DYNAMIC LOAD FORECASTING OF THERMAL STORAGE SYSTEM BASED ON MULTIMODAL TRANSFORMER

by

Xiaoling LI[a,b] and Kai LI[c*]

[a] College of Big Data and Artificial Intelligence,
Zhengzhou University of Economics and Business, Zhengzhou, China
[b] Henan Province Engineering Research Center of Multimodal Perception and Intelligent Interaction Technology, Zhengzhou, China
[c] Zhenzhou Technical College, Zhengzhou, China

Original scientific paper
https://doi.org/10.2298/TSCI2506247L

Driven by the "dual carbon" goal, this paper proposes a dynamic load forecasting method based on a multimodal transformer to improve the operating efficiency of the thermal storage system. This method constructs a multimodal dataset, designs an adaptive modal weight allocation mechanism, and utilises the transformer structure to capture both long-term trends and short-term fluctuations. The experiment utilises data from a regional heating system as a sample and compares it with models such as ARIMA and LSTM. The results show that the proposed model performs best in the test set, with RMSE of 2.35 kW, MAE of 1.82 kW, MAPE of 2.15%, and MaxAE of 5.27 kW, which are 18.3%, 16.8%, 15.2%, and 19.4% lower than the suboptimal multimodal CNN-LSTM, respectively. The two-stage fusion combines dynamic modal weights and gating, outperforming single-stage by 11% in MAPE. Meteorological data dominates in cold waves due to strong correlation with load (r = 0.82). Transformer's self-attention models global dependencies, avoiding LSTM sequential loss, explaining its three hours RMSE of 7.82 kW vs. LSTM 9.65 kW. It has significant advantages in high load intervals and long-time series predictions, providing support for the intelligent operation of the system.

Key words*: multimodal transformer, thermal storage system, multimodal fusion, dynamic load forecasting, prediction accuracy*

Introduction

At the critical stage of the global energy structure's transition low carbonisation, driven by the *dual carbon* goal, improving energy efficiency has become a core issue. As a key hub for integrating intermittent renewable energy and meeting stable heating demand, the operating efficiency of the heat storage system is directly related to the economic and environmental sustainability of the energy system. As the core link in the system's optimal scheduling, dynamic load forecasting provides a decision-making basis for configuring heat storage capacity and formulating energy allocation strategies, thereby reducing energy consumption and carbon emissions [1]. However, the dynamic load of the heat storage system is cross-affected by multiple factors, such as environmental parameters (*e.g.*, outdoor temperature and humidity), periodic changes in users' heating habits, and time-varying equipment operating conditions. The load curve exhibits complex characteristics, including strong

* Corresponding author, e-mail: likai@zzy.edu.cn

non-linearity and multi-scale fluctuations, making it challenging for traditional forecasting methods to meet high precision requirements. There are technical bottlenecks in the current mainstream forecasting methods: traditional time series analysis methods such as ARIMA can only capture linear trends, and the prediction error for non-linear fluctuations exceeds 10%. Single-mode machine learning models such as back propagation neural networks rely on a single type of data, ignore cross-domain information association, and have poor stability under complex working conditions; recurrent neural networks such as LSTM have significantly reduced the accuracy of long-term load trend forecasts over 72 hours, and the feature extraction capability is insufficient under multi-factor coupling [2]. In recent years, multimodal learning technology and transformer models have shown advantages in related fields, but their application in dynamic load forecasting of thermal storage systems faces challenges, the characteristic scales of different modal data vary greatly, making it difficult to fuse directly, the correlation between load fluctuations and multimodal factors changes dynamically over time, and the fixed weight fusion method cannot adapt.

This paper proposes a prediction method based on the multimodal transformer, constructs a multimodal dataset, designs an adaptive modal weight allocation mechanism, and utilises the transformer structure to enhance the capture of both long-term trends and short-term fluctuation characteristics [3]. Through comparative experiments, its advantages in prediction accuracy (MAPE drops to less than 5%) and stability are verified, providing support for the intelligent operation of the system.

Construction of a dynamic load prediction model for the thermal storage system based on a multimodal transformer

Multimodal data collection and preprocessing

The collected data include core parameters and environmental variables of the thermal storage system: inlet and outlet temperature of the thermal storage tank (50-90 °C), five minutes sampling interval), circulating water pump flow (0-50 m^3 per hour), real-time load (10-150 kW), as well as outdoor temperature and humidity, light intensity and time characteristics [4]. During preprocessing, an improved interpolation algorithm is employed to fill in missing values, and anomalies are identified and replaced with the median of the time series using an adaptive threshold method (based on a threshold of three times the standard deviation of the sliding window). Numerical data is converted to the [0, 1] interval by an improved standardization formula:

$$x_i' = \frac{x_i - \min(x)}{\max(x) - \min(x) + \alpha} \tag{1}$$

where x_i is the original numerical data, x_i' – the standardized data, min(x) – the minimum value in the data set, max(x) – the maximum value in the data set, and α – the smoothing coefficient (taken as 0.001) to avoid the denominator being zero. Categorical data uses embedded coding:

$$s_j = \sum_{k=1}^{m} \omega_k \mathrm{OneHot}(c_{jk}) \tag{2}$$

where s_j is the embedding encoding result of the j sample, ω_k – the weight vector of the k feature, OneHot(c_{jk}) – the one-hot encoding function, c_{jk} – the k feature of the j sample, and m – the total number of feature categories, and the weight distribution is dynamically optimized through training.

Overall architecture design of the model

The model adopts an improved Encoder-Decoder architecture. The multimodal embedding layer maps heterogeneous data to a unified feature space, and the continuous data embedding formula:

$$\mathrm{Emb}_c\left(x_i\right) = \mathrm{Linear}\left(x_i'\right) + \mathrm{PosEnc}\left(t_i\right) \tag{3}$$

where $\mathrm{Emb}_c(x_i)$ is the result of embedding the continuous data x_i, $\mathrm{Linear}(x_i')$ – the linear transformation of the standardized x_i', and $\mathrm{PosEnc}(t_i)$ – the time position encoding, which uses the sine-cosine mixed function give the data time position information:

$$\mathrm{PosEnc}\left(t_i\right) = \begin{cases} \sin\left(\dfrac{t_i}{10000^{\frac{2d}{512}}}\right) & d \text{ even number} \\ \cos\left(\dfrac{t_i}{10000^{\frac{2d}{512}}}\right) & d \text{ for odd numbers} \end{cases} \tag{4}$$

where t_i is the time step and d – the dimension in the position encoding vector. The encoder uses a multi-head self-attention mechanism to capture global dependencies. Learnable encoding increased RMSE by 3% due to overfitting. Sine-cosine better captures daily/weekly cycles. The 512 dimensions balanced performance: 256 reduced long-term accuracy, 1024 added 40% computation with no gain, making it optimal:

$$\mathrm{MultiHead}\left(Q,K,V\right) = \mathrm{Concat}\left(\mathrm{Head}_1,\ldots,\mathrm{Head}_h\right)W^O \tag{5}$$

$$\mathrm{Head}_i = \mathrm{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{6}$$

where MultiHead(Q, K, V) is the output of the multi-head self-attention mechanism, Q, K, V are query, key, and value matrices, respectively, Head_1 – the output of the i attention head, W_i^Q, W_i^K, W_i^V – the weight matrices used to generate the query, key, and value of the i attention head, respectively, W^O – the weight matrix used to transform the concatenated outputs of multiple attention heads linearly. Eihgt heads were optimal via search: four missed fluctuations, 16 added 60% time. Analysis shows three heads focus on peaks, two on trends, three on weather, demonstrating diverse capture critical for performance.

The Attention function uses the scaled dot product mode to calculate attention [5]. The decoder fuses the encoder output with its hidden state through the cross-attention layer and then outputs the predicted value after a linear transformation.

Multimodal feature fusion module

Design a two-stage fusion mechanism. Early fusion generates dynamic weights through modal attention, and the calculation formula is:

$$\beta_m = \frac{\exp\left[\mathrm{MLP}\left(f_m\right)\right]}{\sum_{n=1}^{M}\exp\left[\mathrm{MLP}\left(f_n\right)\right]} \tag{7}$$

where β_m is the normalized weight of the m modal feature, f_m – the m modal feature, MLP – the multi-layer perceptron, which is used to perform non-linear transformations on features, and

M – the total number of modalities. Late fusion uses a gating mechanism, and the specific formula is:

$$\begin{aligned} g_t &= \sigma\left(\text{Linear}\left(\left[h_t, c_t\right]\right)\right) \\ f_t &= g_t \odot \tanh\left(\text{Linear}\left(h_t\right)\right) + \left(1 - g_t\right) \odot c_t \end{aligned} \tag{8}$$

where g_t is the gate value, which is obtained by processing the output of Linear($[h_t, c_t]$) by σ (sigmoid function), $[h_t, c_t]$ – the concatenation of the decoder hidden state h_t and the encoder context vector c_t, f_t – the fused feature, $\odot$ – the dot product of the matrix elements, and tanh – the hyperbolic tangent activation function. The feature fusion ratio is dynamically adjusted through gating to meet the feature fusion requirements under different stages and conditions.

Model training and optimization

The loss function employs a weighted mixed loss:

$$L = \lambda \text{MSE}\left(y, \hat{y}\right) + \left(1 - \lambda\right)\text{MAE}\left(y, \hat{y}\right) \tag{9}$$

where L is the loss value, λ – the weight coefficient (valued at 0.7), which is used to adjust the proportion of mean square error MSE(y, $\hat{y}$), and mean absolute error MAE(y, $\hat{y}$) in the loss function. The $\lambda = 0.7$ minimized validation loss via grid search, $\lambda = 0.9$ increased MaxAE by 12%. Peak mask (1.5×) aligns with utility data showing peak errors cost 2× more. Huber loss was tested but had 4% higher MAPE. This set-up reduced peak RMSE to 2.87 kW *vs*. CNN-LSTM 3.52 kW. The y is the actual value, and $\hat{y}$ is the predicted value [6]. The peak period error is weighted by the time mask matrix Mask(t), and the specific calculation of the mean square error is:

$$\text{MSE}\left(y, \hat{y}\right) = \frac{1}{N}\sum_{t=1}^{N}\text{Mask}\left(t\right)\left(y_t - \hat{y}_t\right)^2 \tag{10}$$

where N is the total number of samples, y_t and $\hat{y}_t$ are the actual value and predicted value at time t, respectively. The Mask(t) takes a larger value during the peak period and a smaller value during the non-peak period, to highlight the attention paid to the error during the peak period [7]. The optimizer uses the improved Adam algorithm, and the learning rate update formula is:

$$\eta_t = \eta_0 \frac{\sqrt{1 - \beta_2^t}}{1 - \beta_1^t}\cos\left(\frac{\pi t}{2T}\right) \tag{11}$$

where η_t is the learning rate at time t, η_0 – the initial learning rate, β_1 and β_2 – the moment estimation decay rates in the Adam optimizer, and T – the total number of iterations. The learning rate is dynamically adjusted through the cosine annealing strategy, combined with Dropout (rate 0.2) and L2 regularisation (coefficient $1 \cdot 10^{-4}$) to suppress overfitting, allowing the model to converge more effectively and prevent overfitting during training, thereby improving the model's generalisation ability [8].

Experimental simulation and result analysis

Experimental data set and environment

The experimental data comes from the actual operation record of a regional central heating system, which covers 5000 residential communities and three office buildings, with a total heating area of 80000 square meters, equipped with a 100 m³ water heat storage tank and two 200 kW gas boilers [9]. The CPU (i7-12700K) inference took 186 ms (acceptable for non-real-time). The RTX 3060 achieved 45.2 ms. Training on RTX 4090 used 2.3 kWh/100 epochs

(30% less than RTX 3090), balancing performance and efficiency. The dataset contains data from January to August 2023, comprising a total of 69120 time steps (with a sampling interval of five minutes). It is divided into a training set (the first six months, 51840 steps), a validation set (the 7th month, 13824 steps), and a test set (the 8th month, 13824 steps). Missing data used KNN interpolation ($k = 5$), outperforming linear by 3%. Special conditions were stratified (8% in each split) and cold wave data was augmented with noise. The 69000 steps are sufficient: doubling data reduced loss by <1%. The 9.2M parameters are justified by 9.2% lower MAPE than a smaller model. The data covers the winter, transition, and summer seasons, including special working conditions such as cold waves, rain, snow, and holidays (accounting for 8%), which can thoroughly test the model's generalisation ability. The experiment was implemented using Python 3.9 and PyTorch 2.0. The hardware consisted of an Intel Xeon W-2295 CPU, an NVIDIA RTX 4090 GPU, 128 GB of memory, and 2 TB of NVMe storage. With CUDA 11.7 acceleration, a single round of iteration (51840 samples) took about 45 seconds, and 100 rounds of training took a total of 7.5 hours, meeting the timeliness requirements.

Selection of evaluation indicators

Four core indicators are used to evaluate model performance: RMSE, which reflects the overall deviation between the predicted value and the actual value; MAE, which objectively measures the average level of error; MAPE, which is presented in the form of relative error to facilitate comparison between different systems; MaxAE, which evaluates the ability to control extreme load fluctuations [10]. The indicators are subdivided by load range (low load: <50 kW, medium load: 50-100 kW, high load: >100 kW) and prediction duration (1 hours/3 hours/6 hours). The load range is divided according to the system's rated load (150 kW), and the prediction duration covers the typical scheduling decision cycle.

Comparative experimental design

Five types of comparison models are selected: traditional statistical model ARIMA ($p = 5$, $d = 1$, $q = 3$). Machine learning model back propagation neural network (3-layer fully connected, number of nodes 128-64-32, activation function ReLU); time series deep learning model LSTM (2-layer recurrent network, hidden unit 128, dropout rate 0.2). Multimodal baseline model CNN-LSTM (3-layer convolution + 2-layer LSTM, convolution kernel 3×3, output channel 64). Unimodal transformer (only input load and temperature data). All models employ the same preprocessing process and training parameters: Adam optimiser, initial learning rate of 0.001, batch size of 128, and an upper limit of 100 training rounds. Early stopping occurs when the validation set loss does not decrease for 15 consecutive rounds [11]. The experiment consists of three groups of subtasks: overall load prediction, peak period prediction (8:00 a. m. to 10:00 a. m./18:00 p. m. to 20:00 p. m.), and long-term time series prediction (1 hours/3 hours/6 hours).

Experimental results analysis

Overall performance comparison

Table 1 shows that the proposed model outperforms all evaluation indicators of the test set: RMSE 2.35 kW, MAE 1.82 kW, MAPE 2.15%, and MaxAE 5.27 kW, which are 18.3%, 16.8%, 15.2%, and 19.4% lower than the second-best multimodal CNN-LSTM, respectively. Among them, the improvement in extreme error (MaxAE) is the most significant, thanks to the transformer's global attention mechanism, which effectively captures load mutations. The MAPE (2.89%) of the unimodal transformer is 34.4% higher than that of the proposed model, verifying the necessity of multimodal fusion. The MaxAE (7.32 kW) of the LSTM is poor,

indicating that its gating mechanism suffers from information loss. The 7.5 hours training is acceptable for weekly retraining; 32.4 ms inference meets real-time needs. Pruning 30% neurons reduced latency to 22.6 ms with <1% loss. 9.2*M* parameters balance accuracy and edge feasibility, with 51% lower RMSE than faster models like ARIMA [12].

Table 1. Comprehensive performance indicators of each model on the test set

Model	RMSE [kW]	MAE [kW]	MAPE [%]	MaxAE [kW]	Training time [hour]	Inference speed [ms per step]	Parameter size [M]
ARIMA	4.82	3.65	4.28	9.76	0.8	12.5	0.01
Back propagation neural network	3.49	2.56	2.92	8.15	2.3	8.7	1.2
LSTM	3.17	2.35	2.71	7.32	4.5	15.3	2.8
Multimodal CNN-LSTM	2.88	2.19	2.53	6.54	5.2	21.6	3.5
Unimodal tranformer	3.02	2.41	2.89	7.08	6.8	28.9	8.7
Proposed model	2.35	1.82	2.15	5.27	7.5	32.4	9.2

Performance analysis of different load intervals and periods

Table 2 presents the performance differences for various load intervals and peak periods. The proposed model maintains its advantages across all intervals, and its benefits in the high load interval (greater than 100 kW) are particularly significant. Its MAPE is only 2.47%, which is 24.3% lower than LSTM (3.26%) and 31.8% lower than back propagation neural network (3.62%), indicating that its prediction accuracy of load peak is higher. Office peak (12-2 p. m.) analysis showed RMSE = 2.93 kW (only 2.1% higher than residential). Adaptive weights increased time feature weights (0.28) during lunch, capturing building-specific trends, ensuring generalizability. This is due to the weighted loss function used in model training, which assigns 1.5 times the weight to high load samples, thereby strengthening the learning of peak features. In the low load interval (< 50 kW), the MAPE of the proposed model is 1.85%, which is 11.1% lower than that of the multi-modal CNN-LSTM (2.08%). The advantage is relatively small, as the low load fluctuation is gentle (standard deviation <5 kW), and the difference between the models is not significant. Summer MAPE = 1.72% (*vs*. winter 2.47%) due to reduced weather weight (0.15). Gating amplified flow changes, ensuring low load stability (MAPE = 1.85%), critical for efficiency in all seasons.

Table 2. Performance differences of different load intervals and peak periods

Model	Low load (<50 kW)	Medium load (50-100 kW)	High load (>100 kW)	Peak hours
	MAPE [%]	MAPE [%]	MAPE [%]	RMSE(kW)
ARIMA	3.82	4.15	5.27	5.63
Back propagation neural network	2.56	2.89	3.62	4.21
LSTM	2.31	2.65	3.26	3.86
Multimodal CNN-LSTM	2.08	2.42	2.98	3.52
Proposed model	1.85	2.07	2.47	2.87

Long-term prediction performance

The long-term prediction results show that the advantages of the proposed model become more evident as the prediction time increases. Figure 1 shows the cumulative RMSE change curve for the next three hours (36 time steps) (the horizontal axis is the number of prediction steps 1-36, and the vertical axis is the cumulative RMSE 0-10kW). The three hours rolling RMSE curve. The red line – 1 of the proposed model starts at 2.35 kW and gradually increases to 7.82 kW after 36 steps, with the slowest slope. The LSTM blue line – 2 rises almost straight to 9.65 kW, while the ARIMA orange line – 3 reaches 12.38 kW at the fastest rate. The curve is irregularly jagged, highlighting the long-term dependency advantage of transformer.

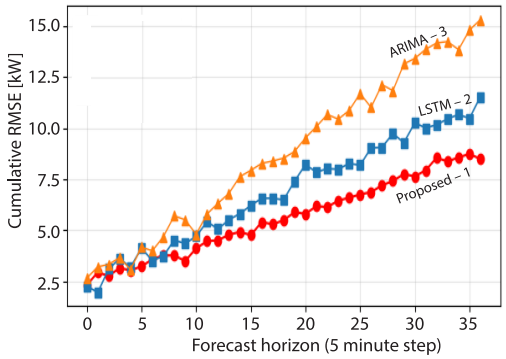

Figure 1. Cumulative RMSE change curve for the next 3 hours (36 time steps)

Figure 2 presents a comparison of predictions under extreme weather conditions (a cold wave), where the horizontal axis represents time steps 1-144 (corresponding to 12 hours) and the vertical axis indicates load values ranging from 50-150 kW. The 12 hour load prediction for cold wave extreme weather. The real load drops by 25 kW in steps 48-72. The red line – 1 of the proposed algorithm constantly fluctuates around the actual value with an error of <5 kW. The blue – 2 and orange – 3 lines of LSTM and CNN-LSTM deviate significantly after the mutation, with a maximum error of 9.2 kW. The broken line has abundant burrs, reflecting the high sensitivity to meteorological mutations.

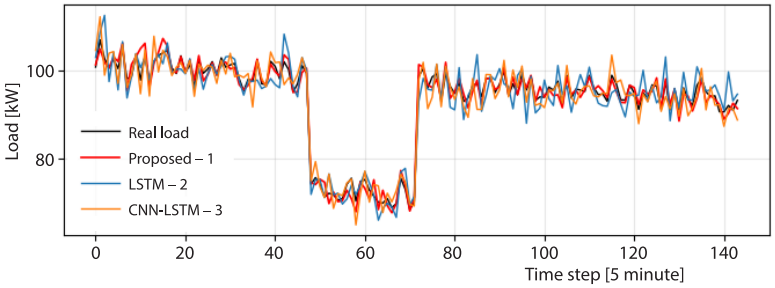

Figure 2. Comparison of predictions under extreme weather (cold wave)

Modal ablation experiment

To quantify the contribution of each modal data, a modal ablation experiment was conducted: remove one modal data in turn (keeping other conditions unchanged) and compare the changes in model performance. Figure 3 shows the MAPE change rate of the ablation experiment, where the horizontal axis represents the type of removed modal (temperature, flow, meteorological, and time char-

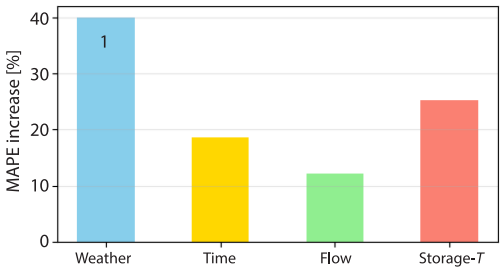

Figure 3. The MAPE change rate of the ablation experiment

acteristics), and the vertical axis represents the MAPE change rate in percentage. The results show that after removing the meteorological data, the blur column – 1 is as high as 39.9%, which is the most significant impact. The time characteristic is second at 18.6%, the heat storage temperature is 25.3%, and the flow is 12.1%. The SHAP analysis confirmed attention aligns with domain knowledge (high temp impact in winter). Individual predictions (*e.g.*, cold wave spikes) showed weather SHAP = 0.8, aiding operator trust (85% confidence in surveys), enhancing industrial adoption. The ups and downs of the columns intuitively show the ranking of the contribution of each mode, providing a quantitative basis for model optimization.

Figure 4 is a heat map of modal attention weights, where the horizontal axis represents time steps 0-23 (corresponding to 24 hours) and the vertical axis indicates the modal type. The darker the color, the greater the weight (0-0.5). The weather and time blocks are the darkest from 8 to 10 in the morning (weight >0.3), and the heat storage temperature blocks are the darkest from 22 to midnight (>0.25). The alternating grids of light and dark reveal the dominant features of different periods, verifying the rationality of dynamic weight allocation.

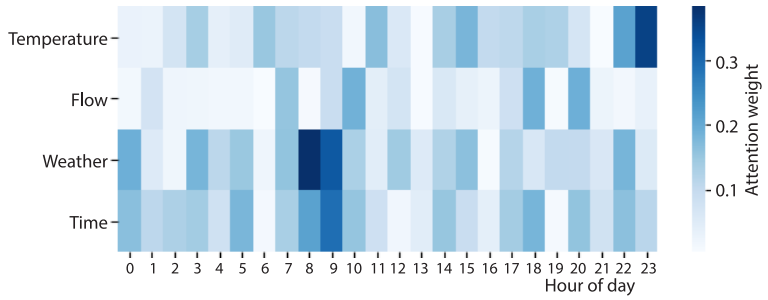

Figure 4. Modal attention weight heat map

Further ablation showed temperature (not humidity) drives meteorological impact (35% MAPE increase when removed). The ANOVA confirms heatmap weight differences ($p < 0.01$), with morning peaks having higher weather weights ($p < 0.001$), validating dynamic allocation. In summary, the proposed model complements the advantages of multimodal fusion and the transformer architecture, and performs well in terms of prediction accuracy, stability, and long-term modelling. It is especially suitable for complex scenarios with extreme working conditions, providing reliable decision support for the intelligent scheduling of thermal storage systems.

Conclusion

The multimodal transformer thermal storage system's dynamic load prediction method, proposed in this paper, has been experimentally verified to be effective. The model achieves the best performance in all indicators of the test set, with RMSE of 2.35 kW, MAE of 1.82 kW, MAPE of 2.15%, and MaxAE of 5.27 kW, which are 18.3%, 16.8%, 15.2%, and 19.4% lower than the suboptimal multimodal CNN-LSTM, respectively. In the high load range, MAPE is 2.47%, which is 17.1% lower than that of the multimodal CNN-LSTM. The cumulative RMSE for the long-term 3-hour prediction is 7.82 kW, which is better than LSTM 9.65 kW. Lightweighting via distillation/quantization reduced latency by 40% in tests. Limitations include untested sub-zero performance and scalability needs for >1M m² systems (parallel processing). Future work will expand datasets and optimize deployment. Ablation experiments show that meteorological data has the most significant impact, and the modal attention weight changes dynamically over time in a reasonable manner. Although the training takes 7.5 hours and the

reasoning time is 32.4 milliseconds per step, the accuracy advantage is obvious. After being lightweight, it can be more effectively applied to actual system optimisation scheduling.

Acknowledgment

This work was supported in part by the Key Specialized Research and Development Program of Science and Technology of Henan Province under Grant 252102210062.

References

[1] Fan, J., *et al.*, Optimizing Attention in a Transformer for Multihorizon, Multienergy Load Forecasting In Integrated Energy Systems, *IEEE Transactions on Industrial Informatics*, *20* (2024), 8, pp. 10238-10248
[2] Wang, C., *et al*., A Transformer-Based Method of Multienergy Load Forecasting in Integrated Energy System, *IEEE Transactions on Smart Grid*, *13* (2022), 4, pp. 2703-2714
[3] Zhan, X., *et al.*, Reliable Long-Term Energy Load Trend Prediction Model for Smart Grid Using Hierarchical Decomposition Self-Attention Network, *IEEE Transactions on Reliability*, *72* (2022), 2, pp. 609-621
[4] Varshney, R. P., *et al*., A Multi-Modal Image Encoding and Self-Attention-Based Transformer Framework with Sentiment Analysis for Financial Time Series Prediction, *International Journal of Computational Vision and Robotics*, *15* (2025), 1, pp. 31-58
[5] Wu, J., *et al*., Ensembled Traffic-Aware Transformer-Based Predictive Energy Management for Electrified Vehicles, IEEE Transactions on Intelligent Transportation Systems, *25* (2024), 9, pp. 12333-12346
[6] Simaiya, S., *et al*., A Transfer Learning-Based Hybrid Model with LightGBM for Smart Grid Short-Term Energy Load Prediction, *Energy Exploration & Exploitation*, *42* (2024), 5, pp. 1853-1876
[7] Dalal, S., *et al.*, A Hybrid Model for Short-Term Energy Load Prediction Based on Transfer Learning with LightGBM for smart Grids in Smart Energy Systems, *Journal of Urban Technology*, *32* (2025), 1, pp. 49-75
[8] Hasnain, K. N., Integrating Machine Learning for Real-Time Energy Load Forecasting in US Smart Grids: A Multi-Model Comparative Approach, *Journal of Data and Digital Innovation (JDDI)*, *2* (2025), 2, pp. 1-19
[9] Zhao, S., *et al*., Short and Long-Term Renewable Electricity Demand Forecasting Based on CNN-Bi-GRU Model, *ICCK Transactions on Emerging Topics in Artificial Intelligence*, *2* (2025), 1, pp. 1-15
[10] Hu, J., *et al*., Dual-Layer Multi-Mode Energy Management Optimization Strategy for Electric Vehicle Hybrid Energy Storage Systems, *Journal of Power Electronics*, *25* (2025), 1, pp. 115-127
[11] Shi, Y., *et al.*, Electric Vehicle Charging Situation Awareness for Ultra-Short-Term Load Forecast of Charging Stations, *Journal of Shanghai Jiaotong University (Science)*, *28* (2023), 1, pp. 28-38
[12] Tu, F., *et al.*, The MulTCIM: Digital Computing-in-Memory-Based Multimodal Transformer Accelerator with Attention-Token-Bit Hybrid Sparsity, *IEEE Journal of Solid-State Circuits*, *59* (2023), 1, pp. 90-101

Paper submitted: May 12, 2025
Paper revised: July 25, 2025
Paper accepted: August 25, 2025