International Journal of Scientific Engineering and Research (IJSER)
ISSN (Online): 2347-3878
SJIF (2025): 8.036

A Patch TST-Bi Transformer Hybrid Framework for Short-Term Power Load Forecasting

Hanyang Wei

Beijing Polytechnic College, School of Information Engineering, ShiMen Road, Beijing, China
Email: *314310676[at]qq.com*

Abstract: ***This paper proposes a hybrid Patch TST-Bi Transformer framework for short-term power load forecasting to simultaneously capture local temporal patterns and long-range dependencies. Historical sequences are partitioned into temporal patches and encoded using a channel-independent Patch TST backbone, while a Bidirectional Transformer models global contextual relationships across patches. Reversible Instance Normalization and moving average decomposition are incorporated to improve robustness against non-stationarity and enhance long-horizon forecasting performance. Experiments conducted on benchmark and real-world power load datasets demonstrate consistent improvements over Autoformer, Informer, FEDformer, and other baseline models. The proposed framework achieves higher forecasting accuracy while maintaining computational efficiency and robust generalization.***

Keywords: PatchTST; BiTransformer; Power Load Forecasting; Time Series Forecasting; Transformer; Deep Learning; RevIN; Smart Grid

1. Introduction

Power load forecasting is one of the fundamental tasks in power system operation and planning, directly affecting power dispatching, generation scheduling, demand response, and the secure and economical integration of renewable energy resources. With the continuous expansion of modern power systems and the increasing complexity of load characteristics, achieving accurate and reliable short-term as well as medium- and long-term load forecasting has become increasingly important.

In practical applications, power load series generally exhibit multiple characteristics simultaneously, including multivariate dependence, high noise, seasonality, periodicity, abrupt fluctuations, and long-term trends0. Moreover, electricity consumption in industrial and residential sectors is significantly influenced by various heterogeneous external factors, such as weather conditions, holidays, and human activities, making accurate forecasting considerably more challenging. Therefore, forecasting models must not only capture short-term local temporal patterns, such as periodic fluctuations and sudden changes, but also effectively model long-range temporal dependencies across different time scales to satisfy the requirements of modern power systems for forecasting accuracy, real-time performance, and robustness.

2. Related Work

Power load forecasting methods have evolved from traditional statistical models to modern machine learning and deep learning approaches. Conventional statistical methods, such as ARIMA and SARIMA, perform well on regular and periodic load series but have limited capability in modeling nonlinear relationships and complex temporal dynamics [2].

With the increasing complexity of power load data, machine learning methods, including Support Vector Machines (SVM) and Random Forests (RF), have been widely applied due to their ability to capture nonlinear patterns. However, these approaches rely heavily on manual feature engineering and are less effective in learning long-range temporal dependencies [3].

Recently, deep learning models, such as CNNs, LSTMs, and Transformer-based architectures, have achieved remarkable performance by automatically extracting temporal features from historical observations[4]. Nevertheless, existing methods still face challenges in simultaneously capturing local temporal patterns, long-range dependencies, and multi-scale characteristics, resulting in limited forecasting performance under complex operating conditions[5] . These limitations motivate the development of more effective hybrid architectures for power load forecasting. To overcome the aforementioned challenges, this paper proposes a hybrid forecasting architecture that combines PatchTST with a Bidirectional Transformer (BiTransformer)

1) A hybrid PatchTST-BiTransformer framework is proposed, where PatchTST extracts local temporal features from patch-wise input sequences and the Bidirectional Transformer captures global contextual dependencies to improve long-range forecasting performance.
2) Reversible Instance Normalization (RevIN) and moving average decomposition are incorporated to reduce the influence of data non-stationarity and enhance model robustness and long-horizon forecasting accuracy.
3) Extensive experiments on public benchmark and real-world power load datasets demonstrate that the proposed method outperforms several state-of-the-art forecasting models, while ablation studies verify the effectiveness of each module.

3. Methodology

3.1 Overall Framework

The proposed PatchTST-BiTransformer framework is designed to jointly capture local temporal patterns and long-range contextual dependencies in power load time series. By combining the efficient local representation

capability of PatchTST with the global modeling ability of a Bidirectional Transformer (BiTransformer), the proposed architecture provides a unified solution for multi-scale feature extraction while maintaining computational efficiency. The overall architecture of the proposed model is illustrated in Figure 1.

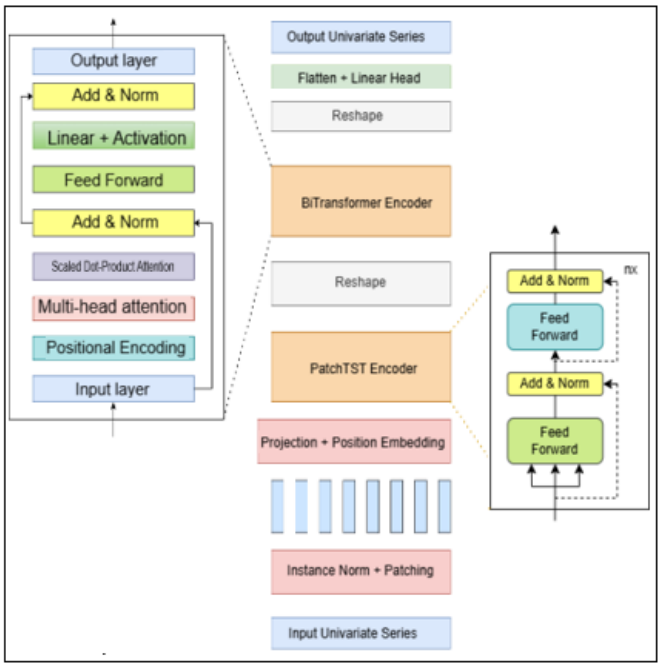

Figure 1: Overall Architecture of the Proposed PatchTST-BiTransformer

The forecasting process consists of four successive stages: robust data preprocessing, local feature extraction, global context modeling, and prediction generation. Given a historical multivariate load sequence, the input data are first processed through a reversible normalization module to alleviate the influence of distribution shifts across different samples. Meanwhile, moving average decomposition is employed to separate long-term trends from short-term fluctuations, enabling the model to learn different temporal characteristics more effectively.

After preprocessing, the normalized sequence is partitioned into a series of fixed-length temporal patches using a sliding-window strategy. Instead of directly processing individual time points, the model regards each patch as a basic modeling unit, which significantly shortens the input sequence while preserving local temporal continuity. The generated patches are subsequently projected into a latent embedding space and encoded by a channel-independent PatchTST backbone to extract discriminative local representations.

To further enhance the modeling capability of long-range dependencies, the local representations generated by PatchTST are fed into a Bidirectional Transformer. Unlike conventional autoregressive Transformers, the BiTransformer adopts a non-causal self-attention mechanism, allowing every patch to interact with all other patches within the input sequence. Consequently, both historical and future contextual information inside the observation window can be fully exploited to model periodic patterns and global temporal correlations.

Finally, the globally encoded features are transformed into the target forecasting horizon through a lightweight prediction head. The predicted sequence is restored to the original numerical scale using the inverse operation of reversible normalization, producing the final power load forecasting results. Through the collaborative learning of local and global temporal information, the proposed framework effectively improves forecasting accuracy and robustness under complex operating conditions.

3.2 RevIN and Series Decomposition

Power load data usually exhibit strong non-stationarity caused by seasonal variations, weather conditions, holidays, and operational uncertainties. These factors often result in significant distribution shifts across different time periods, which may reduce the stability and generalization capability of deep learning models. To alleviate these issues, a robust preprocessing module is introduced before feature extraction. The proposed preprocessing consists of two complementary components: Reversible Instance Normalization (RevIN) and moving average series decomposition.

RevIN is first applied to normalize each input sequence independently. Unlike conventional normalization methods that estimate global statistics over the entire dataset, RevIN computes sequence-specific statistics and performs reversible affine normalization. This strategy effectively mitigates distribution discrepancies among different samples while preserving the intrinsic temporal characteristics of each sequence. The normalization process is defined by Equation (1).

$$X_{norm}=RevIN_{norm}(X)=\frac{X-\mu x}{\sigma x+\epsilon}\odot\gamma+\beta(2-X) \quad (1)$$

A key advantage of RevIN is its reversible design. During inference, the normalization statistics recorded in the preprocessing stage are reused to recover the predicted values to their original numerical scale. Consequently, the forecasting model benefits from normalized feature learning without sacrificing the physical interpretability of the prediction results.

Although normalization improves training stability, the normalized sequence still contains both slowly varying trends and rapidly changing fluctuations. Directly modeling these heterogeneous temporal patterns may increase the learning difficulty of the forecasting network. Therefore, a moving average decomposition strategy is further introduced to separate the normalized sequence into trend and residual components. The decomposition process is described by Equation (2).

$$X_{norm}=T+R \quad (2)$$

The trend component represents long-term evolution in electricity demand, while the residual component mainly reflects local variations and short-term fluctuations. By explicitly separating these two temporal characteristics, the subsequent feature extraction module can focus on learning dynamic local patterns from the residual sequence while preserving the underlying long-term trend information. The

final forecasting result is obtained by combining the predicted trend and residual components before the inverse normalization process.

The combination of RevIN and series decomposition enhances the robustness of the proposed framework against distribution shifts and temporal non-stationarity. More importantly, this preprocessing strategy provides cleaner and more stable feature representations for the subsequent PatchTST encoder, thereby improving both optimization efficiency and forecasting accuracy under diverse operating conditions.

3.3 Local Representation Learning Using PatchTST

Following the preprocessing stage, the normalized sequence is transformed into local feature representations using the PatchTST encoder. Instead of directly processing point-wise observations, PatchTST divides the original time series into a sequence of fixed-length temporal patches, allowing the model to learn meaningful local patterns while significantly reducing the computational complexity associated with long input sequences. The overall architecture of the PatchTST module is illustrated in Figure 2.

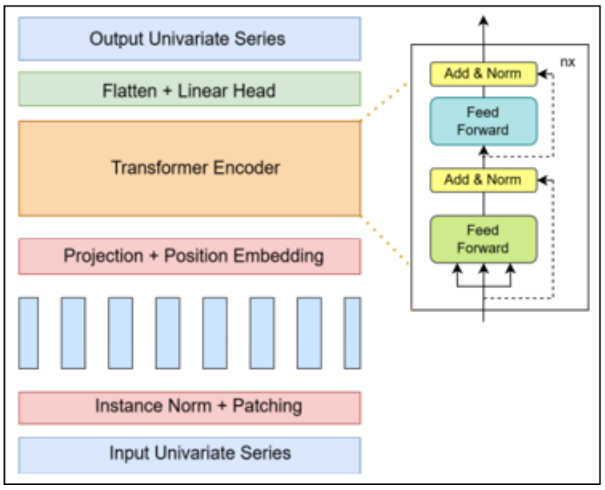

Figure 2: PatchTST Model Architecture

Given an input sequence with length L, a sliding-window strategy is adopted to partition the sequence into multiple overlapping or non-overlapping patches. Let P denote the patch length and S represent the stride between two adjacent patches. The total number of generated patches is determined according to Equation (3).

$$N=\left\lfloor\frac{(L-P)}{S}\right\rfloor+2 \tag{3}$$

Compared with conventional point-level tokenization, patch-based representation preserves local temporal continuity by grouping neighboring observations into a unified semantic unit. Consequently, each patch contains richer contextual information than an individual sampling point, enabling the model to capture short-term load fluctuations and periodic variations more effectively. Moreover, converting a long sequence into a much shorter patch sequence substantially decreases the computational burden of self-attention, making the proposed framework more suitable for long-horizon power load forecasting.

After patch partitioning, every temporal patch is projected into a latent feature space through a learnable linear projection layer to generate patch embeddings. This transformation converts the original numerical observations into token-like representations that can be efficiently processed by the Transformer encoder. The embedding operation is determined according to Equation (4).

$$E_c=\text{LinearProj}(P_c)\in R^{N\times D_{model}} \tag{4}$$

Unlike conventional multivariate Transformers that jointly process all variables, PatchTST adopts a channel-independent learning strategy. Each feature channel is encoded separately while sharing the same network parameters across all variables. This design enables the encoder to focus exclusively on temporal dependency modeling within individual variables, thereby reducing interference caused by heterogeneous feature distributions. At the same time, parameter sharing effectively controls model complexity and improves generalization capability when dealing with high-dimensional multivariate load data.

The embedded patches are subsequently processed by stacked Transformer encoder layers to learn local temporal representations. Each encoder layer consists of a standard multi-head self-attention module followed by a feed-forward network with residual connections and layer normalization. Rather than modeling relationships among individual sampling points, self-attention is performed on the patch sequence, allowing neighboring patches to exchange contextual information efficiently. Consequently, the encoder is capable of capturing short-term fluctuations, local periodic patterns, and temporal transitions that frequently appear in electricity load series.

The hierarchical representation learned by the PatchTST encoder serves as the local feature extractor of the proposed framework. Through patch-level modeling, the encoder effectively preserves fine-grained temporal dynamics while maintaining computational efficiency for long input sequences. However, because the channel-independent architecture primarily emphasizes local temporal evolution within each variable, its capability to model long-range contextual interactions across distant patches remains limited. To further exploit global temporal dependencies, the extracted local representations are subsequently forwarded to the BiTransformer encoder for global context modeling.

3.4 Global Context Modeling and Forecast Generation

Although the PatchTST encoder effectively captures local temporal dynamics through patch-based representation learning, its channel-independent design mainly focuses on intra-variable temporal evolution and provides limited capability for modeling global contextual relationships among distant patches. To overcome this limitation, a Bidirectional Transformer (BiTransformer) is incorporated as the second-stage encoder to further exploit long-range temporal dependencies.

The local representations generated by the PatchTST encoder are first reshaped and reorganized as the input sequence of the BiTransformer. The overall architecture of the global

context encoder is illustrated in Figure 3. Unlike conventional autoregressive Transformers, the proposed BiTransformer removes the causal attention constraint and performs non-causal self-attention over the entire patch sequence. Consequently, each patch can interact with every other patch within the observation window, allowing the model to simultaneously utilize historical and future contextual information during feature extraction.

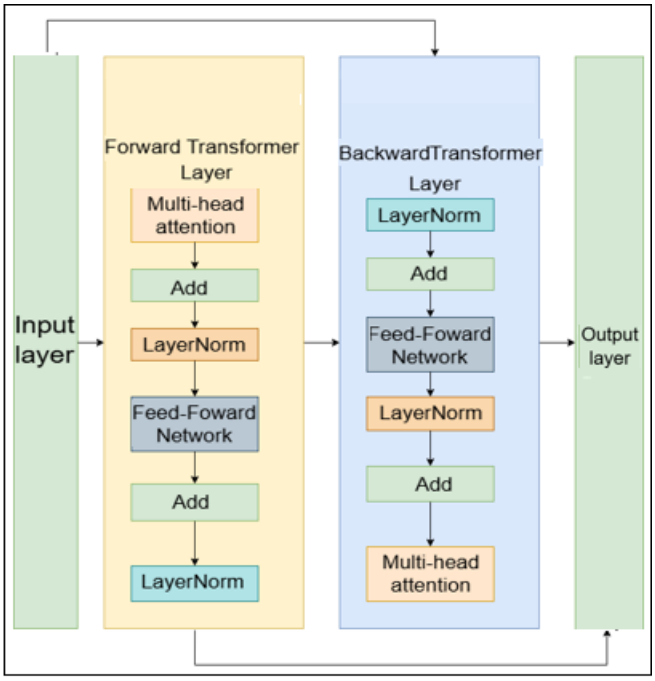

Figure 3: BiTransformer Architecture

The attention mechanism follows the standard scaled dot-product formulation shown in Equation (5).

$$\text{Attention}(Q,K,V)=\text{Softmax}(\frac{Q(K)^T}{\sqrt{d_K}})V \quad (5)$$

The bidirectional attention mechanism enables the encoder to establish global relationships among local representations extracted from different temporal locations. Such a design is particularly suitable for power load forecasting, where electricity demand is strongly affected by recurring daily and weekly cycles, seasonal variations, and long-range temporal dependencies. By integrating contextual information from the entire observation window, the BiTransformer effectively complements the local modeling capability of PatchTST and generates more informative feature representations for subsequent prediction.

After global context encoding, the output features are flattened into a one-dimensional representation before being projected to the target forecasting horizon through a linear prediction head. The flattening operation is described by Equation (6), while the forecasting process is performed according to Equation (7).

$$F=\text{Flatten}(H_{global})\in R^{1\times D_{flat}} \quad (6)$$

$$\hat{Y}_{norm}^{flat}=W_{out}F^T+b_{out}\ (2\text{-}B)\in R^{\hat{L}\times C} \quad (7)$$

Compared with multi-stage decoding strategies adopted by conventional sequence-to-sequence forecasting models, the lightweight linear prediction head directly maps the encoded global representations to the future load sequence, reducing model complexity while maintaining efficient end-to-end optimization. This design also minimizes error accumulation during multi-step forecasting and improves computational efficiency for practical applications.

Finally, the predicted sequence is transformed back to its original numerical scale through the inverse operation of RevIN using the statistical information recorded during preprocessing. The denormalization process is expressed by Equation (8).

$$\hat{Y}=\text{RevIn}_{denorm}(\hat{Y}_{norm})=\left(\frac{\hat{Y}_{norm}-\beta}{\gamma}\right)\odot(\sigma x+\epsilon)+\mu x \quad (8)$$

The reversible transformation guarantees that the final forecasting results remain consistent with the physical magnitude of the original power load data while preserving the numerical stability achieved during model training. By combining PatchTST-based local feature extraction with BiTransformer-based global dependency modeling, the proposed framework effectively integrates fine-grained temporal information and long-range contextual knowledge within a unified architecture. The lightweight prediction head and reversible denormalization further enhance computational efficiency and forecasting robustness, making the proposed model suitable for short-term power load forecasting under complex operating conditions.

4. Experiments and Results Analysis

4.1 Datasets

4.1.1 Dataset Description

To evaluate the proposed PatchTST-BiTransformer framework, experiments were conducted on the Panama Electricity Load Dataset, a widely used benchmark for short-term power load forecasting. The dataset records the hourly electricity demand of Panama from January 2015 to June 2020 and includes multiple meteorological variables, such as temperature, humidity, and wind speed, together with calendar information including holidays and school schedules.

Before training, the data were preprocessed to remove abnormal records and arranged in chronological order. Following the standard practice in time series forecasting, the dataset was sequentially divided into training, validation, and testing subsets to prevent information leakage. The training set was used for model optimization, the validation set for hyperparameter tuning, and the testing set for final performance evaluation.

The rich temporal patterns and diverse exogenous variables make this dataset suitable for evaluating the proposed framework in practical short-term power load forecasting tasks.

4.1.2 Experimental Environment

The proposed model was implemented using PyTorch and trained with the Adam optimizer. Mean Squared Error (MSE) was adopted as the loss function to optimize

forecasting accuracy. The initial learning rate was set to 1×10^{-4}, and the batch size was fixed at 32. An early stopping strategy with a patience of 15 epochs was employed to prevent overfitting.

For all experiments, the input sequence length was set to 168 time steps, while the forecasting horizon was 96 time steps. The historical sequence was partitioned into temporal patches and processed by the PatchTST encoder, followed by the BiTransformer module for global context modeling. The final forecasting results were generated through a linear prediction head and restored to the original scale using the inverse operation of RevIN.

4.2 Evaluation Metrics

To comprehensively evaluate forecasting performance, three widely used evaluation metrics are adopted, namely Mean Squared Error (MSE), Mean Absolute Error (MAE), and Relative Squared Error (RSE).

Mean Squared Error (MSE) measures the average squared difference between the predicted values and the corresponding ground-truth observations. Since the prediction errors are squared, MSE imposes a larger penalty on significant forecasting deviations, making it particularly sensitive to large errors.

The mathematical definition of MSE is given in Equation (9), where N denotes the number of samples, y_i represents the ground-truth value, and $\hat{y}_i$ denotes the corresponding prediction.

$$MSE=\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2 \quad (9)$$

Mean Absolute Error (MAE) calculates the average absolute difference between predicted values and actual observations.

Unlike MSE, MAE maintains the same physical unit as the original data, providing better interpretability while exhibiting lower sensitivity to outliers. The formulation of MAE is presented in Equation (10).

$$MAE=\frac{1}{N}\sum_{i=1}^{N}|y_i-\hat{y}_i| \quad (10)$$

Relative Squared Error (RSE) is a dimensionless evaluation metric that compares the forecasting error of a prediction model with that of a simple baseline predictor.

An RSE value smaller than one indicates that the forecasting model outperforms the baseline model based on the mean value of the observations. The corresponding formulation is presented in Equation (11), where $\bar{y}$ denotes the average value of the ground-truth observations.

$$RSE=\frac{\sqrt{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}}{\sum_{i=1}^{N}(y_i-\bar{y})^2} \quad (11)$$

4.3 Comparative Experiments and Results Analysis

To ensure a fair comparison, three representative Transformer-based forecasting models, namely Autoformer, Informer, and FEDformer, were selected as baseline methods. All baseline models were implemented following the optimal parameter settings reported in their original publications and were trained using the same input sequence length and forecasting horizon described in Section 3.4. Model performance was consistently evaluated using three metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), and Relative Squared Error (RSE).

The experimental results on the Panama electricity load dataset are presented in Table 1.

Table 1: Experimental Results on the Panama Electricity Load Dataset

Dataset	*Panama Electricity Load Dataset*		
Evaluation Metric	*MSE*	*MAE*	*RSE*
PatchTST+BiTransformer(ours)	0. 249	0. 398	0. 499
Autoformer	0. 288	0. 415	0. 513
Informer	0. 381	0. 454	0. 521
FEDformer	0. 289	0. 416	0. 521

Despite the presence of numerous exogenous variables, including meteorological observations and calendar information, the proposed PatchTST-BiTransformer model still achieves the best overall forecasting performance among all compared methods.

The superior performance can be attributed to the strong representation capability of the proposed framework, which effectively integrates intrinsic temporal patterns with external driving factors, enabling more accurate prediction of national-scale electricity demand.

Figure 4 illustrates representative forecasting results obtained on the Panama Electricity Load Dataset. In the figure, the horizontal axis represents the forecasting horizon, while the vertical axis denotes the normalized load value. The blue curve corresponds to the ground-truth observations, whereas the orange curve represents the predictions generated by the proposed model.

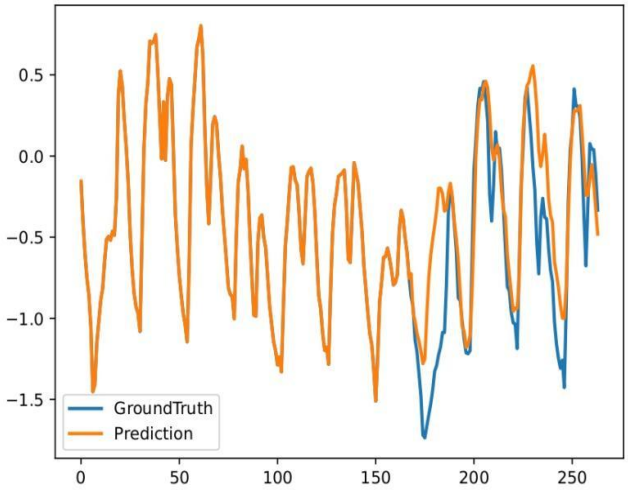

Figure 4: Comparison of Forecasting Results on the Panama Electricity Load Dataset

Overall, the predicted curves closely follow the actual load variations across different datasets, demonstrating the proposed model's capability to accurately capture both local fluctuations and long-term temporal trends.

4.4 Ablation Study

To quantitatively evaluate the contribution of the two key components of the proposed hybrid architecture, namely the PatchTST local encoder and the BiTransformer global encoder, comprehensive ablation studies were conducted.

Three model variants were considered for comparison:

- BiTransformer-only: Only the Bidirectional Transformer is retained, while the PatchTST module is removed.
- PatchTST-only: Only the PatchTST encoder is employed, without incorporating the BiTransformer module.
- PatchTST-BiTransformer: The complete hybrid framework integrating both local feature extraction and global contextual modeling.

Table 2: Ablation Study Results on the Panama Electricity Load Dataset

Dataset	*Panama Electricity Load Dataset*		
Evaluation Metric	*MSE*	*MAE*	*RSE*
PatchTST+BiTransformer(ours)	0. 249	0. 398	0. 499
PatchTST-only	0. 259	0. 401	0. 502
BiTransformer-only	0. 261	0. 404	0. 510

The complete model achieves the best overall forecasting accuracy, demonstrating that both the PatchTST encoder and the BiTransformer encoder contribute positively to the final prediction performance even in scenarios involving abundant external influencing factors.

Overall, the results of the ablation study clearly demonstrate the effectiveness of the proposed hybrid architecture. Although the PatchTST encoder is capable of extracting informative local temporal patterns, its channel-independent design limits its ability to capture global contextual relationships across patches and feature channels. By introducing bidirectional self-attention along the patch dimension, the BiTransformer effectively models long-range temporal dependencies and injects global contextual information into the forecasting process.

Consequently, the integration of the two modules significantly improves forecasting accuracy and consistently outperforms either individual component across all benchmark datasets. These findings verify that local feature extraction and global dependency modeling complement each other, leading to a more robust and accurate power load forecasting framework.

5. Conclusion

This study presents a hybrid PatchTST-BiTransformer framework for short-term power load forecasting that effectively combines local temporal representation learning with global contextual dependency modeling. By integrating RevIN and moving average decomposition, the proposed framework improves robustness under non-stationary conditions while maintaining efficient computation. Experimental results demonstrate superior forecasting performance compared with representative state-of-the-art methods, and the ablation study confirms the contribution of each model component. Overall, the proposed framework provides an effective and practical solution for accurate power load forecasting and offers a promising foundation for future research on intelligent energy forecasting systems.

References

[1] Xavier S, Marco B, Jean-Michel C, et al. A new interval prediction methodology for short-term electric load forecasting based on pattern recognition[J]. Applied Energy, 2021, 297 DOI:10. 1016/J. APENERGY. 2021. 117173.

[2] Jinran W, You-Gan W, Yu-Chu T, et al. Support vector regression with asymmetric loss for optimal electric load forecasting[J]. Energy, 2021, 223 DOI:10. 1016/J. ENERGY. 2021. 119969.

[3] Yuanyuan C, Peiyong D, Junqing L. Hourly electric load forecasting for buildings using hybrid intelligent modelling[J]. IOP Conference Series: Earth and Environmental Science,2021,669 (1):012022-. DOI:10. 1088/1755-1315/669/1/012022.

[4] Dongyeon J, Chiwoo P, Myoung Y K. Short-term electric load forecasting for buildings using logistic mixture vector autoregressive model with curve registration[J]. Applied Energy,2021,282 (PB): DOI:10. 1016/j. apenergy. 2020. 116249.

[5] Hariharan R, Kumar K M. A Research on Electric Load Forecasting Factors Effecting and Methods Involved[J]. International Journal of Innovative Technology and Exploring Engineering (IJITEE),2019,8 (12):1462-1466.