. This mechanism empowers the model with dynamic focusing capabilities, enabling it to adaptively adjust the attention window at different forecast times, enhancing awareness of key areas such as load jump precursors and sudden meteorological changes. Furthermore, the multi-head architecture concurrently learns different time-feature dependency patterns in multiple representation subspaces, significantly improving the model's ability to resolve complex coupling relationships.

Multi-Scale Feature Fusion and Prediction Output Generation

In order to further integrate global semantic information and local context perception results, a multi-level feature splicing and nonlinear mapping path are designed to complete the regression prediction of the final load value. The context vector $\mathbf{a}_t$ output by the attention mechanism has encoded the weighted dynamic features of the key time step, but its representation is limited to the local evolution path and lacks deep semantic support for long-term trends. To this end, the high-level semantic feature $\mathbf{z}_t$ output by the last layer of the Transformer module is applied. This vector contains the global periodic pattern and cross-time correlation structure after self-attention extraction. The two are spliced in the feature dimension to form a joint representation that integrates long-term dependencies and dynamic focus:

$$\mathbf{R}_t = [\mathbf{a}_t; \mathbf{z}_t] \tag{13}$$

$\mathbf{R}_t$ represents the final fused feature vector, and the semicolon denotes vector concatenation. This fusion strategy avoids semantic loss during hierarchical transmission, ensuring the preservation of complementarity between global and local information, and between static patterns and dynamic responses. The concatenated high-dimensional vector is fed into a multi-layer fully connected network. Each layer is followed by a GELU (Gaussian Error Linear Unit) activation function and Dropout regularization, progressively compressing the feature dimensions and enhancing nonlinear fitting capabilities. The final layer outputs a single-dimensional real number, corresponding to the load forecast value at the next target time. The entire fusion process is optimized through end-to-end training, enabling the coordinated evolution of attention weights, feature mapping parameters, and prediction functions, achieving a seamless transition from multi-

scale time series representation to precise numerical regression. This structure not only improves the model's responsiveness to sudden changes in scenarios but also enhances its generalization stability under complex coupled meteorological and social behavior conditions.

MULTIDIMENSIONAL COMPARATIVE EVALUATION OF PREDICTION PERFORMANCE

Experimental Data

The experimental data were obtained from the operational records of a load sub-region in a large textile industrial park between 2019 and 2022. This sub-region aggregated a total load of approximately 100 megawatts, containing 15-minute resolution load data and synchronized meteorological observations. This study focuses on grid load forecasting for energy-intensive manufacturing sectors such as the textile industry. The selected region encompasses multiple large-scale textile industrial parks, whose electricity consumption exhibits typical industrial characteristics: pronounced daytime peaks, distinct nighttime troughs, temperature sensitivity (e.g., dyeing and finishing processes requiring constant temperature and humidity), and sudden start-stop cycles with non-stationary fluctuations influenced by order cycles and fast fashion trends. Consequently, this dataset effectively captures the dynamic power demand patterns in textile production scenarios. The original load series undergoes outlier cleaning and linear interpolation to remove extreme jump points caused by communication outages and equipment failures. Meteorological variables, including hourly temperature, humidity, wind speed, and weather type, are spatially aligned to each load collection node. Temporal feature encoding includes hour, day of the week, holiday identifiers, and seasonality to capture social activity cycles. The training set uses data from the first three years, while the validation and test sets cover the first four and last two months of the last six months, respectively, ensuring that the model evaluation is conducted on the true time series evolution path. All input variables are uniformly normalized to the interval [0, 1]. A sliding window with a window length of 96 time steps (24 hours) is used to construct the sample series. The prediction target is the load value at the next moment. Data division strictly follows chronological order to avoid future information leakage and ensure the credibility of experimental results and the reproducibility of practical applications. To comprehensively evaluate model performance, the baseline methodology includes classical statistical models (ARIMA), machine learning models (Support Vector Regression, SVR), mainstream deep architectures (Bidirectional Long Short-Term

Memory, BiLSTM; Convolutional Neural Network-LSTM, CNN-LSTM), and standalone Transformer models. BiLSTM and Transformer respectively represent the current best single-architecture solutions for local dynamic modeling and global dependency modeling.

Load Dynamics and Multidimensional Feature Attention Analysis

To balance model input complexity with practical scheduling requirements (typically hourly), the original 15-minute load data was resampled by averaging hourly values, while meteorological variables were aligned using average values from the same time window. This approach preserves intra-day peak-valley patterns while effectively suppressing high-frequency noise without introducing future information. An original dataset is constructed, consisting of hourly series of power load and meteorological characteristics (temperature and humidity). The data is then preprocessed: a sliding window approach is used to segment the continuous time series to capture intra-day and intra-week patterns. Each feature is normalized to the [0, 1] interval to eliminate dimensional differences. The load curves are then aligned with the meteorological characteristic curves and visually compared on the same timeline, with weekend backgrounds annotated to highlight cyclical differences. This visual assessment aims to reveal correlations, time lags, and short-term fluctuations between load and meteorological variables, providing structural evidence for subsequent modeling.

Figure 5 shows the short-term electricity load curve for 7 days (168 hours) and the time-series variations of its main external drivers—temperature and humidity. The horizontal axis represents time (unit: hours), and the vertical axis represents the normalized magnitude. The solid blue line represents the electricity load curve, showing clear intraday peaks and valleys: a diurnal load cycle with superimposed random fluctuations. The dashed red line represents temperature (scaled for comparison), showing a phase difference between it and the load. Load peaks are particularly pronounced during high-temperature periods, indicating the driving effect of increased cooling demand on electricity consumption. The dashed green line represents humidity, also showing low-frequency fluctuations across weeks, providing a background influence on load. Furthermore, the gray shading marks the weekend period, showing an overall decrease in load, reflecting the difference between residential and industrial electricity consumption during weekdays and weekends.

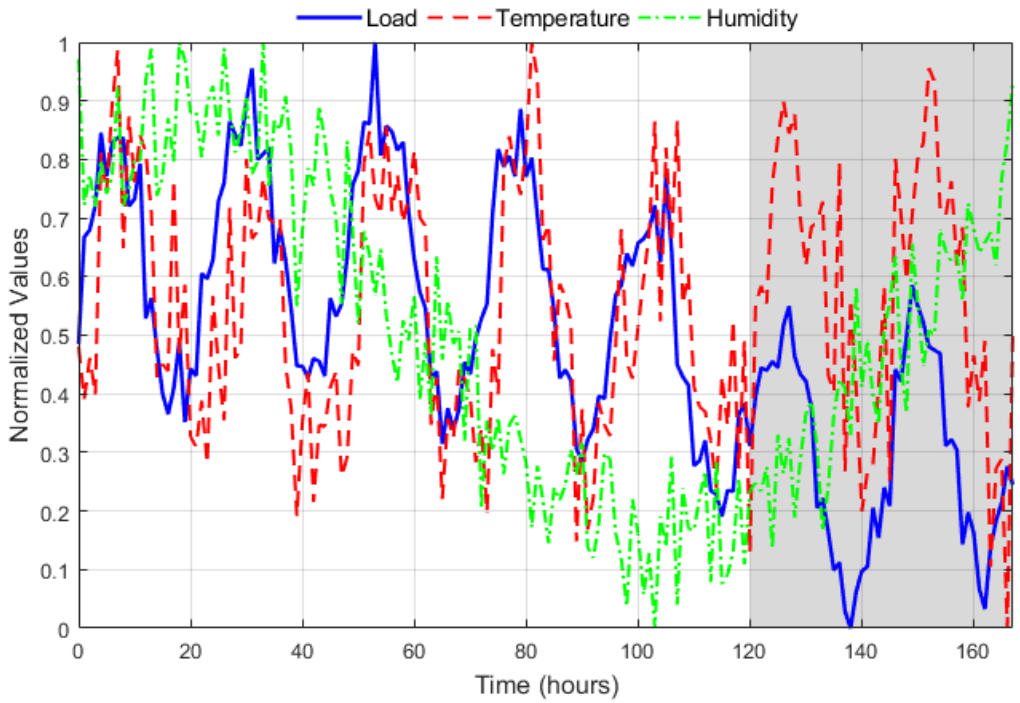

Figure 5. Cyclical normalized load and weather characteristics

Figure 6(a) shows a normalized short-term electricity load series, with the horizontal axis representing the 24-hour day (hour of day) and the vertical axis representing the normalized load value (0-1). The data demonstrates a typical intraday load fluctuation pattern: a rapid rise in the morning as residential and industrial electricity consumption begins, reaching a peak of 1 at noon, followed by a subsequent decline, and a gradual return to a low point at night. Normalization not only eliminates absolute value differences but also highlights the intraday dynamics of load fluctuations. Peak labeling emphasizes peak load periods. This load variability is primarily due to the cyclical nature of electricity consumption, industrial production schedules, and the lagged effects of external meteorological factors such as temperature on load. Figure 6(b) shows a heat map of the Transformer's three-head self-attention weights. The horizontal axis represents hours, and the vertical axis represents the three-head attention (Heads 1–3). Color indicates weight. Head 1 focuses on the morning peak, Head 2 on afternoon load fluctuations, and Head 3 on the nighttime trough.

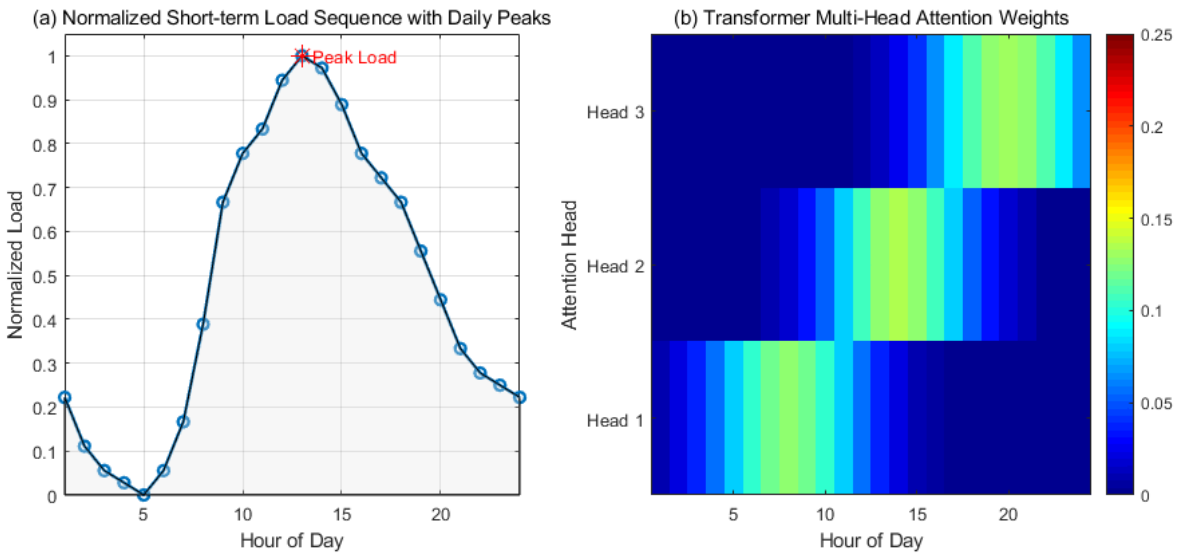

Figure 6. Normalized load and multi-head attention distribution

Comparison of MAE and RMSE at Different Time Scales

For short-term (15 minutes), medium-term (1 hour), and long-term (6 hours) forecast steps, it calculates the mean absolute error (MAE) and root mean square error (RMSE) of the Transformer-LSTM model proposed in this paper compared with the currently popular ARIMA (Autoregressive Integrated Moving Average Model), SVR (Support Vector Regression), Transformer, Bi-directional Long Short-Term Memory (BiLSTM), and CNN-LSTM. It specifically compares the changes in forecast error among these models under short-term, high-frequency load fluctuations. All error metrics are based on three independent training/test runs, with ± indicating standard deviation.

Figure 7 shows the multi-scale error performance of the proposed model and mainstream methods at different forecast time domains. The MAE in Figure 7(a) reflects the overall model bias. The Transformer-LSTM has a significant advantage in short-term forecasting, with an MAE of (0.75±0.04) MW, thanks to its ability to fine-tune the modeling of high-frequency load fluctuations. Traditional models such as ARIMA lack a nonlinear fitting mechanism, and errors accumulate rapidly as the forecast step size increases. The RMSE in Figure 7(b) reveals the model's sensitivity to abnormal errors, the root mean square error (RMSE) of Transformer-LSTM in short-term prediction is (0.90 ± 0.06) MW. The BiLSTM and SVR have significantly higher RMSEs in medium- and long-term forecasts, indicating that they experience response lags or overshoots at load mutation points, resulting in the square amplification of occasional large errors. It is worth noting that

the BiLSTM model does not achieve the highest MAE, but its performance declines in RMSE. This indicates that while its forecast results have a small average deviation, they suffer from localized extreme inaccuracies, exposing the instability of capturing time series dynamics.

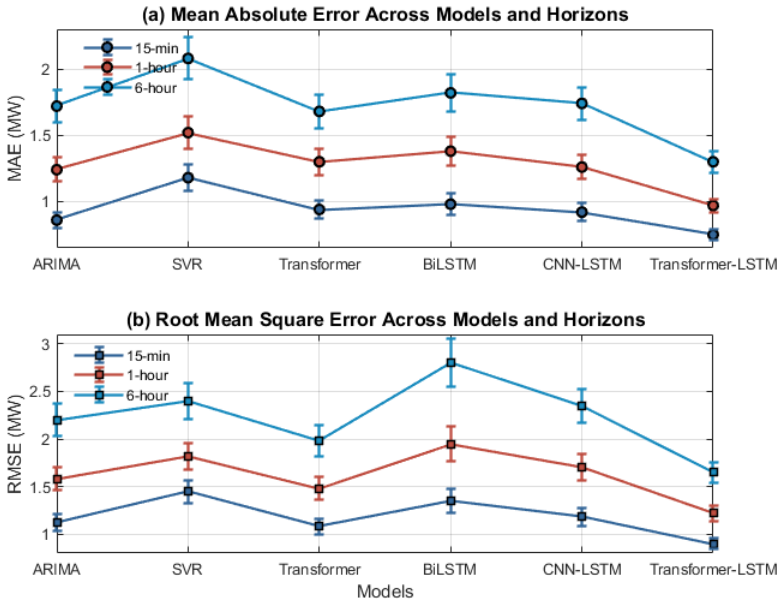

Figure 7. Multi-scale error performance under different prediction time domains

Analysis of Differences Between MAPE and R^2 in Different Meteorological Regions

To systematically evaluate the model's generalization capabilities under diverse climate environments, three typical meteorological regions—the hot south, the cold north, and the temperate central region—are selected as experimental sites. A cross-regional forecasting performance comparison is conducted based on a unified model configuration and training strategy. All models are independently trained and tested in each region, and the input features include localized meteorological variables and time encoding to ensure accurate representation of external driving factors. MAPE (Mean Absolute Percentage Error) is used to measure the relative magnitude of forecast deviations to overcome evaluation bias caused by differences in load bases. The coefficient of determination (R^2) is used to quantify the model's ability to explain actual load fluctuations. The mean and standard deviation of the performance indicators are obtained through three repeated experiments, and the discreteness of the results is presented as error bars. The focus is on analyzing

the impact of climate extremes on model robustness, revealing the differences in adaptability of different architectures to temperature-sensitive load responses.

Figure 8 shows the differences in forecast stability between different models under three typical climate conditions: high temperatures in the south, cold temperatures in the north, and mild temperatures in the central region. The MAPE indicator shows that traditional model errors increase significantly in extreme climate regions, particularly during periods of sharp increases in heating loads in northern winter and peak cooling loads in southern summer. ARIMA and SVR exhibit significant systematic biases due to their difficulty in capturing the nonlinear coupling between temperature and load. The Transformer-LSTM model maintained the lowest error across all three regions, with an average MAPE of 3.2%-4.7%. This demonstrates its enhanced cross-regional adaptability by integrating meteorological characteristics with time-dependent modeling. R^2 analysis further revealed that the explanatory power of BiLSTM and CNN-LSTM decreased significantly in cold regions, reflecting their lack of balance between long-term dependencies and sudden changes. The R^2 of Transformer-LSTM is 0.94-0.97.

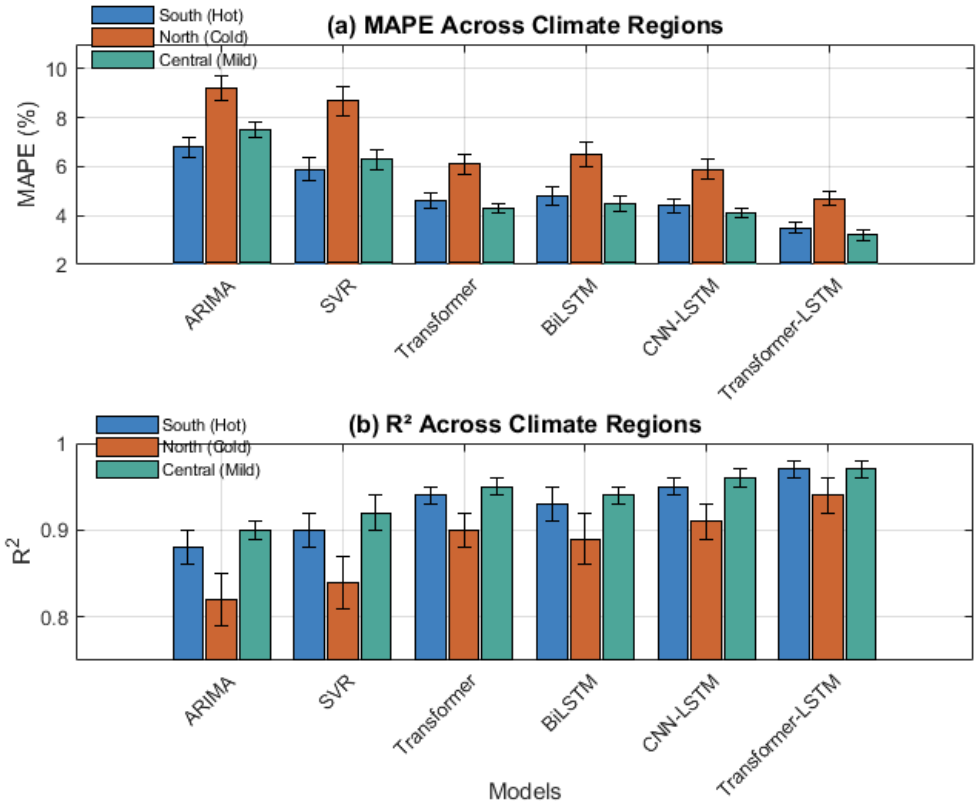

Figure 8. Cross-regional performance stability under climate change

Robustness Testing in Weekday and Holiday Scenarios

To evaluate the model's adaptability in scenarios with sudden changes in load behavior, a comparative experiment is conducted on two time periods with significant differences in electricity consumption patterns: typical working days and statutory holidays. Workday loads are driven by industrial and commercial activities and exhibit a regular peak-valley structure. Holiday loads, on the other hand, exhibit atypical characteristics such as extended low-peak periods, peak-time shifts, and sudden load fluctuations due to changes in social behavior patterns. Under the same training-testing split, each model independently makes predictions in the two scenarios, calculating their mean absolute percentage error (MAPE) and coefficient of determination (R^2) to quantify the fitting accuracy and stability of different load dynamics. The focus is on analyzing whether the model experienced a cliff-like drop in performance during the mode switching process, revealing its ability to perceive social temporal disturbances and behavioral heterogeneity, and further testing its robustness boundaries in actual scheduling applications.

Table 2 compares the prediction performance of various models under two typical electricity consumption scenarios, weekdays and holidays, to assess their adaptability to sudden changes in load patterns. Weekday loads are characterized by strong cyclicality and clear peak-valley patterns, leading all models to demonstrate high accuracy, with the Transformer-LSTM performing the best. However, in holiday scenarios, due to changes in residential behavior, reduced industrial load, and increased randomness in electricity consumption, the performance of all models generally declines, with a significant increase in MAPE and a decrease in R^2, highlighting the challenges of modeling irregular load fluctuations. Traditional models such as ARIMA and SVR, due to their reliance on linear trend and fixed-cycle assumptions, experience significant increases in error during holidays, exposing their vulnerability to sudden changes in patterns. BiLSTM and Transformer still exhibit response lags during peak-time offsets and periods of sustained low load. Notably, Transformer-LSTM maintains the lowest error and highest coefficient of determination in both scenarios, with an average MAPE of no more than 5.12% and an average R^2 of no less than 0.928.

Table 2 Robustness test

Model	Scenario	MAPE (%)	R^2
ARIMA	Workday	5.82 ± 0.31	0.912
	Holiday	9.65 ± 0.58	0.803
SVR	Workday	5.37 ± 0.29	0.928
	Holiday	8.92 ± 0.51	0.826
Transformer	Workday	4.15 ± 0.24	0.956
	Holiday	6.78 ± 0.42	0.887
BiLSTM	Workday	4.08 ± 0.26	0.958
	Holiday	7.21 ± 0.45	0.873
CNN-LSTM	Workday	3.92 ± 0.22	0.961
	Holiday	6.53 ± 0.39	0.894
Transformer-LSTM	Workday	3.41 ± 0.19	0.973
	Holiday	5.12 ± 0.33	0.928

Comparison of Convergence Rounds and Inference Time under Different Data Missing Rate Interference

To evaluate the training stability and operational efficiency of the model in an environment with degraded data quality, a multi-level data missing interference experiment is designed. Key eigenvalues in the historical input sequence are randomly set to zero at ratios of 10%, 20%, and 30%, simulating data missing problems caused by communication failures or sensor failures in actual acquisition systems. All models are trained under the same initialization conditions, and the evolution of their validation set performance is monitored. The number of training rounds (convergence rounds) required to achieve the minimum validation error is recorded to measure the stability and learning efficiency of the parameter optimization path. At the same time, the average time consumed for a single forward inference is calculated to reflect the model's

responsiveness in edge deployment scenarios. By comparing the convergence speed and time overhead of each model under different missing rates, their tolerance for incomplete input and the growth trend of computational complexity are analyzed, revealing the profound impact of structural design on training robustness and inference efficiency.

Table 3 compares the training convergence and inference efficiency of various models under varying data missingness rates, assessing their robustness and practicality under conditions of degraded data quality. As the missingness rate increases from 10% to 30%, the number of convergence rounds for all models increases, indicating that incomplete data significantly interferes with the parameter optimization process. Traditional models such as ARIMA, which rely on complete sequences for parameter estimation, experience a significant decrease in convergence speed. However, the Transformer-LSTM, benefiting from its gradient-driven, fault-tolerant learning mechanism, exhibits greater training stability. The proposed Transformer-LSTM maintains the lowest number of convergence rounds under various missingness conditions, particularly when the missingness rate is 30%, achieving a convergence round of 134 ± 9 and an inference time of 40.3 ms.

Table 3 Convergence rounds and inference time

Model	Missing Rate	Convergence Rounds	Inference Time (ms)
ARIMA	10%	186 ± 12	3.2
	20%	214 ± 15	3.3
	30%	257 ± 21	3.4
SVR	10%	142 ± 9	8.7
	20%	168 ± 11	9.1
	30%	203 ± 14	9.5
Transformer	10%	98 ± 6	42.6
	20%	125 ± 8	43.1
	30%	162 ± 12	44
BiLSTM	10%	112 ± 7	28.3

	20%	146 ± 9	29
	30%	189 ± 13	30.2
	10%	105 ± 6	35.8
CNN-LSTM	20%	137 ± 8	36.5
	30%	176 ± 11	37.3
	10%	89 ± 5	38.5
Transformer-LSTM	20%	108 ± 6	39.1
	30%	134 ± 9	40.3

[Note: For ARIMA and SVR, "convergence epoch" refers to the number of independent model evaluations completed during hyperparameter optimization (such as the number of attempts at grid search or automated modeling), not the number of training epochs in the neural network. All experiments were conducted on an NVIDIA RTX 3090 GPU, Intel Xeon Gold 6248R CPU, and 64 GB of memory, with batch size uniformly set to 64. The prediction task was configured as single-step ahead (i.e., forecasting the load value at the next time step)]

CONCLUSION

This paper addresses the challenges of insufficient modeling of multi-scale temporal dependencies and neglect of dynamic changes in feature importance in short-term power load forecasting. The ability to overcome these modeling limitations holds significant promise for improving complex time series predictions, such as those related to production and demand variability within the textile industry. By integrating the Transformer-LSTM module with a multi-head time-feature attention mechanism, this hybrid forecasting framework combines long-term global dependencies captured by the Transformer module and fine-tuned local dynamic evolution by combining the bidirectional LSTM module. This framework achieves a coordinated representation of cross-period regularities and transient fluctuations in load series. Furthermore, a multi-head time-feature joint attention mechanism is applied to adaptively weight key historical moments and dominant influencing factors, enhancing the model's contextual awareness in complex external

environments. Experimental results demonstrate that the proposed model significantly outperforms mainstream methods such as ARIMA, SVR, Transformer, and BiLSTM across various time scales, climate regions, and electricity usage scenarios. It demonstrates greater robustness and stability under sudden load changes, holiday mode switching, and high-missing data interference. Multi-dimensional comparisons validate its comprehensive advantages in metrics such as MAE, RMSE, MAPE, and R^2, while also demonstrating good convergence efficiency and acceptable inference overhead. This architecture effectively addresses the imbalance between global and local feature extraction in a single model, a capability particularly vital for the textile industry where fluctuating production schedules require precise and adaptable power planning, thereby providing a scalable technical path for high-precision, highly generalizable short-term load forecasting. It has the potential for practical deployment in smart grid dispatching and energy management systems, particularly enabling textile manufacturers to optimize the highly energy-intensive dyeing and finishing processes by allowing for precision load shifting and real-time power consumption control.

Author Contributions

Conceptualization – Minjing Yang, Binghong Su, Jiaxin Zhuo and Xuanli Lan; methodology – Minjing Yang, Binghong Su, Qinwei Duan, Yashan Zhong, Jiaxin Zhuo and Xuanli Lan; investigation – Minjing Yang, Binghong Su, Qinwei Duan, Yashan Zhong, Jiaxin Zhuo and Xuanli Lan; writing-original draft preparation – Minjing Yang, Binghong Su, Qinwei Duan, Yashan Zhong, Jiaxin Zhuo and Xuanli Lan. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

Funding

The research is supported by the Science and Technology Project of China Southern Power Grid Corporation: Research and Application Development of Load Forecasting Technology in Regions with New Energy Permeability Exceeding 40% - Topic 3: Research on Intelligent Bus Load Forecasting Technology Considering Source Load Interaction and Application Development of Regional Load Intelligent Forecasting System (No.

036000KK52210065(GDKJXM20210096)).

Acknowledgements

Not applicable.

REFERENCE

[1] Dong X, Deng S, Wang D. A short-term power load forecasting method based on k-means and SVM. Journal of Ambient Intelligence and Humanized Computing. 2022; 13(11):5253-5267. doi: 10.1007/s12652-021-03444-x

[2] Kim N, Park H, Lee J, Choi JK. [Yy10.1]Short-term electrical load forecasting with multidimensional feature extraction. IEEE Transactions on Smart Grid. 2022; 13(4):2999-3013. doi: 10.1109/TSG.2022.3158387

[3] Veeramsetty V, Reddy KR, Santhosh M, Mohnot A, Singal G. Short-term electric power load forecasting using random forest and gated recurrent unit. Electrical Engineering. 2022; 104(1):307-329. doi: 10.1007/s00202-021-01376-5

[4] Li J, Wei S, Dai W. Combination of manifold learning and deep learning algorithms for mid-term electrical load forecasting. IEEE Transactions on Neural Networks and Learning Systems. 2021; 34(5):2584-2593. doi: 10.1109/TNNLS.2021.3106968

[5] Neeraj N, Mathew J, Agarwal M, Behera RK. Long short-term memory-singular spectrum analysis-based model for electric load forecasting. Electrical Engineering. 2021; 103(2):1067-1082. doi: 10.1007/s00202-020-01135-y

[6] Wu D, Lin W. Efficient residential electric load forecasting via transfer learning and graph neural networks. IEEE Transactions on Smart Grid. 2022; 14(3):2423-2431. doi: 10.1109/TSG.2022.3208211

[7] Ozer I, Efe SB, Ozbay H. A combined deep learning application for short term load forecasting. Alexandria Engineering Journal. 2021; 60(4):3807-3818. doi: 10.1016/j.aej.2021.02.050

[8] Lin X, Zamora R, Baguley CA, Srivastava AK. A hybrid short-term load forecasting approach for individual residential customer. IEEE Transactions on Power Delivery. 2022; 38(1):26-37. doi: 10.1109/TPWRD.2022.3178822

[9] Lin W, Wu D, Boulet B. Spatial-temporal residential short-term load forecasting via graph neural networks. IEEE Transactions on Smart Grid. 2021; 12(6):5373-5384. doi: 10.1109/TSG.2021.3093515

[10] Chen Z, Zhang D, Jiang H, Wang L, Chen Y, Xiao Y, et al. Load forecasting based on LSTM neural network and applicable to loads of "replacement of coal with electricity". Journal of Electrical Engineering & Technology. 2021; 16(5):2333-2342. doi: 10.1007/s42835-021-00768-8

[11] Wei T, Pan T. Short-term power load forecasting based on LSTM neural network optimized by improved PSO. Journal of System Simulation. 2021; 33(8):1866-1874.

[12] Xia M, Shao H, Ma X, De Silva CW. A stacked GRU-RNN-based approach for predicting renewable energy and electricity load for smart grid operation. IEEE Transactions on Industrial Informatics. 2021; 17(10):7050-7059. doi: 10.1109/TII.2021.3056867

[13] Agga FA, Abbou SA, El Houm Y, Labbadi M. Short-term load forecasting based on CNN and LSTM deep neural networks. IFAC-PapersOnLine. 2022; 55(12):777-781. doi: 10.1016/j.ifacol.2022.07.407

[14] Zhang X, Chau TK, Chow YH, Fernando T, Iu H. A novel sequence to sequence data modelling based CNN-LSTM algorithm for three years ahead monthly peak load forecasting. IEEE Transactions on Power Systems. 2023; 39(1):1932-1947. doi: 10.1109/TPWRS.2023.3271325