ISSN 2616-5767 Vol.8, Issue 4: 79-86, DOI: 10.25236/AJETS.2025.080411

Multi-Variable Time Series Forecasting for Wind Power Using Improved Transformer Architecture

Liyu Wu[a,*,#], Yanxin Liu[#], Xinyu Gao[#]

School of Communication and Artificial Intelligence, School of Integrate Circuits, Nanjing Institute of Technology, Nanjing, China
[a]15122006186@139.com
[#]These Authors Contributed Equally to This Work

Abstract: *This paper proposes an improved Transformer-based model MVTformer for multivariate time series forecasting tasks. The model introduces a sparse self-attention mechanism and a time series feature extraction module to improve the modeling capabilities of cross-variable dependencies and long-term dynamic features. MVTformer implemented on the PyTorch platform has good computational efficiency and scalability, and is suitable for complex industrial time series data scenarios. The experiment was carried out on a real wind power dataset. Compared with mainstream models such as LSTM, GRU, TCN and standard Transformer, MVTformer performed best in multiple evaluation indicators, fully verifying its accuracy and robustness in sequence prediction.*

Keywords: *Machine Learning; Deep Learning; Time Series Modeling; Transformer Architecture; Self-Attention Mechanism; Wind Power Forecasting; Sequence Learning*

1. Introduction

With the widespread application of renewable energy, especially the rapid development of wind energy, modern power systems are facing unprecedented challenges in stability and reliability [1]. Since wind energy is intermittent and volatile, accurate short-term wind power forecasting is of great significance for achieving efficient scheduling, economic operation and safe operation of power grids [2]. Traditional statistical methods (such as ARIMA, autoregressive smoothing models, etc.) are often difficult to effectively model the complex time series dependencies of high-dimensional, multivariate, and strongly nonlinear wind power data [3].

In recent years, deep learning methods (such as recurrent neural networks (RNNs), long short-term memory networks (LSTMs), and temporal convolutional networks (TCNs) have achieved remarkable results in time series forecasting. However, these models have certain limitations, such as insufficient ability to model long-term dependencies and low model training efficiency. The Transformer architecture was originally used for natural language processing tasks. Its self-attention mechanism and high parallelism make it show great potential in capturing long-term dependencies and complex time series patterns, and it has gradually been introduced into time series forecasting tasks [4].

This paper proposes a multivariate time series prediction model based on an improved Transformer, named MVTformer, which is specifically used for wind power prediction tasks. The model introduces a feature selection mechanism and an enhanced attention module in its structure, which can effectively extract the nonlinear relationship between input variables such as wind speed, wind direction, temperature, and blade angle [5]. Through empirical analysis on a real wind power dataset, this paper compares the proposed model with a variety of mainstream prediction methods. The results show that MVTformer is superior to existing methods in terms of prediction accuracy and generalization ability, and has good engineering practicality and deployment prospects [6].

2. Related Work

With the continuous increase in the installed capacity of wind farms, wind power prediction has gradually become a research hotspot in smart grids and new energy access systems. The current mainstream prediction methods can be roughly divided into three categories: physical models, statistical models, and data-driven models. The physical model relies on meteorological parameters and wind

ISSN 2616-5767 Vol.8, Issue 4: 79-86, DOI: 10.25236/AJETS.2025.080411

turbine structure information. Although it has a certain theoretical basis, it is limited by the difficulty in obtaining external data and the high computational complexity. Statistical models such as ARIMA and support vector regression (SVR) are suitable for short-term predictions, and the modeling process is relatively simple. However, when faced with high-dimensional, nonlinear, and strongly coupled wind power time series data, the prediction performance is often difficult to meet actual needs [7].

With the development of deep learning, data-driven methods have gradually become mainstream. Among them, recurrent neural networks (RNNs) and their variants, such as long short-term memory networks (LSTMs) and gated recurrent units (GRUs), are widely used in time series modeling. Structures such as LSTM have improved the model's memory capacity to a certain extent and can handle dependency problems with longer time spans. However, due to its inherent "sequential calculation" characteristics, it has problems with low computational efficiency and difficulty in parallelization. In addition, when faced with multi-variable inputs, the model has limited ability to model the interaction relationship between variables [8].

In order to further improve the prediction accuracy and model efficiency, the Transformer architecture has been introduced into the field of time series prediction in recent years. This structure can effectively capture long-distance time series dependencies through the self-attention mechanism and has good parallel computing capabilities. Improved versions such as Informer, Autoformer, FEDformer, etc [9]. have further improved performance by introducing sparse attention, trend decomposition, frequency domain enhancement and other mechanisms. This type of method has achieved remarkable results in the fields of power load forecasting and traffic flow forecasting. Some studies have also applied it to wind power forecasting, showing good development potential [10].

However, there are still some shortcomings in the current Transformer-based wind power forecasting research. For example, most models directly use the original multivariable inputs, do not model the importance of the variables, and easily introduce redundant information; at the same time, some methods have complex structures and high training costs, which are not conducive to deployment in actual wind farms. Therefore, designing a Transformer prediction model with a simple structure, strong modeling capabilities, and suitable for multivariate wind power data is still a problem with practical significance and research value [11].

3. Proposed Method

In order to solve the problems of long-term dependency modeling and insufficient variable coupling expression in wind power time series forecasting, this paper proposes an improved Transformer model structure, named MVTformer (Multi-Variable Time-series Transformer) [12]. Before introducing the structure of this model, this paper first reviews the traditional Transformer architecture and analyzes its limitations in time series modeling [13].

3.1 Traditional Transformer Structure

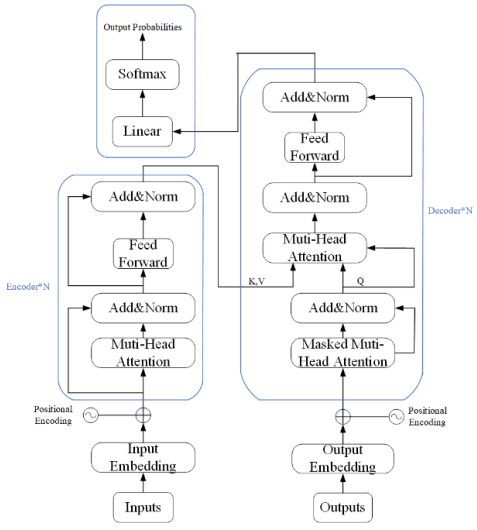

Fig.1 Transformer original structure diagram

ISSN 2616-5767 Vol.8, Issue 4: 79-86, DOI: 10.25236/AJETS.2025.080411

Transformer was first proposed by Vaswani et al. in the field of natural language processing, and its typical structure is shown in Figure 1. The model mainly consists of modules such as Multi-Head Attention, Feed Forward Network, and Add & Norm. After the input sequence passes through the embedding layer and position encoding, it is processed by the encoder and decoder networks in turn, and finally outputs the prediction result [14].

Although this architecture performs well in modeling long text contexts in natural language, it has the following problems when directly applied to wind power prediction tasks: Redundant calculation problem: The computational efficiency of the fully connected self-attention mechanism is low when processing long sequences; Insufficient feature fusion: The original structure does not explicitly consider the coupling relationship between multiple variables; Weak generalization ability: Faced with the problems of high noise and periodic instability in wind power data, the original Transformer lacks targeted design [15].

3.2 Model Overview

In order to improve the accuracy of wind power prediction, MVTformer introduces multivariate embedding, feature selection module and sparse attention mechanism based on the traditional Transformer, which effectively enhances the model's time series modeling ability and feature expression ability. The overall architecture of the model is shown in Figure 2.

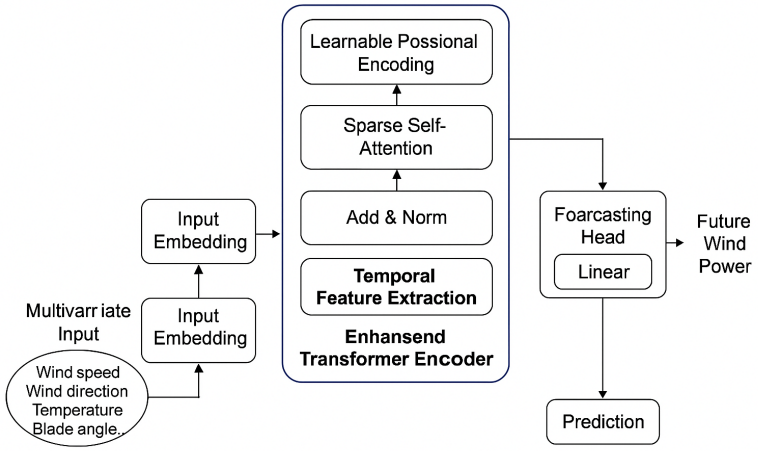

Fig.2 MVTformer structure diagram

3.3 Input Representation

Let the multivariate time series be denoted as:

$$X = \{x_1, x_2, \ldots, x_T\} \in \mathbb{R}^{T \times d} \tag{1}$$

Where T is the length of the time series (number of time steps), and d is the number of input features at each time step (e.g., wind speed, wind direction, temperature, blade angle, etc.).

Each input vector x_t is first projected into a common latent space of dimension d_{model} using a linear transformation:

$$E_t = W_e x_t + b_e \quad , \quad E \in \mathbb{R}^{T \times d_{model}} \tag{2}$$

To preserve temporal order information, we add a learnable positional encoding to the embedded sequence:

$$Z_0 = E + P \tag{3}$$

This final embedded representation Z_0 is then fed into the subsequent Transformer encoder layers for temporal modeling and feature extraction.

ISSN 2616-5767 Vol.8, Issue 4: 79-86, DOI: 10.25236/AJETS.2025.080411

3.4 Enhanced Transformer Encoder

The standard Transformer relies on global self-attention to compute dependencies between all time steps, resulting in a quadratic computational complexity of $O(T^2)$, where T is the sequence length. To improve efficiency and adapt the architecture to time series forecasting, we introduce a sparse self-attention mechanism, inspired by the Informer architecture, which preserves only the top-k attention scores.

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\text{Top}_k \left(\frac{QK^\top}{\sqrt{d_k}} \right) \right) V \tag{4}$$

Here, the query Q ,key K, and value V matrices are computed as:

$$Q = ZW_Q, \quad K = ZW_K, \quad V = ZW_V \tag{5}$$

To stabilize training and preserve feature propagation, we apply residual connections and layer normalization after each attention and feedforward block:

$$Z_{l+1} = \text{LayerNorm}(Z_l + \text{Attention}(Z_l)) \tag{6}$$

$$Z_{l+1} = \text{LayerNorm}(Z_{l+1} + \text{FeedForward}(Z_{l+1})) \tag{7}$$

In our architecture, we also include a temporal feature extraction module before the attention block, which enhances the model's ability to capture short-term patterns and multiscale dependencies.

By integrating sparse attention, residual learning, and temporal enhancement, the encoder is able to capture both local and global dependencies more efficiently and effectively, even in high-dimensional wind power time series data.

3.5 Forecasting Head

After passing through the Transformer encoder layers, the final hidden representation at the last time step Z_T is fed into a fully connected layer for regression to obtain the wind power prediction for the next H time steps.

$$\hat{y}_{t+1:t+H} = \text{MLP}(Z_T) \tag{8}$$

For multi-step forecasting scenarios, the output is a sequence of predicted values:

$$\hat{Y} = \{\hat{y}_{t+1}, \hat{y}_{t+2}, \ldots, \hat{y}_{t+H}\} \tag{9}$$

To improve the model's forecasting ability, we optionally adopt a recursive or direct prediction strategy. In the recursive approach, the output of each time step is fed back as part of the input for the next prediction. In the direct approach, the model is trained to predict all HHH future steps simultaneously.

3.6 Loss Function

To optimize the forecasting performance, we adopt the Mean Squared Error (MSE) as the primary loss function, which measures the average squared difference between the predicted and actual wind power values over the prediction horizon H. The loss function is defined as:

$$\mathcal{L}_{MSE} = \frac{1}{H} \sum_{i=1}^{H} \left(\hat{y}_{t+i} - y_{t+i} \right)^2 \tag{10}$$

This loss function penalizes large deviations more heavily and encourages the model to make accurate multi-step predictions. During training, the model parameters are updated via backpropagation using gradient descent–based optimizers such as Adam.

ISSN 2616-5767 Vol.8, Issue 4: 79-86, DOI: 10.25236/AJETS.2025.080411

4. Experiments and Results

4.1 Dataset Description

In this study, we utilize real-world operational data collected from a wind farm located in northern China. The dataset spans over 12 months and consists of high-frequency time series measurements related to wind turbine performance and environmental conditions. Each data point represents a 10-minute interval, resulting in over 50,000 timestamped records.

The dataset contains both meteorological variables and turbine operating parameters. These variables are used as input features to predict the future wind power output. Table 1 summarizes a portion of the dataset structure.

Table 1 Main Variables and Descriptions

Timestamp	Wind Speed (m/s)	Wind Direction (°)	Temperature (°C)	Blade Angle (°)	Historical Power (kW)	Target Power +1h (kW)
2023-05-01 00:00	5.24	198.3	16.2	3.5	302.6	355.1
2023-05-01 00:10	4.97	202.7	16.0	3.7	310.8	360.3
2023-05-01 00:20	5.18	205.2	15.9	3.6	328.2	374.0
...	...	...	...	...	...	...

In order to improve the generalization ability and convergence of the model, all input features are normalized to zero mean and unit variance. The dataset is divided into training set, validation set and test set according to the ratio of 8:1:1.

4.2 Experimental Setup

All experiments were conducted using Python 3.10 and the PyTorch 2.0 deep learning framework on a workstation equipped with an Intel Xeon Gold 5218 CPU, NVIDIA RTX 3090 GPU (24GB), and 128GB RAM running Ubuntu 20.04. The input sequence length was set to T = 36 (6 hours of historical data), and the prediction horizon was H = 6 (1 hour ahead). All models were trained using the Adam optimizer with early stopping based on validation loss.

Table 2 shows the parameter settings of the model.

Table 2 Hyperparameters and training settings

Parameter	Value
Input sequence length (T)	36
Forecast horizon (H)	6
Batch size	64
Learning rate	0.001
Optimizer	Adam
Number of epochs	100
Hidden dimension	128
Transformer layers	2
Attention heads	4
Dropout rate	0.1
Positional encoding	Learnable
Early stopping patience	10 epochs

We adopted the following commonly used metrics to assess the model performance:

Mean Absolute Error (MAE):

$$\mathrm{MAE} = \frac{1}{H}\sum_{i=1}^{H}\left|\hat{y}_{t+i} - y_{t+i}\right| \tag{11}$$

Root Mean Squared Error (RMSE):

$$\mathrm{RMSE} = \sqrt{\frac{1}{H}\sum_{i=1}^{H}\left(\hat{y}_{t+i} - y_{t+i}\right)^2} \tag{12}$$

Mean Absolute Percentage Error (MAPE):

ISSN 2616-5767 Vol.8, Issue 4: 79-86, DOI: 10.25236/AJETS.2025.080411

$$\text{MAPE} = \frac{1}{H}\sum_{i=1}^{H}\left|\frac{\hat{y}_{t+i} - y_{t+i}}{y_{t+i}}\right| \times 100\% \tag{13}$$

These metrics collectively reflect the average prediction error, the severity of large deviations, and the relative accuracy across different scales of wind power output.

4.3 Experimental Results and Analysis

To evaluate the effectiveness of the proposed MVTformer model, we compare it against several commonly used baselines, including traditional machine learning and deep learning approaches. These include ARIMA, LSTM, GRU, TCN, and a standard Transformer. All models were trained on the same dataset with identical input-output configurations for fairness.

We summarize the quantitative prediction results on the test set in Figure 3. Our proposed MVTformer outperforms all baseline models in terms of MAE, RMSE, and MAPE, demonstrating superior accuracy and stability.

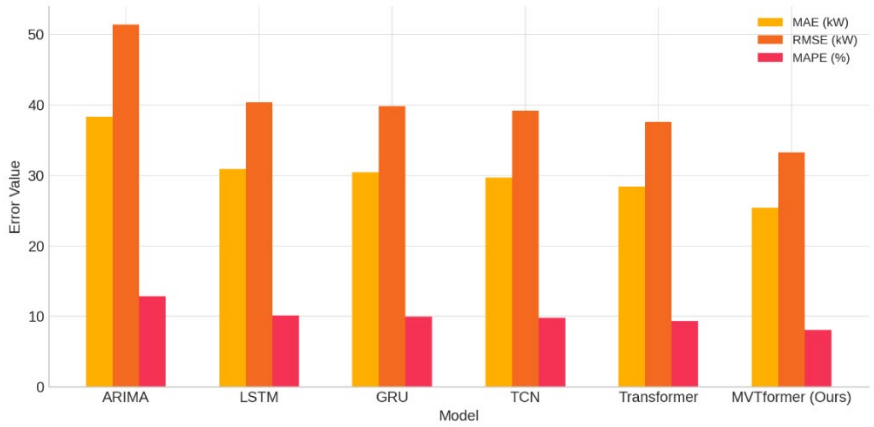

Fig.3 Performance comparison of different models on the test set

To better illustrate the forecasting performance, we visualize the predicted wind power versus the actual ground truth on several randomly selected time windows. As shown in Figure 4, MVTformer yields more accurate and smoother predictions with reduced lag and overshooting, especially during rapid fluctuations.

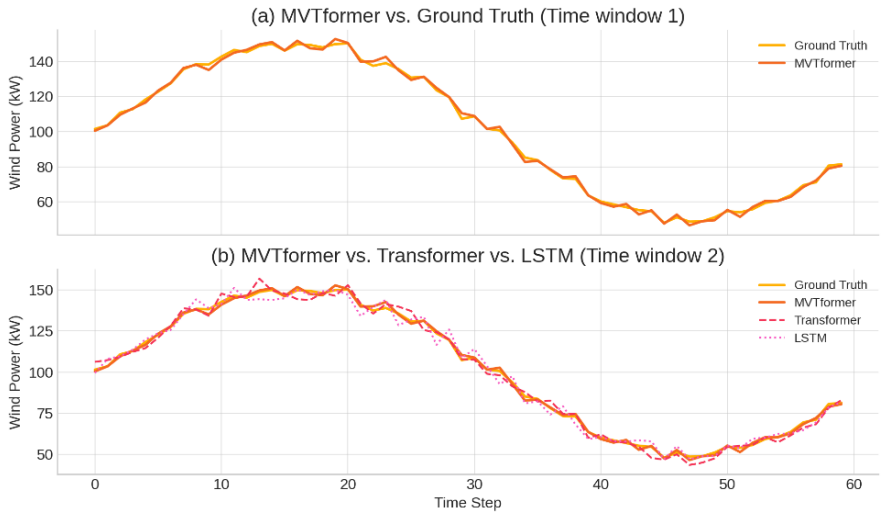

Fig.4 Prediction curve visualization

We also evaluate how the forecast accuracy changes with longer forecast horizons. As shown in Figure 5, MVTformer maintains lower MAE and RMSE at longer forecast steps (e.g., 1 to 6 steps ahead), indicating its stronger long-term modeling capability compared to the baseline method.

ISSN 2616-5767 Vol.8, Issue 4: 79-86, DOI: 10.25236/AJETS.2025.080411

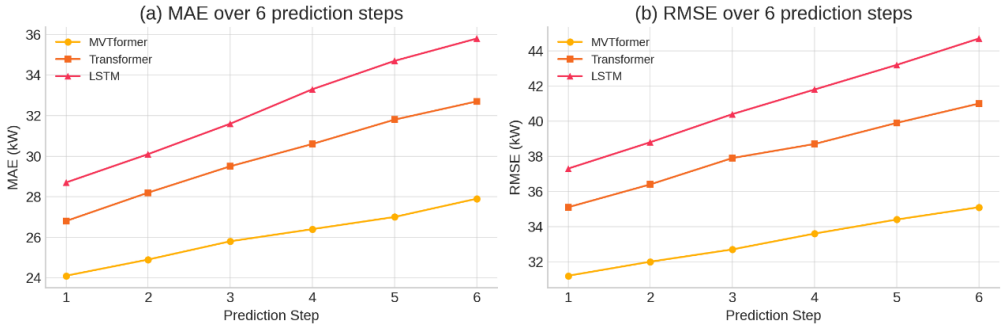

Fig.5 Multi-step forecast error trend chart

5. Conclusion

This paper proposes a new Transformer model for multivariate time series forecasting, MVTformer, which performs structural optimization on the strong multivariate dependency and time series characteristics in wind power forecasting. Compared with the traditional Transformer structure, MVTformer introduces a sparse self-attention mechanism and a time series feature extraction module, which significantly improves the modeling ability of local time series changes and long-term dependency patterns, and has stronger sequence feature learning capabilities.

In terms of implementation, MVTformer is developed based on the PyTorch deep learning framework, has a good modular structure and parallel computing capabilities, and is compatible with multi-GPU training environments. The design of the entire model fully reflects the scalability and efficiency of modern neural network engineering, and provides a generalizable reference architecture for complex industrial time series modeling.

Through experimental evaluation on a real wind farm dataset, MVTformer outperforms benchmark methods such as LSTM, GRU, TCN, and standard Transformer in terms of MAE, RMSE, and MAPE. Especially in multi-step forecasting tasks, the model shows a smoother error growth trend and stronger robustness, effectively alleviating the drift problem of traditional methods in long-term forecasting, and verifying its adaptability in actual complex scenarios.

In future work, we will further expand the capabilities of MVTformer, including introducing external meteorological information, enhancing the model's online learning capabilities to adapt to real-time deployment scenarios, and exploring model compression and distillation methods to adapt to edge deployment. These directions will promote the implementation and development of the Transformer structure in key applications such as smart grids and industrial control.

References

[1] D. Yang, B. Gao, S. Wang, and H. Xiang, "Robustness test for fouling state identification of homogeneous pressure electrodes based on confidence ellipsoids," IEICE Electron. Express, p. 22. 20240283, 2025, doi: 10.1587/elex.22.20240283.

[2] E. Arslan Tuncar, Ş. Sağlam, and B. Oral, "A review of short-term wind power generation forecasting methods in recent technological trends," Energy Reports, vol. 12, pp. 197–209, Dec. 2024, doi: 10.1016/j. egyr. 2024.06.006.

[3] N. Li, J. Dong, L. Liu, H. Li, and J. Yan, "A novel EMD and causal convolutional network integrated with transformer for ultra short-term wind power forecasting," Int. J. Electr. Power Energy Syst., vol. 154, p. 109470, Dec. 2023, doi: 10.1016/j.ijepes.2023.109470.

[4] S. Wang, J. Shi, W. Yang, and Q. Yin, "High and low frequency wind power prediction based on transformer and BiGRU-attention," Energy, vol. 288, p. 129753, Feb. 2024, doi: 10.1016/j.energy. 2023. 129753.

[5] J. Cheng, X. Luo, and Z. Jin, "Integrating domain knowledge into transformer for short-term wind power forecasting," Energy, vol. 312, p. 133511, Dec. 2024, doi: 10.1016/j.energy.2024.133511.

[6] T. B. M. J. Ouarda and F. Houndekindo, "LSTM and transformer-based framework for bias correction of ERA5 hourly wind speeds," 2025, SSRN. doi: 10.2139/ssrn.5125439.

[7] H. Zhou, "Informer: beyond efficient transformer for long sequence time-series forecasting," Proc. AAAI Conf. Artif. Intell., vol. 35, no. 12, Art. no. 12, May 2021, doi: 10.1609/aaai.v35i12.17325.

ISSN 2616-5767 Vol.8, Issue 4: 79-86, DOI: 10.25236/AJETS.2025.080411

[8] H. Eskandari, M. Imani, and M. P. Moghaddam, "A deep residual network integrating entropy-based wavelet packet ensemble model for short-term electrical load forecasting," Energy, vol. 314, p. 134168, Jan. 2025, doi: 10.1016/j.energy.2024.134168.

[9] A. Ijadi Maghsoodi, A. E. Torkayesh, L. C. Wood, E. Herrera-Viedma, and K. Govindan, "A machine learning driven multiple criteria decision analysis using LS-SVM feature elimination: sustainability performance assessment with incomplete data," Eng. Appl. Artif. Intell., vol. 119, p. 105785, Mar. 2023, doi: 10.1016/j.engappai.2022.105785.

[10] S. Abolhosseini, M. Khorashadizadeh, M. Chahkandi, and M. Golalizadeh, "A modified ID3 decision tree algorithm based on cumulative residual entropy," Expert Syst. Appl., vol. 255, p. 124821, Dec. 2024, doi: 10.1016/j.eswa.2024.124821.

[11] E. G. S. Nascimento, T. A. C. De Melo, and D. M. Moreira, "A transformer-based deep neural network with wavelet transform for forecasting wind speed and wind energy," Energy, vol. 278, p. 127678, Sep. 2023, doi: 10.1016/j.energy.2023.127678.

[12] A. Redekar, H. S. Dhiman, D. Deb, and S. M. Muyeen, "On reliability enhancement of solar PV arrays using hybrid SVR for soiling forecasting based on WT and EMD decomposition methods," Ain Shams Eng. J., vol. 15, no. 6, p. 102716, Jun. 2024, doi: 10.1016/j.asej.2024.102716.

[13] M. Barman and N. B. Dev Choudhury, "Season specific approach for short-term load forecasting based on hybrid FA-SVM and similarity concept," Energy, vol. 174, pp. 886–896, May 2019, doi: 10.1016/j.energy.2019.03.010.

[14] A. Dupré, P. Drobinski, B. Alonzo, J. Badosa, C. Briard, and R. Plougonven, "Sub-hourly forecasting of wind speed and wind energy," Renewable Energy, vol. 145, pp. 2373–2379, Jan. 2020, doi: 10.1016/j.renene.2019.07.161.

[15] J. G. Martin, J. R. D. Frejo, J. M. Maestre, and E. F. Camacho, "Spatio-temporal kriging for spatial irradiance estimation with short-term forecasting in a thermosolar power plant," Heliyon, vol. 10, no. 20, p. e39247, Oct. 2024, doi: 10.1016/j.heliyon.2024.e39247.