A Hybrid Transformer-MLP Approach for Short-Term Electric Load Forecasting

Tuan Anh Nguyen [1*], Thanh Ngoc Tran [2]
[1, 2] Faculty of Electrical Engineering Technology, Industrial University of Ho Chi Minh City, Ho Chi Minh City, Viet Nam
Email: [1] nguyenanhtuan@iuh.edu.vn, [2] tranthanhngoc@iuh.edu.vn
*Corresponding Author

Abstract—**Short-term electric load forecasting plays a vital role in ensuring the stability and efficiency of smart grid operations. However, accurately predicting demand remains challenging due to nonlinearity, volatility, and long-term temporal dependencies in consumption patterns. The research proposes a lightweight hybrid deep learning model that integrates a Transformer encoder with a multi-layer perceptron (MLP) to enhance prediction accuracy and robustness for short-term load forecasting. The proposed model employs a Transformer to extract long-range temporal features through self-attention mechanisms, while the MLP captures complex nonlinear mappings at the output stage. A real-world electricity load dataset collected from three Australian states (NSW, QLD, VIC) between 2009 and 2014 is used for evaluation. To assess model performance, mean absolute percentage error (MAPE), mean squared error (MSE), and Root Mean Squared Error (RMSE) are used. Experimental results demonstrate that the proposed transformer-MLP model consistently achieves the lowest forecasting error across all regions. MAPE ranges from 0.69% to 0.95%, outperforming standard deep learning models, including LSTM, CNN, and MLP. Despite its shallow architecture and reduced computational complexity, the hybrid model effectively captures both temporal dependencies and nonlinear variations. This study provides a practical, deployable forecasting solution for smart grids. Future work will extend the model to multi-step forecasting, incorporate exogenous variables such as weather and calendar effects, and explore deeper Transformer variants further to enhance prediction accuracy and generalization across diverse load conditions.**

Keywords—Short-Term Load Forecasting; Hybrid Deep Learning Model; Transformer; MLP.

I. Introduction

Electric load forecasting plays a pivotal role in the operation and management of power systems, enabling operators and electricity providers to plan for generation, distribution, and consumption effectively. Accurate forecasting of future electricity demand supports optimal resource utilization, reduces operational costs, and enhances grid stability. With the increasing demand for electricity and the rapid integration of renewable energy sources and demand-side management strategies, load forecasting has become increasingly complex. Forecasting errors may lead to supply-demand imbalances, resulting in electricity shortages or energy waste, thus negatively impacting overall system efficiency.

Historically, electric load forecasting relied on linear statistical methods such as linear regression, moving average, and autoregressive integrated moving average (ARIMA) [1]-[10]. While these approaches provide satisfactory performance on stable and stationary time series data, they are inadequate for modeling modern power systems' nonlinear and highly dynamic nature.

To overcome these limitations, machine learning techniques such as decision tree regression [11]-[14], support vector regression (SVR) [15]-[23], random forest, and K-nearest neighbors (KNN) [24]-[27] have been widely adopted. These models offer greater flexibility in capturing nonlinear relationships and integrating high-dimensional inputs. However, they often fall short in effectively modeling temporal dependencies in time series data, which are crucial for load forecasting.

Recent advancements in deep learning have introduced a new paradigm for load forecasting. Architectures such as multilayer perceptron (MLP) [28]-[33], recurrent neural networks (RNN) [34]-[45], long short-term memory (LSTM) [46]-[59], gated recurrent unit (GRU) [60]-[67], and convolutional neural networks (CNN) [68]-[79] have demonstrated superior performance in learning complex patterns and modeling long temporal sequences. LSTM and GRU effectively capture long-term dependencies via controlled memory mechanisms, while CNN is known for extracting local temporal features efficiently. Nevertheless, deep learning models often demand significant computational resources, long training times, and are prone to overfitting, especially when data is noisy or insufficiently diverse.

To address these challenges, hybrid models have emerged, combining the strengths of multiple architectures to improve forecasting accuracy and model robustness. A notable example is integrating the transformer architecture—initially developed for natural language processing—with traditional structures like MLP. The Transformer excels in modeling long-range dependencies via the Self-Attention mechanism [80]-[87], and its ability to process sequences in parallel significantly reduces training time compared to sequential models like LSTM. Meanwhile, MLP enhances nonlinear mapping and feature integration capabilities at the output stage.

In this context, the motivation for selecting the transformer-MLP [88]-[95] architecture lies in its ability to combine the advantages of both components: the Transformer captures global temporal patterns effectively through attention mechanisms. At the same time, MLP complements this by enhancing local nonlinear interactions and stabilizing the output layer. This design ensures that both long-term dependencies and short-term variations in load profiles are well represented, which is critical for real-world electricity consumption data that exhibit both trends and volatility.

The research contribution is developing a lightweight yet accurate hybrid forecasting model that demonstrates high generalization capability across different load conditions while maintaining computational efficiency. This is particularly relevant for innovative grid systems requiring real-time responses and adaptable deployment in varying scenarios.

This study aims to develop and evaluate the effectiveness of a Transformer-MLP hybrid model for electric load forecasting. The proposed model leverages the Transformer's ability to capture long-term temporal dependencies and the MLP's strength in nonlinear data representation, thereby improving the accuracy and adaptability of load forecasting. The model's performance is empirically compared with standard deep learning models on real-world load datasets, including MLP, LSTM, and CNN. Results indicate that the proposed hybrid model achieves superior forecasting performance, with significantly lower MAE, RMSE, and MAPE values. This research contributes theoretically and provides a practical forecasting solution for modern power systems, particularly in the transition toward renewable energy and the growing need for flexible load management.

II. Theoretical Basis

A. MLP Model

Multi-layer perceptron (MLP) networks, a subclass of artificial neural networks within the broader domain of machine learning, are widely utilized for classification and regression tasks. As a Deep Neural Network (DNN) family member, MLPs can learn and represent complex nonlinear relationships, offering substantial advantages in addressing high-dimensional and nonlinear problems.

An MLP consists of multiple layers of interconnected neurons, organized into three main components: the input layer, one or more hidden layers, and the output layer.

Input layer: This layer is responsible for receiving raw input data and forwarding it to subsequent layers in the network. Each neuron within the input layer corresponds to a specific feature in the dataset.

Hidden layers: MLPs may include one or several hidden layers, where each neuron is connected to neurons in both the preceding and succeeding layers. These layers are critical in extracting hierarchical and abstract patterns from the data. The input to each neuron in a hidden layer is typically a weighted linear combination of outputs from the previous layer, followed by applying an activation function to introduce nonlinearity.

Output layer: This layer delivers the final predictions the network generates. In classification tasks, the number of output neurons corresponds to the number of target classes. In contrast, for regression problems, the output layer generally comprises a single neuron that produces a continuous value.

Overall, the MLP architecture enables robust function approximation and pattern recognition capabilities, making it a foundational model in deep learning applications. The operation process of a Multi-Layer Perceptron (MLP) can be divided into two main stages. Forward propagation begins with input data being fed into the input layer. At each layer, the input is transmitted through the neurons by linearly combining the outputs from the previous layer with corresponding weights and biases. The result is then passed through an activation function to produce the output. The input to each neuron is computed as (1):

$$z^l = W^l a^{(l-1)} + b^l \tag{1}$$

In which, z^l is the Linear combination between the pre-layer output and weights, plus bias; W^l is the weighted matrix between class-1 and class l; a^{l-1} is the output of class -1 l; b^l is the bias of the class l.

After the calculation, we use the σ trigger function to calculate the output of the class: $z^l l$.

$$a^l = \sigma(z^l) \tag{2}$$

In which a^l is the output of the second class, which is the result after applying the trigger function l and σ is the trigger function.

The output of this layer is relayed through the successive layers up to the network's output layer.

Backpropagation: Backpropagation is an optimization method for updating the network weights after calculating the error. The steps in the backpropagation process include:

Calculation of errors at the output layer: Errors at the output layer are calculated using (3),

$$Loss = L(\hat{y}, y) \tag{3}$$

In which, $\hat{y}$ is the Predicted output value and y is the Actual value.

Calculate the gradient of the loss function: After calculating the error, the error gradient for the weights and bias will be calculated.

The gradient of the error for the output is calculated as follows:

$$\frac{\partial L}{\partial a^l} = \frac{2}{m}(\hat{y} - y) \tag{4}$$

In which $\frac{\partial L}{\partial a^l}$ is the derivative of the loss function L for class output l, m is the number of samples in the training dataset, $\hat{y}$ is the Predicted output value, and y is the Actual value.

Gradient for calculation via trigger function: z^l.

$$\frac{\partial L}{\partial z^l} = \frac{\partial L}{\partial a^l} \sigma' z^l \tag{5}$$

In which, $\frac{\partial L}{\partial z^l}$ is the gradient of the loss function for pure input z^l and $\frac{\partial L}{\partial a^l}$ is the gradient of the loss function to the output of the class l.

Updated weights and biases: Weights and biases will be updated according to the Gradient Descent formula:

$$W^l = W^l - \eta \frac{\partial L}{\partial W^l} \tag{6}$$

In which, W^l is the Weighted matrix in the second class l, η is the Learning rate, and $\eta \frac{\partial L}{\partial W^l}$ is the derivative (gradient) of the loss function L by weight W^l.

$$b^l = b^l - \eta \frac{\partial L}{\partial b^l} \tag{7}$$

In which, η is the speed of learning, b^l is the Bias of the class, making the model more flexible when learning l, $\frac{\partial L}{\partial b^l}$ is the derivative of the loss function L for the bias at the class l.

This process is repeated until the desired accuracy is achieved.

Activation function: Activation functions are used in MLP to create nonlinearity in the network, helping the network learn the complex relationships between inputs and outputs. Some standard trigger functions:

Sigmoid: The sigmoid function converts the xxx input value into a value in the range (0, 1), This is a nonlinear trigger function, commonly used in binary classification problems.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{8}$$

Hyperbolic Tangent: A tangent function that converts an input value into a value in the range (-1, 1): This function is often used in problems with negative and positive output values.

$$tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{9}$$

Rectified linear unit (ReLU): The ReLU function is a very popular trigger function that preserves positive values and converts all negative values to 0:

$$ReLU(x) = max(0, x) \tag{10}$$

Leaky ReLU: This is a variant of ReLU, with a small portion of the negative value retained to avoid "neuro death":

$$Leaky\ ReLU(x) = max(\alpha x, x) \tag{11}$$

Loss Function: The loss function measures predicted and actual values. Standard loss functions in MLP are:

Cross-Entropy Loss: This is a common loss function in classification problems, calculated as follows:

$$\mathcal{L} = -\sum_{i=1}^{m} y^{(i)} \log(\hat{y}^{(i)}) + (1 - y^{(i)}) \log(1 - y^{(i)}) \tag{12}$$

Mean squared error (MSE): MSE is a common loss function in regression problems, calculated as follows:

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y^{(i)} - \hat{y}^{(i)})^2 \tag{13}$$

Multilayer perceptron (MLP) networks represent a powerful approach in machine learning. They offer practical solutions for classification and regression tasks because they can learn and model nonlinear relationships within data. The mathematical operations within an MLP involve a sequence of computations, including input aggregation, application of activation functions, error calculation, and weight optimization through the backpropagation algorithm. The selection of appropriate activation and loss functions plays a crucial role in enhancing the performance and convergence behavior of the MLP model.

B. *Transformer Model*

The transformer is a deep neural network architecture that Vaswani *et al.* proposed. It is now considered foundational for many modern models in sequence-related tasks, especially in natural language processing and time series forecasting. Unlike traditional sequential models such as RNNs or LSTMs, the transformer does not process data in order. Instead, it utilizes a Self-Attention mechanism to learn dependencies between elements in a sequence regardless of distance. This enables parallel computation and efficient modeling of long-term relationships. In Self-Attention, each input vector is mapped into three vectors: Query (Q), Key (K), and Value (V). The attention mechanism is computed as:

$$Attention\ (Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{14}$$

Where d_k is the dimensionality of the Key vectors. To enhance its capacity to capture diverse patterns, the Transformer employs multiple attention heads in parallel, known as Multi-Head Attention, calculated as:

$$M(Q, K, V) = Concat(head_1, ..., head_h)W^0 \tag{15}$$

Each head is computed independently using the attention formula with separate learned weight matrices. Since the Transformer does not operate on sequence order, it incorporates Positional Encoding to inject position information into the input embeddings, defined as:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \tag{16}$$

$$PE_{(\mathrm{pos,2i/1})} = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \tag{17}$$

These encodings allow the model to capture the relative positions of sequence elements. Each Transformer Encoder block consists of two main components. First is the Multi-Head Self-Attention layer, followed by a residual connection and Layer Normalization, expressed as:

$$LayerNorm\ (x + MultiHead(x)) \qquad (18)$$

Next, the output passes through a feedforward neural network (FFN) with two fully connected layers and a nonlinear activation function (such as ReLU or GELU), represented as:

$$FFN(x) = max(0, xW_1 +\ b_1)\, W_2 + b_2 \qquad (18)$$

which is also followed by a residual connection and normalization:

$$LayerNorm(x + FFN(x)) \qquad (19)$$

Stacking these encoder blocks allows the Transformer to learn multilevel representations from the input sequence. In electricity load forecasting, the Transformer effectively models dependencies among historical data points, especially those reflecting seasonal and long-term trends. Moreover, the architecture can be seamlessly integrated with models such as the multi-layer perceptron (MLP) to form a hybrid transformer-MLP model, where the Transformer extracts temporal features and the MLP performs nonlinear mapping at the output stage. This hybrid approach enhances prediction accuracy in complex time series forecasting tasks.

III. Suggested Methodology

A. *Proposed Hybrid Model Architecture*

This study proposes a hybrid model combining the transformer Encoder architecture with multi-layer Perceptron (MLP) layers. The objective is to leverage the transformer's strength in learning long-range dependencies and the powerful nonlinear mapping capability of MLPs to enhance forecasting accuracy in complex time series problems such as electricity load prediction. The main components of the proposed model include:

Input embedding layer: The input data has the shape (B, S, 1), where B is the batch size, S is the sequence length, and 1 represents the number of features. This data is projected into a higher-dimensional space with dmodel=16d using a linear transformation:

$$Embedding(x) = xW_e + b_e \qquad (20)$$

Learnable positional encoding: Since the Transformer does not inherently capture the order of time steps, a learnable positional encoding matrix of shape (1, S, d_{model}) is added to the embedded representation:

$$x_{input} = Embedding(x) + PositionalEncoding \qquad (21)$$

Transformer encoder: The embedded and positionally encoded input is passed through a Transformer Encoder block. A transformer encoder layer is used with the following hyperparameters: d_{model}=16; n_{head}=1; $dim_{feedforward}$ =128; dropout = 0.1.

Then, a transformer encoder is constructed by stacking the encoder layer num_layers=1 time. This encoder captures temporal patterns and dependencies across the sequence:

$$x_{tramsformer} = TransformerEncoder(x_{input}) \qquad (22)$$

Extracting the Last Time Step: Since the task is to predict a single output value, the model only uses the representation from the last time step:

$$x_{last} =\ x_{transformer}[:, -1, :] \qquad (23)$$

Multi-layer perceptron (MLP): The output from the transformer is passed through a stack of fully connected layers with ReLU activations, structured as follows: 16→32→64→12816

Each layer is defined as:

$$x_{mlp} = RELU(xW_1 + b_1)RELU(xW_2 + b_2)\ RELU(xW_3 + b_3) \qquad (24)$$

Output layer: Finally, a linear layer maps the 128-dimensional vector to a single output, suitable for regression tasks:

$$\hat{y} = xW_0 + b_0 \qquad (25)$$

Model training: The model is trained using the mean squared error (MSE) loss function and the Adam optimizer with a learning rate of lr=0.001 over 500 epochs and a batch size 32. The training data is organized into a DataLoader to ensure efficient mini-batch gradient updates during training.

B. *Algorithm Flowchart*

Fig. 1 illustrates the algorithmic workflow adopted for training and evaluating the Transformer-MLP model in the context of short-term electricity load forecasting. The process begins with a time series of electricity demand data Y1, Y2, ..., Yn, which is first preprocessed through normalization and reshaping into a supervised learning format using a sliding window technique. The preprocessed dataset is divided into two subsets: a training set (X_train, Y_train) and a testing set (X_test, Y_test). The Transformer-MLP model is trained on the training data, generating predictions Ypred based on the input Xtest. These predictions are subsequently compared with the actual target values Ytest, and standard error metrics—including MAE, MSE, and RMSE —are computed to assess the model's forecasting performance quantitatively. This structured workflow ensures consistency, reproducibility, and transparency in the experimental evaluation process.

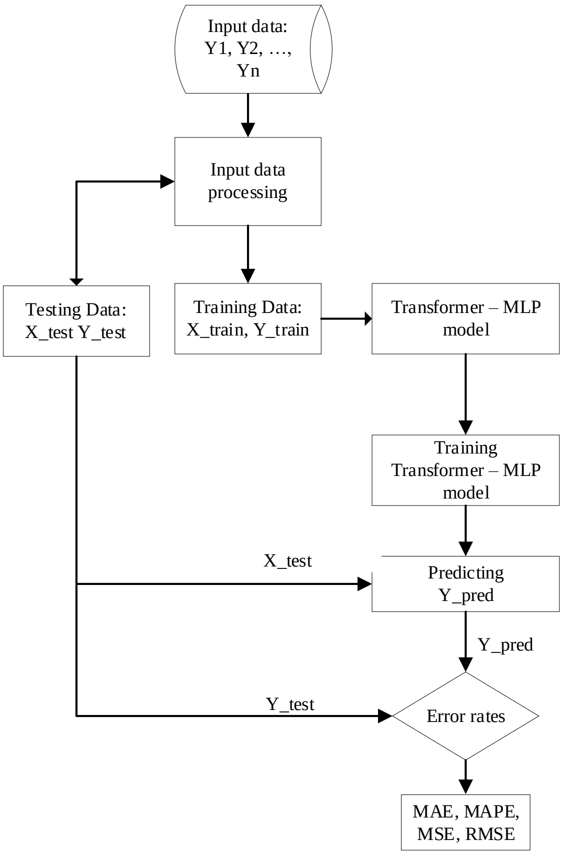

Fig. 1. Algorithm flowchart

IV. PERFORMANCE EVALUATION AND DISCUSSION

A. Experimental Data

Table I, Table II, and Table III present the electricity consumption data for three Australian states: New South Wales (NSW), Queensland (QL), and Victoria (VI). Each dataset contains time-stamped electricity demand values (TOTALDEMAND) recorded at 30-minute intervals. The data covers a continuous period from May 1, 2009, to May 31, 2014, yielding 1,857 consecutive days for each state. This constitutes a long-term time series dataset, a robust foundation for applying machine learning and deep learning models in the context of Short-Term Load Forecasting (STLF). The uniformity in the duration and structure of the data across the three states enables meaningful comparative analyses, facilitates the exploration of geographical consumption trends, and supports the development of national and region-specific forecasting models.

During the data preprocessing stage, the most recent 35 days of data were extracted from each dataset to emphasize recent patterns relevant for model training. The 30-minute sampling rate (equivalent to 48 data points per day) results in 1,680 time-ordered samples per state. The datasets were then restructured into input–output pairs using a sliding window technique with a fixed window length 48. Each input sequence comprises 48 consecutive values, and the corresponding output is the immediate next value in the sequence. The resulting data was split chronologically into training (80%) and testing (20%) sets to preserve the temporal dependencies inherent in time series data. This preprocessing strategy ensures consistency, temporal integrity. It facilitates effective generalization of deep learning models such as long short-term memory (LSTM) networks and transformer architectures when applied to the STLF task.

TABLE I. HISTORICAL LOAD DATA OF NSW

DATE	H0	H1	………	H46	H47
01/05/09	8724.1	8565.6	………	8798.7	8765.0
02/05/09	8608.93	8422.36	………	8162.89	8128.82
……..	……..	……..	………	……..	…..
30/05/14	7747.18	7556.37	………	7840.4	7829.76
31/05/14	7848.03	7654.02	………	7328.17	7355.55

TABLE II. HISTORICAL LOAD DATA OF QUEENSLAND

DATE	H0	H1	………	H46	H47
01/05/09	5318.73	5112.69	………	5572.11	5466.93
02/05/09	5267.31	5074.75	………	5360.74	5302.82
……..	……..	…………	………	…………	…………
30/05/14	5346.41	5171.66	………	5581.69	5458.63
31/05/14	5247.4	5126.62	………	5378.7	5328.07

TABLE III. HISTORICAL LOAD DATA OF VICTORIA

DATE	H0	H1	………	H46	H47
01/05/09	5700.4	5486.66	………	5797.02	5992.16
02/05/09	5794.11	5612.12	………	5426.44	5640.43
………	………	………	………	………	………
30/05/2014	5760.19	5470.23	………	5493.71	5813.51
31/05/2014	5751.21	5506.47	………	5050.1	5379.3

B. Model Parameters

The model is designed based on a hybrid architecture combining a Transformer encoder with a multi-layer perceptron (MLP) to handle univariate time series data with a sequence length of 48. The input, shaped as (batch_size, 48, 1), is first passed through a linear embedding layer that maps the single input feature to a 16-dimensional space (d_model = 16). To incorporate temporal position information, the model employs a learnable positional encoding of shape (1, 48, 16). The Transformer encoder consists of one encoder layer (num_layers = 1) with a single attention head (nhead = 1) and an internal feedforward layer size of 128 (dim_feedforward = 128). After the Transformer processes the input, only the final time step’s output is retained and passed through an MLP consisting of three hidden layers with dimensions: 16 → 32 → 64 → 128, each followed by a ReLU activation. Finally, a linear output layer maps the 128-dimensional vector to a single output value for prediction. The model is trained using the Mean Squared Error (MSE) loss function and optimized with the Adam optimizer, with a learning rate of lr = 0.001, over 500 epochs and a batch size of 32.

We intentionally adopted a lightweight Transformer encoder configuration with only one attention head and a single encoder layer to ensure a practical balance between model complexity and forecasting efficiency. This minimalistic architecture was selected based on preliminary tests, which showed that deeper Transformer variants (e.g., using 2–4 layers or multiple heads) slightly improved performance but significantly increased training time, risk of overfitting, and memory usage. Given that our target application includes deployment in innovative grid environments—where real-time response and computational resources may be constrained—this simplified configuration offers sufficient temporal feature extraction without compromising operational feasibility. The design thus prioritizes model interpretability, efficiency, and scalability, which are essential for load forecasting in real-world power system operations.

In addition to the lightweight transformer encoder, the multi-layer perceptron (MLP) component was deliberately designed with a relatively simple architecture, consisting of three hidden layers with increasing dimensions (from 16 to 128). This configuration was chosen to provide sufficient nonlinear feature extraction at the output stage without introducing excessive depth that might result in overfitting or prolonged training. Empirical testing confirmed that deeper or more complex MLP structures did not yield noticeable improvements in forecasting accuracy, especially after the Transformer had already encoded long-range temporal patterns. Thus, the selected MLP structure effectively balances complexity and performance, making the model suitable for deployment in real-time energy management systems with limited computational resources.

C. *Results*

Fig. 2, Fig. 3, and Fig. 4 present the short-term electricity load forecasting results obtained using the Multi-Layer Perceptron (MLP) model for three Australian states: New South Wales (NSW), Queensland (QLD), and Victoria (VIC). The visual comparisons between the predicted values (*y_pred*) and the actual observed values (*y_test*) reveal that the MLP model effectively captures the temporal patterns and trends in the electricity load data. The predicted curves in NSW and QLD align with the actual load values, indicating high forecasting accuracy and minimal deviation. This strong performance can be attributed to the relatively smooth and consistent oscillatory behavior of the load signals in these regions, which the MLP model is well-suited to learn and generalize from.

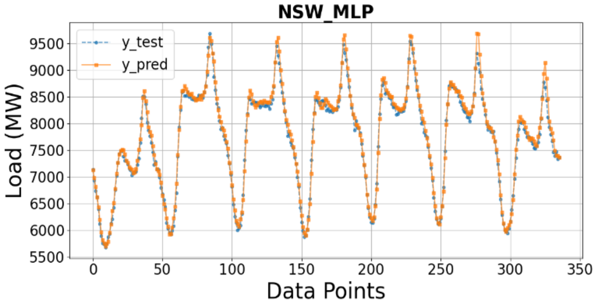

Fig. 2. Electricity load forecasting in NSW using MLP

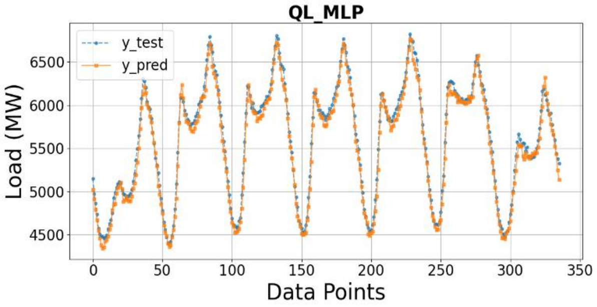

Fig. 3. Electricity load forecasting in QL using MLP

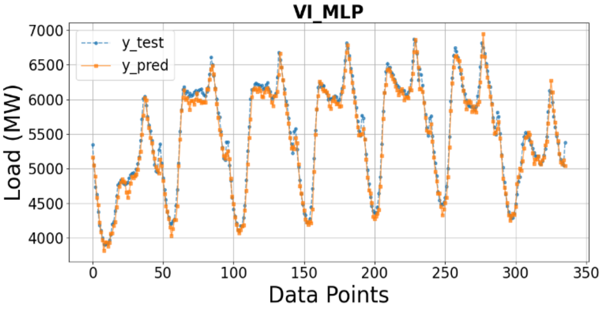

Fig. 4. Electricity load forecasting in VI using MLP

However, while the overall trend is still well-followed in the VI dataset, the model exhibits noticeable discrepancies at specific time steps. These deviations are primarily due to the more volatile nature of the load profile in Victoria, which includes rapid fluctuations and sharp peaks that are inherently more difficult to predict. Despite these challenges, the MLP model still demonstrates reasonable robustness and adaptability. Moreover, its straightforward architecture and ease of implementation make it a practical choice for load forecasting tasks, particularly in scenarios where the complexity of the input signal is moderate or low. Overall, the results confirm the MLP model’s potential as a baseline method for electricity demand prediction across various regions.

Fig. 5, Fig. 6, and Fig. 7 present the electricity load forecasting results using the LSTM model for NSW, QLD, and VI states. Unlike MLP, the LSTM model is superior in learning temporal dependencies. However, in the displayed plots, the model does not achieve a high level of alignment with the actual data. In NSW (Fig. 4), while the LSTM captures the overall trend, it fails to identify several critical load peaks. For QL and VI (Fig. 5 and Fig. 6), the forecasts show noticeable deviations from the actual values, especially at points with intense fluctuations.

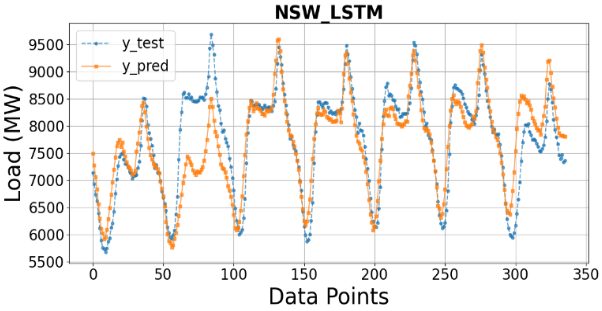

Fig. 5. Electricity load forecasting in NSW using LSTM

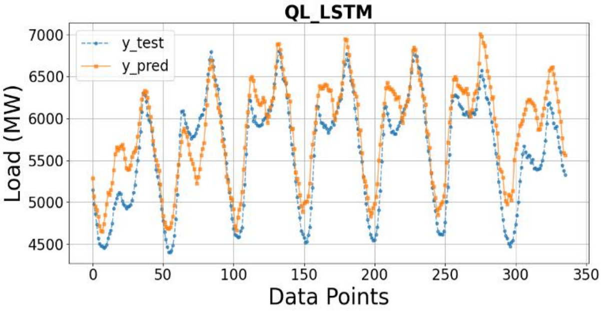

Fig. 6. Electricity load forecasting in QL using LSTM

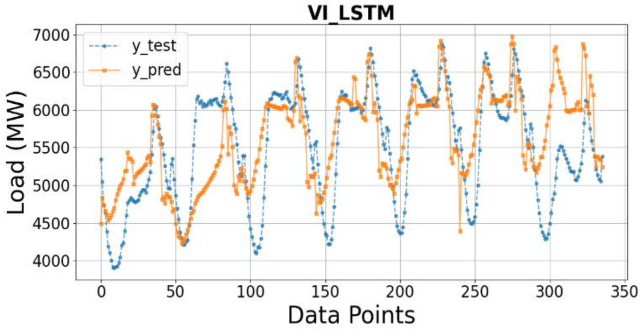

Fig. 7. Electricity load forecasting in VI using LSTM

Although the long short-term memory (LSTM) model is inherently well-suited for handling sequential and time-dependent data, the experimental results indicate that its current performance in electricity load forecasting is suboptimal. Therefore, additional hyperparameter tuning, architectural adjustments, and training refinements are required to improve its applicability and accuracy.

Fig. 8, Fig. 9, and Fig. 10 depict the electricity load forecasting results generated by the Convolutional Neural Network (CNN) model for the Australian states of New South Wales (NSW), Queensland (QLD), and Victoria (VIC). The visualizations demonstrate that the CNN model produces predictions highly consistent with the actual load values across all three regions. Specifically, the predicted curves (y_pred) closely follow the ground truth values (y_test), including during periods of sharp fluctuations, sudden peaks, or rapid transitions. This level of alignment indicates that the CNN model can learn the general trend in the data and respond to local variations and short-term anomalies. The model delivers superior predictive accuracy in NSW and QLD, with minimal deviation observed. In VIC, where the load signal exhibits more complex and irregular behavior, CNN maintains robust and stable tracking of the real values. These findings highlight the model's ability to effectively extract and utilize local temporal features. Due to its high performance, computational efficiency, and ease of implementation, CNN is considered a practical and reliable approach for short-term electricity load forecasting across diverse regional datasets.

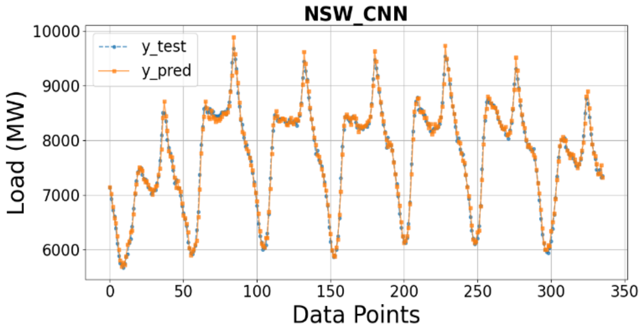

Fig. 8. Electricity load forecasting in NSW using CNN

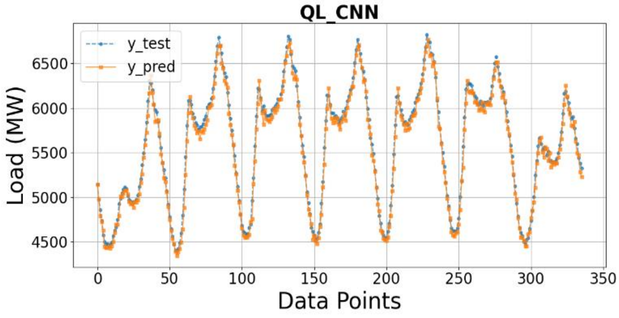

Fig. 9. Electricity load forecasting in QL using CNN

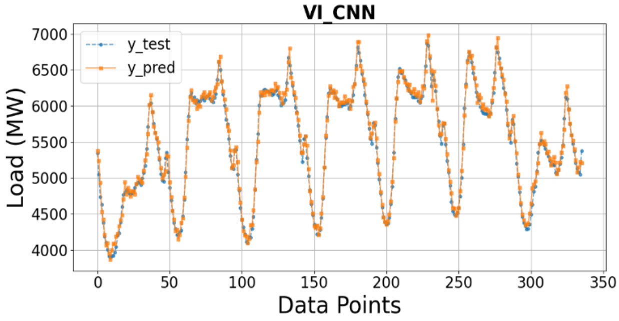

Fig. 10. Electricity load forecasting in VI using CNN

Fig. 11, Fig. 12, and Fig. 13 illustrate the electricity load forecasting outcomes for the states of New South Wales (NSW), Queensland (QLD), and Victoria (VIC) using the proposed Transformer-MLP hybrid model. By integrating the self-attention mechanism of the Transformer with the nonlinear learning capability of the MLP, the model achieves exceptional predictive performance. The forecasted curves (y_pred) align almost perfectly with the actual load values (y_test) across all three states. In particular, the model demonstrates excellent accuracy in NSW (Fig. 11) and QLD (Fig. 12), effectively capturing sharp peaks and minor fluctuations in the data. Even in the case of VI (Fig. 13), where the load profile is highly irregular and volatile, the Transformer-MLP maintains a high level of accuracy. These results confirm the model's strong ability to learn complex temporal dependencies and adaptability to diverse and challenging load patterns, making it a powerful tool for short-term load forecasting.

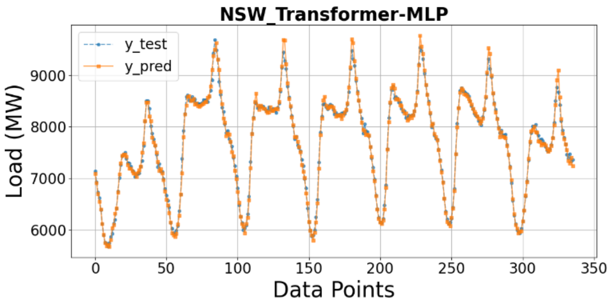

Fig. 11. Electricity load forecasting in NSW using transformer – MLP

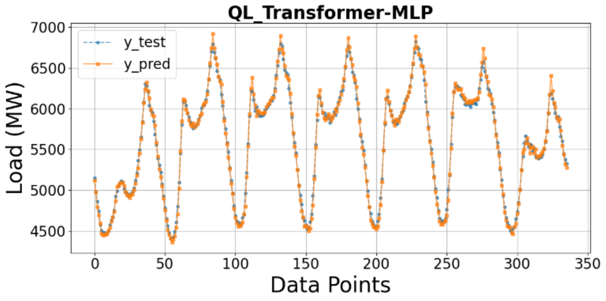

Fig. 12. Electricity load forecasting in QL using transformer – MLP

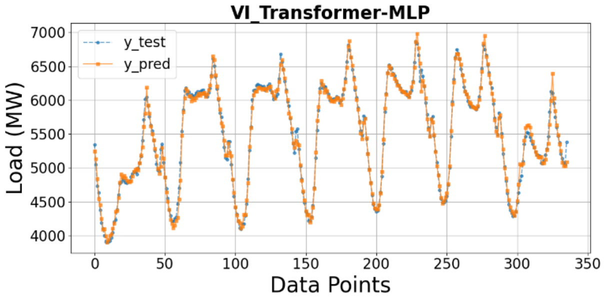

Fig. 13. Electricity load forecasting in VI using transformer – MLP

Fig. 14, Fig. 15, and Fig. 16 present a comparative analysis of the mean absolute percentage error (MAPE) across different deep learning models—LSTM, MLP, CNN, and Transformer-MLP—for the NSW, QL, and VI regions. Across all areas, the Transformer-MLP model consistently achieves the lowest MAPE values, highlighting its superior forecasting accuracy and stability compared to the other models. Specifically, the Transformer-MLP yields MAPE values ranging from only 0.69% to 0.95%, significantly lower than traditional models such as LSTM. Notably, LSTM records the highest error rates across all regions: 4.82% (NSW), 5.38% (QL), and 9.44% (VI). CNN and MLP models exhibit intermediate performance, with MAPE values typically ranging from 0.89% to 1.59%. Although these figures are lower than those of LSTM, both models fall short of matching the accuracy of the Transformer-MLP, particularly in the VI region, where MLP records a MAPE of 1.59%, nearly double that of the Transformer-MLP (0.95%).

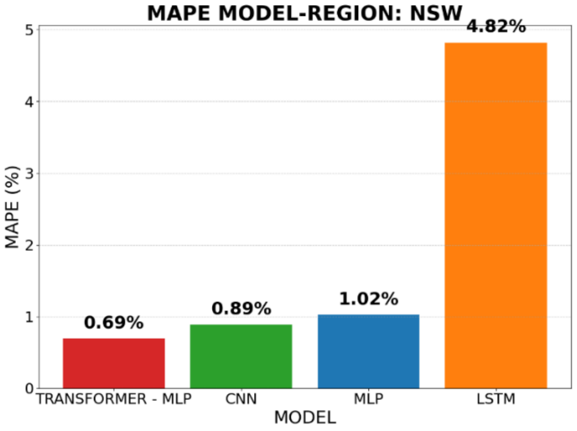

Fig. 14. MAPE comparison across models – region: NSW

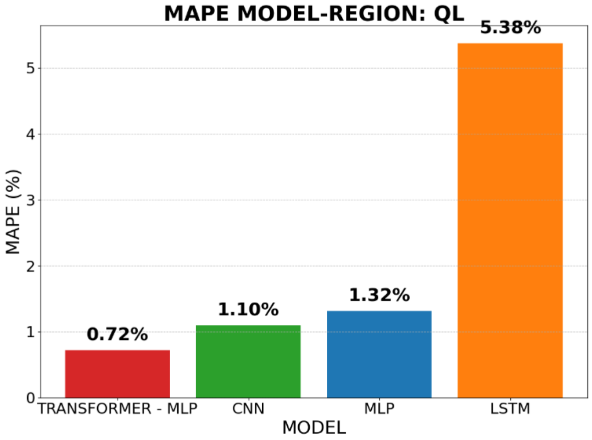

Fig. 15. MAPE comparison across models – region: QL

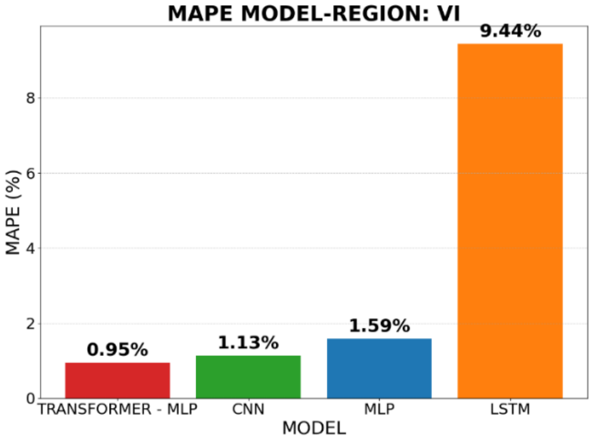

Fig. 16. MAPE Comparison across models – region: VI

Fig. 17 compares the execution time of four deep learning models—Transformer, MLP, CNN, and LSTM—across three geographical regions: NSW, QLD, and VI. Each group of bars represents a region and shows the execution time for the corresponding four models. The results indicate that LSTM consistently exhibits the highest execution time across all areas, ranging from 1,266 to 1,389 seconds. This reflects the inherently sequential nature and architectural complexity of LSTM, which demands longer training time and higher computational resources than other models. In contrast, MLP is consistently the fastest model, requiring only about 58–101 seconds. This underscores the simplicity of the traditional feedforward neural network architecture, making it suitable for applications where quick response time and computational efficiency are critical. CNN and Transformer-MLP demonstrate intermediate execution performance. CNN's execution time ranges from 148 to 183 seconds, while Transformer-MLP exhibits more variable runtime, between 176 and 681 seconds, with a noticeable increase in NSW. This variation may be attributed to differences in input sequence length or the complexity of the datasets across regions.

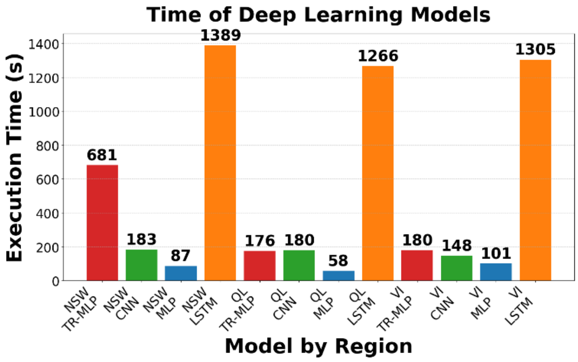

Fig. 17. Execution time of deep learning models across regions

D. Discussion

The findings of this study highlight the effectiveness of the proposed Transformer-MLP model in short-term electric load forecasting by demonstrating consistently lower prediction errors compared to traditional deep learning models. This superior performance can be attributed to the model's ability to integrate the transformer's self-attention mechanism—capable of capturing long-range temporal dependencies—with the MLP's capacity for nonlinear mapping and output stabilization. Unlike LSTM, which suffers from high training time and potential overfitting in highly volatile regions, the Transformer-MLP achieves high accuracy and computational efficiency with a shallow architecture. While CNN excels in extracting local patterns, it lacks global sequence awareness, limiting its adaptability in datasets with complex dynamics. MLP, though fast and lightweight, struggles to handle sequential correlations effectively. The Transformer–MLP strikes a balance, offering a robust, generalizable, and low-complexity solution suitable for deployment in innovative grid systems. These results suggest that attention-based hybrid models may provide a valuable direction for future load forecasting systems. However, limitations remain, such as excluding exogenous variables (e.g., weather or calendar effects) and focusing on one-step-ahead prediction. Future research could address these gaps by incorporating external features, exploring multi-step forecasting, and evaluating deeper Transformer variants further to improve accuracy and generalization under diverse load conditions.

V. Conclusion

This study proposed a lightweight hybrid deep learning model that integrates a Transformer encoder with a multi-layer perceptron (MLP) to improve the accuracy of short-term electricity load forecasting. The model was evaluated on real-world datasets from three Australian states—NSW, QLD, and VIC—and demonstrated superior performance to conventional models, including MLP, LSTM, and CNN. Among all models, Transformer-MLP consistently achieved the lowest MAPE values, ranging from 0.69% to 0.95%, confirming its robustness across stable and volatile load scenarios. The key strength of the proposed model lies in its ability to simultaneously capture long-term temporal dependencies through the Transformer and handle nonlinear patterns using the MLP. Its shallow architecture also ensures computational efficiency, making it suitable for real-time applications in innovative grid systems.

However, this study presents certain limitations. The model was designed for one-step-ahead forecasting and relied solely on historical load data without considering exogenous variables such as weather conditions, holidays, or socioeconomic factors. Additionally, only a single-layer Transformer with one attention head was used, which may restrict the model's capacity to learn deeper hierarchical patterns in complex scenarios. Future research will expand the model to multi-step forecasting tasks, integrate external influencing factors (e.g., temperature, calendar events), and experiment with deeper transformer architectures or multi-head attention mechanisms to enhance accuracy and generalizability in real-world deployments.

References

[1] M. Abdurohman and A. G. Putrada, "Forecasting Model for Lighting Electricity Load with a Limited Dataset using XGBoost," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, vol. 8, no. 2, pp. 571–580, 2023, doi: 10.22219/kinetik.v8i2.1687.

[2] J. Mo, R. Wang, M. Cao, K. Yang, X. Yang, and T. Zhang, "A hybrid temporal convolutional network and Prophet model for power load forecasting," *Complex and Intelligent Systems*, vol. 9, no. 4, pp. 4249–4261, 2023, doi: 10.1007/s40747-022-00952-x.

[3] X. Guo, Q. Zhao, D. Zheng, Y. Ning, and Y. Gao, "A short-term load forecasting model of multi-scale CNN-LSTM hybrid neural network considering the real-time electricity price," *Energy Reports*, vol. 6, pp. 1046–1053, 2020, doi: 10.1016/j.egyr.2020.11.078.

[4] A. Mansouri, A. H. Abolmasoumi, and A. A. Ghadimi, "Weather sensitive short term load forecasting using dynamic mode decomposition with control," *Electric Power Systems Research*, vol. 221, p. 109387, 2023, doi: 10.1016/j.epsr.2023.109387.

[5] R. Lu *et al.*, "A novel sequence-to-sequence-based deep learning model for multistep load forecasting," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 1, pp. 638–652, 2025, doi: 10.1109/TNNLS.2023.3329466.

[6] P. D. Lê *et al.*, "Applying statistical analysis for assessing the reliability of input data to improve the quality of short-term load forecasting for a Ho Chi Minh City distribution network," *Science & Technology Development Journal - Engineering and Technology*, vol. 2, no. 4, pp. 223–239, 2020, doi: 10.32508/stdjet.v2i4.614.

[7] S. Aoufi, A. Derhab, M. Guerroumi, H. Guemmouma, and H. Lazali, "LITE-FORT: Lightweight three-stage energy theft detection based on time series forecasting of consumption patterns," *Electric Power Systems Research*, vol. 225, p. 109840, 2023, doi: 10.1016/j.epsr.2023.109840.

[8] Q. Liu, J. Cao, J. Zhang, Y. Zhong, T. Ba, and Y. Zhang, "Short-Term Power Load Forecasting in FGSM-Bi-LSTM Networks Based on Empirical Wavelet Transform," *IEEE Access*, vol. 11, pp. 105057–105068, 2023, doi: 10.1109/ACCESS.2023.3316516.

[9] E. A. Siqueira-Filho, M. F. A. Lira, A. Converti, H. V. Siqueira, and C. J. A. Bastos-Filho, "Predicting Thermoelectric Power Plants Diesel/Heavy Fuel Oil Engine Fuel Consumption Using Univariate Forecasting and XGBoost Machine Learning Models," *Energies*, vol. 16, no. 7, 2023, doi: 10.3390/en16072942.

[10] Y. Miao, Z. Chen, J. Zhu, S. Li, H. Dong, and X. Wen, "Short-term Load Forecasting Based on Echo State Network and LightGBM," in *2023 IEEE International Conference on Predictive Control of Electrical Drives and Power Electronics, PRECEDE 2023*, no. 52177087, pp. 1–6, 2023, doi: 10.1109/PRECEDE57319.2023.10174609.

[11] A. Ghasemieh, A. Lloyed, P. Bahrami, P. Vajar, and R. Kashef, "A novel machine learning model with Stacking Ensemble Learner for predicting emergency readmission of heart-disease patients," *Decision Analytics Journal*, vol. 7, p. 100242, 2023, doi: 10.1016/j.dajour.2023.100242.

[12] J. Luo, Y. Zheng, T. Hong, A. Luo, and X. Yang, "Fuzzy support vector regressions for short-term load forecasting," *Fuzzy Optimization and Decision Making*, vol. 23, no. 3, pp. 363–385, 2024, doi: 10.1007/s10700-024-09425-x.

[13] B. Chen and Y. Wang, "Short-Term Electric Load Forecasting of Integrated Energy System Considering Nonlinear Synergy between Different Loads," *IEEE Access*, vol. 9, pp. 43562–43573, 2021, doi: 10.1109/ACCESS.2021.3066915.

[14] H. Hou *et al.*, "Load Forecasting Combining Phase Space Reconstruction and Stacking Ensemble Learning," *IEEE Transactions on Industry Applications*, vol. 59, no. 2, pp. 2296–2304, 2023, doi: 10.1109/TIA.2022.3225516.

[15] X. Yang and Z. Chen, "A Hybrid Short-Term Load Forecasting Model Based on CatBoost and LSTM," in *2021 IEEE 6th International Conference on Intelligent Computing and Signal Processing, ICSP 2021*, pp. 328–332, 2021, doi: 10.1109/ICSP51882.2021.9408768.

[16] S. Atef and A. B. Eltawil, "Assessment of stacked unidirectional and bidirectional long short-term memory networks for electricity load forecasting," *Electric Power Systems Research*, vol. 187, p. 106489, 2020, doi: 10.1016/j.epsr.2020.106489.

[17] V. K. Saini, A. S. Al-Sumaiti, and R. Kumar, "Data driven net load uncertainty quantification for cloud energy storage management in residential microgrid," *Electric Power Systems Research*, vol. 226, p. 109920, 2024, doi: 10.1016/j.epsr.2023.109920.

[18] Z. Wen, L. Xie, Q. Fan, and H. Feng, "Long term electric load forecasting based on TS-type recurrent fuzzy neural network model," *Electric Power Systems Research*, vol. 179, p. 106106, 2020, doi: 10.1016/j.epsr.2019.106106.

[19] H. Liu, Y. Tang, Y. Pu, F. Mei, and D. Sidorov, "Short-term Load Forecasting of Multi-Energy in Integrated Energy System Based on Multivariate Phase Space Reconstruction and Support Vector Regression Mode," *Electric Power Systems Research*, vol. 210, p. 108066, 2022, doi: 10.1016/j.epsr.2022.108066.

[20] G. F. Fan, Y. R. Liu, H. Z. Wei, M. Yu, and Y. H. Li, "The new hybrid approaches to forecasting short-term electricity load," *Electric Power Systems Research*, vol. 213, p. 108759, 2022, doi: 10.1016/j.epsr.2022.108759.

[21] S. Rai and M. De, "Ensemble-based Load Forecasting for Smart Metered System," in *2022 2nd International Conference on Emerging Frontiers in Electrical and Electronic Technologies, ICEFEET 2022*, pp. 1–4, 2022, doi: 10.1109/ICEFEET51821.2022.9847955.

[22] M. Li and Y. Wang, "Power load forecasting and interpretable models based on GS_XGBoost and SHAP," *Journal of Physics: Conference Series*, vol. 2195, no. 1, 2022, doi: 10.1088/1742-6596/2195/1/012028.

[23] G. Yan, J. Wang, and M. Thwin, "A new Frontier in electric load forecasting: The LSV/MOPA model optimized by modified orca predation algorithm," *Heliyon*, vol. 10, no. 2, p. e24183, 2024, doi: 10.1016/j.heliyon.2024.e24183.

[24] I. U. Khalil, A. Ul Haq, and N. Ul Islam, "A deep learning-based transformer model for photovoltaic fault forecasting and classification," *Electric Power Systems Research*, vol. 228, p. 110063, 2024, doi: 10.1016/j.epsr.2023.110063.

[25] M. S. Tahsin, M. Al Karim, M. U. Ahmed, Y. Rahman, F. Tafannum, and S. Abdullah, "Comparative Analysis of Weather Prediction Using Ensemble Learning Models and Neural Network," in *Proceedings - 2021 19th OITS International Conference on Information Technology, OCIT 2021*, pp. 325–330, 2021, doi: 10.1109/OCIT53463.2021.00071.

[26] M. Dostmohammadi, M. Z. Pedram, S. Hoseinzadeh, and D. A. Garcia, "A GA-stacking ensemble approach for forecasting energy consumption in a smart household: A comparative study of ensemble methods," *Journal of Environmental Management*, vol. 364, p. 121264, 2024, doi: 10.1016/j.jenvman.2024.121264.

[27] S. Singh, A. Yassine, and R. Benlamri, "Internet of Energy: Ensemble Learning through Multilevel Stacking for Load Forecasting," in *Proceedings - IEEE 18th International Conference on Dependable, Autonomic and Secure Computing, IEEE 18th International Conference on Pervasive Intelligence and Computing, IEEE 6th International Conference on Cloud and Big Data Computing and IEEE 5th Cyber Science and Technology Congress, DASC/PiCom/CBDCom/CyberSciTech 2020*, pp. 658–664, 2020, doi: 10.1109/DASC-PICom-CBDCom-CyberSciTech49142.2020.00113.

[28] J. H. Kim, B. S. Lee, and C. H. Kim, "A Study on the development of long-term hybrid electrical load forecasting model based on MLP and statistics using massive actual data considering field applications," *Electric Power Systems Research*, vol. 221, p. 109415, 2023, doi: 10.1016/j.epsr.2023.109415.

[29] A. Faustine, N. J. Nunes, and L. Pereira, "Efficiency through simplicity: MLP-based approach for net-load forecasting with uncertainty estimates in low-voltage distribution networks," *IEEE Transactions on Power Systems*, vol. 40, no. 1, pp. 46–56, 2025, doi: 10.1109/TPWRS.2024.3400123.

[30] A. P. Wibawa, A. B. P. Utama, H. Elmunsyah, U. Pujianto, F. A. Dwiyanto, and L. Hernandez, "Time-series analysis with smoothed Convolutional Neural Network," *Journal of Big Data*, vol. 9, no. 1, 2022, doi: 10.1186/s40537-022-00599-y.

[31] I. C. Figueiró, A. R. Abaide, N. K. Neto, L. N. F. Silva, and L. L. C. Santos, "Bottom-Up Short-Term Load Forecasting Considering Macro-Region and Weighting by Meteorological Region," *Energies*, vol. 16, no. 19, pp. 1191–1198, 2023, doi: 10.3390/en16196857.

[32] S. Ryu and Y. Yu, "Quantile-Mixer: A Novel Deep Learning Approach for Probabilistic Short-Term Load Forecasting," *IEEE Transactions on Smart Grid*, vol. 15, no. 2, pp. 2237–2250, 2024, doi: 10.1109/TSG.2023.3290180.

[33] S. S. Subbiah and J. Chinnappan, "Deep learning based short term load forecasting with hybrid feature selection," *Electric Power Systems Research*, vol. 210, p. 108065, 2022, doi: 10.1016/j.epsr.2022.108065.

[34] A. Ajitha, M. Goel, M. Assudani, S. Radhika, and S. Goel, "Design and development of Residential Sector Load Prediction model during COVID-19 Pandemic using LSTM based RNN," *Electric Power Systems Research*, vol. 212, p. 108635, 2022, doi: 10.1016/j.epsr.2022.108635.

[35] M. Abumohsen, A. Y. Owda, and M. Owda, "Electrical Load Forecasting Using LSTM, GRU, and RNN Algorithms," *Energies*, vol. 16, no. 5, pp. 1–31, 2023, doi: 10.3390/en16052283.

[36] M. Zhang, Z. Yu, and Z. Xu, "Short-term load forecasting using recurrent neural networks with input attention mechanism and hidden connection mechanism," *IEEE Access*, vol. 8, pp. 186514–186529, 2020, doi: 10.1109/ACCESS.2020.3029224.

[37] H. Shahinzadeh, H. Sadrarhami, M. M. Hayati, H. Majidi-Gharehnaz, M. Abapour, and G. B. Gharehpetian, "Review and Comparative Analysis of Deep Learning Techniques for Smart Grid Load Forecasting," in *2024 20th CSI International Symposium on Artificial Intelligence and Signal Processing, AISP 2024*, pp. 1–9, 2024, doi: 10.1109/AISP61396.2024.10475303.

[38] O. A. Lawal and J. Teh, "Assessment of dynamic line rating forecasting methods," *Electric Power Systems Research*, vol. 214, p. 108807, 2023, doi: 10.1016/j.epsr.2022.108807.

[39] A. K. Mishra, P. Mishra, and H. D. Mathur, "A deep learning assisted adaptive nonlinear deloading strategy for wind turbine generator integrated with an interconnected power system for enhanced load frequency control," *Electric Power Systems Research*, vol. 214, 2023, doi: 10.1016/j.epsr.2022.108960.

[40] S. Singh and M. M. Tripathi, "A Comparative Analysis of Extreme Gradient Boosting Technique with Long Short-Term Memory and Layered Recurrent Neural Network for Electricity Demand Forecas," in *2021 6th International Conference on Recent Trends on Electronics, Information, Communication and Technology, RTEICT 2021*, pp. 297–302, 2021, doi: 10.1109/RTEICT52294.2021.9573988.

[41] Z. Xu *et al.*, "PhaCIA-TCNs: Short-term load forecasting using temporal convolutional networks with parallel hybrid activated convolution and input attention," *IEEE Transactions on Network Science and Engineering*, vol. 11, no. 1, pp. 427–438, 2024, doi: 10.1109/TNSE.2023.3300744.

[42] G. Gürses-Tran, T. A. Körner, and A. Monti, "Introducing explainability in sequence-to-sequence learning for short-term load forecasting," *Electric Power Systems Research*, vol. 212, p. 108366, 2022, doi: 10.1016/j.epsr.2022.108366.

[43] S. Luo, Y. Rao, J. Chen, H. Wang, and Z. Wang, "Short-Term Load Forecasting Model of Distribution Transformer Based on CNN and LSTM," in *7th IEEE International Conference on High Voltage Engineering and Application, ICHVE 2020 - Proceedings*, pp. 1–4, 2020, doi: 10.1109/ICHVE49031.2020.9279813.

[44] A. K. A. Penaloza, A. Balbinot, and R. C. Leborgne, "Review of deep learning application for short-term household load forecasting," in *2020 IEEE PES Transmission and Distribution Conference and Exhibition - Latin America, T and D LA 2020*, pp. 1–6, 2020, doi: 10.1109/TDLA47668.2020.9326148.

[45] Y. Liu, X. Wang, S. Wang, and Z. Xu, "Short-term Power Load Forecasting Based on Temporal Convolutional Network," in *Proceedings of the 2022 International Conference on Information,*

Control, and Communication Technologies, ICCT 2022, pp. 1–4, 2022, doi: 10.1109/ICCT56057.2022.9976543.

[46] J. Wen, Y. Peng, W. Zhang, X. Huang, and Z. Wang, "Short-term power load forecasting based on TCN-LSTM model," in *Proceedings of 2024 IEEE 6th International Conference on Civil Aviation Safety and Information Technology, ICCASIT 2024*, vol. 13, pp. 734–738, 2024, doi: 10.1109/ICCASIT62299.2024.10827868.

[47] S. Özdemir, Y. Demir, and Ö. Yildirim, "The effect of input length on prediction accuracy in short-term multi-step electricity load forecasting: A CNN-LSTM Approach," *IEEE Access*, vol. 13, pp. 28419–28432, 2025, doi: 10.1109/ACCESS.2025.3540636.

[48] C. Li, R. Hu, C. Y. Hsu, and Y. Han, "Short-term Power Load Forecasting based on Feature Fusion of Parallel LSTM-CNN," in *2022 IEEE 4th International Conference on Power, Intelligent Computing and Systems, ICPICS 2022*, no. 4, pp. 448–452, 2022, doi: 10.1109/ICPICS55264.2022.9873566.

[49] C. Cai, Y. Tao, Q. Ren, and G. Hu, "Short-term load forecasting based on MB-LSTM neural network," in *Proceedings - 2020 Chinese Automation Congress, CAC 2020*, pp. 5402–5406, 2020, doi: 10.1109/CAC51589.2020.9326696.

[50] H. Kuang, Q. Guo, S. Li, and H. Zhong, "Short-term Power Load Forecasting Method in Rural Areas Based on CNN-LSTM," in *Proceedings of 2021 IEEE 4th International Electrical and Energy Conference, CIEEC 2021*, pp. 7–11, 2021, doi: 10.1109/CIEEC50170.2021.9510777.

[51] T. H. Bao Huy, D. N. Vo, K. P. Nguyen, V. Q. Huynh, M. Q. Huynh, and K. H. Truong, "Short-Term Load Forecasting in Power System Using CNN-LSTM Neural Network," in *Conference Proceedings - 2023 IEEE Asia Meeting on Environment and Electrical Engineering, EEE-AM 2023*, pp. 1–6, 2023, doi: 10.1109/EEE-AM58328.2023.10395221.

[52] S. Chen, R. Lin, and W. Zeng, "Short-term load forecasting method based on ARIMA and LSTM," in *International Conference on Communication Technology Proceedings, ICCT*, vol. Nov 2022, pp. 1913–1917, 2022, doi: 10.1109/ICCT56141.2022.10073051.

[53] S. H. Rafi, N. Al-Masood, S. R. Deeba, and E. Hossain, "A short-term load forecasting method using integrated CNN and LSTM network," *IEEE Access*, vol. 9, pp. 32436–32448, 2021, doi: 10.1109/ACCESS.2021.3060654.

[54] B. Farsi, M. Amayri, N. Bouguila, and U. Eicker, "On short-term load forecasting using machine learning techniques and a novel parallel deep LSTM-CNN approach," *IEEE Access*, vol. 9, pp. 31191–31212, 2021, doi: 10.1109/ACCESS.2021.3060290.

[55] F. Xu, G. Weng, Q. Ye, and Q. Xia, "Research on Load Forecasting Based on CNN-LSTM Hybrid Deep Learning Model," in *2022 IEEE 5th International Conference on Electronics Technology, ICET 2022*, pp. 1332–1336, 2022, doi: 10.1109/ICET55676.2022.9824615.

[56] C. Fan, G. Li, L. Xiao, L. Yi, and S. Nie, "Short-Term Power Load Forecasting in City Based on ISSA-BiTCN-LSTM," *Cognitive Computation*, vol. 17, no. 1, 2025, doi: 10.1007/s12559-024-10401-1.

[57] W. G. Buratto, R. N. Muniz, A. Nied, and G. V. Gonzalez, "Seq2Seq-LSTM With Attention for Electricity Load Forecasting in Brazil," *IEEE Access*, vol. 12, pp. 30020–30029, 2024, doi: 10.1109/ACCESS.2024.3365812.

[58] M. Xue, L. Wu, Q. P. Zhang, J. X. Lu, X. Mao, and Y. Pan, "Research on Load Forecasting of Charging Station Based on XGBoost and LSTM Model," *Journal of Physics: Conference Series*, vol. 1757, no. 1, 2021, doi: 10.1088/1742-6596/1757/1/012145.

[59] M. Alhussein, K. Aurangzeb, and S. I. Haider, "Hybrid CNN-LSTM model for short-term individual household load forecasting," *IEEE Access*, vol. 8, pp. 180544–180557, 2020, doi: 10.1109/ACCESS.2020.3028281.

[60] S. Yin, Z. Chen, W. Liu, and Z. Su, "Ultra Short-Term Charging Load Forecasting Based on Improved Data Decomposition and Hybrid Neural Network," *IEEE Access*, vol. 13, pp. 58778–58789, 2025, doi: 10.1109/ACCESS.2025.3555737.

[61] B. Li, Y. Mo, F. Gao, and X. Bai, "Short-term probabilistic load forecasting method based on uncertainty estimation and deep learning model considering meteorological factors," *Electric Power Systems Research*, vol. 225, p. 109804, 2023, doi: 10.1016/j.epsr.2023.109804.

[62] H. Shi, L. Wang, R. Scherer, M. Wozniak, P. Zhang, and W. Wei, "Short-Term Load Forecasting Based on Adabelief Optimized Temporal Convolutional Network and Gated Recurrent Unit Hybrid Neural Network," *IEEE Access*, vol. 9, pp. 66965–66981, 2021, doi: 10.1109/ACCESS.2021.3076313.

[63] H. Hu and B. Zheng, "Short-term electricity load forecasting based on CEEMDAN-FE-BiGRU-Attention model," *International Journal of Low-Carbon Technologies*, vol. 19, pp. 988–995, 2024, doi: 10.1093/ijlct/ctae040.

[64] T. A. Nguyen and T. N. Tran, "Improving short-term electrical load forecasting with dilated convolutional neural networks: a comparative analysis," *Journal of Robotics and Control (JRC)*, vol. 6, no. 2, pp. 560–569, 2025, doi: 10.18196/jrc.v6i2.24967.

[65] R. Liu, T. Chen, G. Sun, S. M. Muyeen, S. Lin, and Y. Mi, "Short-term probabilistic building load forecasting based on feature integrated artificial intelligent approach," *Electric Power Systems Research*, vol. 206, p. 107802, 2022, doi: 10.1016/j.epsr.2022.107802.

[66] A. Parizad and C. J. Hatziadoniu, "A Real-Time Multistage False Data Detection Method Based on Deep Learning and Semisupervised Scoring Algorithms," *IEEE Systems Journal*, vol. 17, no. 2, pp. 1753–1764, 2023, doi: 10.1109/JSYST.2023.3265021.

[67] J. Gan, L. Pan, Y. Jin, Q. Liu, and X. Liu, "A Load Forecasting Approach Based on Graph Convolution Neural Network," in *Proceedings of the 2022 IEEE International Conference on Dependable, Autonomic and Secure Computing, International Conference on Pervasive Intelligence and Computing, International Conference on Cloud and Big Data Computing, International Conference on Cyber Science and Technology Congress, DASC/PiCom/CBDCom/CyberSciTech 2022*, pp. 1–3, 2022, doi: 10.1109/DASC/PiCom/CBDCom/Cy55231.2022.9927829.

[68] Y. Lu, G. Wang, X. Huang, S. Huang, and M. Wu, "Probabilistic load forecasting based on quantile regression parallel CNN and BiGRU networks," *Applied Intelligence*, vol. 54, no. 15–16, pp. 7439–7460, 2024, doi: 10.1007/s10489-024-05540-9.

[69] W. Xiong, L. Han, and X. Qu, "Bus Load Forecasting Based on Maximum Information Coefficient and CNN-LSTM Model," in *2023 IEEE International Conference on Image Processing and Computer Applications, ICIPCA 2023*, pp. 659–663, 2023, doi: 10.1109/ICIPCA59209.2023.10257944.

[70] J. Zhang, Z. Zhu, and Y. Yang, "Electricity Load Forecasting Based on CNN-LSTM," in *2023 IEEE International Conference on Electrical, Automation and Computer Engineering, ICEACE 2023*, pp. 1385–1390, 2023, doi: 10.1109/ICEACE60673.2023.10442217.

[71] S. Wu *et al.*, "Power Load Forecasting Method Based on Random Matrix Theory and CNN-LSTM Model," in *2022 IEEE 2nd International Conference on Digital Twins and Parallel Intelligence, DTPI 2022*, pp. 1–6, 2022, doi: 10.1109/DTPI55838.2022.9998910.

[72] O. Rubasinghe, X. Zhang, T. K. Chau, Y. H. Chow, T. Fernando, and H. H. C. Iu, "A Novel Sequence to Sequence Data Modelling Based CNN-LSTM Algorithm for Three Years Ahead Monthly Peak Load Forecasting," *IEEE Transactions on Power Systems*, vol. 39, no. 1, pp. 1932–1947, 2024, doi: 10.1109/TPWRS.2023.3271325.

[73] K. Aurangzeb, M. Alhussein, K. Javaid, and S. I. Haider, "A Pyramid-CNN based deep learning model for power load forecasting of similar-profile energy customers based on clustering," *IEEE Access*, vol. 9, pp. 14992–15003, 2021, doi: 10.1109/ACCESS.2021.3053069.

[74] M. Aouad, H. Hajj, K. Shaban, R. A. Jabr, and W. El-Hajj, "A CNN-Sequence-to-Sequence network with attention for residential short-term load forecasting," *Electric Power Systems Research*, vol. 211, p. 108152, 2022, doi: 10.1016/j.epsr.2022.108152.

[75] A. Irankhah, S. R. Saatlou, M. H. Yaghmaee, S. Ershadi-Nasab, and M. Alishahi, "A parallel CNN-BiGRU network for short-term load forecasting in demand-side management," in *2022 12th International Conference on Computer and Knowledge Engineering, ICCKE 2022*, pp. 511–516, 2022, doi: 10.1109/ICCKE57176.2022.9960036.

[76] S. Luo, Z. Ni, X. Zhu, P. Xia, and H. Wu, "A novel methanol futures price prediction method based on multicycle CNN-GRU and Attention Mechanism," *Arabian Journal for Science and Engineering*, vol. 48, no. 2, pp. 1487–1501, 2023, doi: 10.1007/s13369-022-06902-6.

[77] X. Li, H. Guo, L. Xu, and Z. Xing, "Bayesian-Based Hyperparameter Optimization of 1D-CNN for Structural Anomaly Detection," *Sensors*, vol. 23, no. 11, 2023, doi: 10.3390/s23115058.

[78] H. Hua, M. Liu, Y. Li, S. Deng, and Q. Wang, "An ensemble framework for short-term load forecasting based on parallel CNN and GRU with improved ResNet," *Electric Power Systems Research*, vol. 216, p. 109057, 2023, doi: 10.1016/j.epsr.2022.109057.

[79] Z. Tian, W. Liu, W. Jiang, and C. Wu, "CNNs-Transformer based day-ahead probabilistic load forecasting for weekends with limited data availability," *Energy*, vol. 293, p. 130666, 2024, doi: 10.1016/j.energy.2024.130666.

[80] C. Wang, Y. Wang, Z. Ding, and K. Zhang, "Probabilistic Multi-Energy Load Forecasting for Integrated Energy System Based on Bayesian Transformer Network," *IEEE Transactions on Smart Grid*, vol. 15, no. 2, pp. 1495–1508, 2024, doi: 10.1109/TSG.2023.3296647.

[81] Y. Shang *et al.*, "Loss of life estimation of distribution transformers considering corrupted AMI data recovery and field verification," *IEEE Transactions on Power Delivery*, vol. 36, no. 1, pp. 180–190, 2021, doi: 10.1109/TPWRD.2020.2978809.

[82] K. Qu, G. Si, Z. Shan, Q. Wang, X. Liu, and C. Yang, "Forwardformer: Efficient Transformer With Multi-Scale Forward Self-Attention for Day-Ahead Load Forecasting," *IEEE Transactions on Power Systems*, vol. 39, no. 1, pp. 1421–1433, 2024, doi: 10.1109/TPWRS.2023.3266369.

[83] P. Ran, K. Dong, X. Liu, and J. Wang, "Short-term load forecasting based on CEEMDAN and Transformer," *Electric Power Systems Research*, vol. 214, p. 108885, 2023, doi: 10.1016/j.epsr.2022.108885.

[84] I. Diahovchenko, A. Chuprun, and Z. Čonka, "Assessment and mitigation of the influence of rising charging demand of electric vehicles on the aging of distribution transformers," *Electric Power Systems Research*, vol. 221, 2023, doi: 10.1016/j.epsr.2023.109455.

[85] C. Xu and G. Chen, "Interpretable transformer-based model for probabilistic short-term forecasting of residential net load," *International Journal of Electrical Power and Energy Systems*, vol. 155, p. 109515, 2024, doi: 10.1016/j.ijepes.2023.109515.

[86] A. Ahmad, X. Xiao, H. Mo, and D. Dong, "TFTformer: A novel transformer based model for short-term load forecasting," *International Journal of Electrical Power and Energy Systems*, vol. 166, p. 110549, 2025, doi: 10.1016/j.ijepes.2025.110549.

[87] W. Zeng *et al.*, "Hybrid CEEMDAN-DBN-ELM for online DGA serials and transformer status forecasting," *Electric Power Systems Research*, vol. 217, 2023, doi: 10.1016/j.epsr.2023.109176.

[88] H. Tong and J. Liu, "MFformer: An improved transformer-based multi-frequency feature aggregation model for electricity load forecasting," *Electric Power Systems Research*, vol. 243, p. 111492, 2025, doi: 10.1016/j.epsr.2025.111492.

[89] Z. Tang, T. Ji, J. Kang, Y. Huang, and W. Tang, "Learning global and local features of power load series through transformer and 2D-CNN: An image-based multi-step forecasting approach incorporating phase space reconstruction," *Applied Energy*, vol. 378, p. 124786, 2025, doi: 10.1016/j.apenergy.2024.124786.

[90] T. Quanwei, X. Guijun, and X. Wenju, "Cakformer: Transformer model for long-term heat load forecasting based on Cauto-correlation and KAN," *Energy*, vol. 324, p. 135460, 2025, doi: 10.1016/j.energy.2025.135460.

[91] T. Bashir, H. Wang, M. Tahir, and Y. Zhang, "Wind and solar power forecasting based on hybrid CNN-ABiLSTM, CNN-transformer-MLP models," *Renewable Energy*, vol. 239, p. 122055, 2025, doi: 10.1016/j.renene.2024.122055.

[92] S. Mo *et al.*, "From global to local: A lightweight CNN approach for long-term time series forecasting," *Computers and Electrical Engineering*, vol. 123, p. 110192, 2025, doi: 10.1016/j.compeleceng.2025.110192.

[93] D. Wang, D. Peng, D. Huang, H. Zhao, and B. Qu, "MMEMformer: A multi-scale memory-enhanced transformer framework for short-term load forecasting in integrated energy systems," *Energy*, vol. 322, p. 135762, 2025, doi: 10.1016/j.energy.2025.135762.

[94] X. Cao *et al.*, "From Dense to Sparse: Event Response for Enhanced Residential Load Forecasting," *IEEE Transactions on Instrumentation and Measurement*, vol. 74, pp. 1–12, 2025, doi: 10.1109/TIM.2025.3544349.

[95] J. Liu *et al.*, "Temporal patterns decomposition and Legendre projection for long-term time series forecasting," *Journal of Supercomputing*, vol. 80, no. 16, pp. 23407–23441, 2024, doi: 10.1007/s11227-024-06313-4.