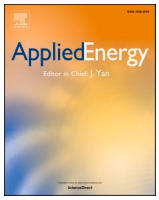
Applied Energy 374 (2024) 124035

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy



Interpretable feature-temporal transformer for short-term wind power forecasting with multivariate time series

Lei Liu [a], Xinyu Wang [a], Xue Dong [b], Kang Chen [a], Qiuju Chen [a,c], Bin Li [a,*]

[a] School of Information Science and Technology, University of Science and Technology of China, Hefei, 230022, China
[b] Key Laboratory of Far-Shore Wind Power Technology of Zhejiang Province, Hangzhou, 311122, China
[c] Laboratory for Big Data and Decision, Changsha, 410037, China

HIGHLIGHTS

- This paper introduces the interpretable feature-temporal transformer (IFTT) model, which enhances wind power forecasting by effectively integrating historical information and future prior information from multiple variables.
- The designed decoupled feature-temporal self-attention (DFTA) module and variable attention network (VAN) ensure the interpretability of temporal information and multi-variable inputs, allowing for the extraction of important features.
- Experimental results on multiple datasets in different geographical locations demonstrate the superior performance of the proposed IFTT algorithm compared to various advanced methods, highlighting its potential to improve the accuracy of WPF.
- The interpretability of the IFTT model provides valuable insights for ensuring the safe and reliable utilization of wind power, enabling informed decision-making and risk assessment in the context of wind power integration into the grid.

ARTICLE INFO

Keywords:
Short-term
Wind power forecasting
Interpretable
Transformer
Self-attention

ABSTRACT

The inherent randomness and volatility of wind power generation present significant challenges to the reliable and secure operation of the power system. Therefore, it is crucial to have interpretable wind power forecasting (WPF) to ensure seamless grid integration and effective risk assessment. Existing forecasting models often focus on improving WPF performance and ignore the interpretability of the model, resulting in ambiguous forecasting. In this paper, the interpretable feature-temporal transformer (IFTT) for short-term wind power forecasting with multivariate time series is presented. The model uses an encoder-decoder architecture to effectively integrate historical information and future prior information from multiple variables. The designed decoupled feature-temporal self-attention (DFTA) module and variable attention network (VAN) effectively realize the interpretability of temporal information and multi-variable inputs while extracting important features. The Auxiliary Forecasting Network (AFN) plays a key role in providing pseudo-future wind speed predictions, which serve as an essential input for the model's decoder, and enhancing forecasting accuracy through multi-task learning. Experimental results on multiple datasets in different geographical locations show that the proposed algorithm is superior to various advanced methods. Besides, the interpretability of the IFTT model offers valuable insights for ensuring the safety of wind power utilization and supporting informed risk decision-making.

1. Introduction

In response to the global climate change resulting from the conventional use of fossil fuels, the global energy sector is undergoing an energy revolution [1]. Wind energy, as a clean and renewable energy source, holds several advantages, including zero emissions, no pollution, and no fuel cost, positioning it as the third-largest energy source worldwide, following thermal power and hydroelectric power [2]. However, the inherent randomness and intermittency of wind power generation present significant challenges to the secure and stable operation of the power grid. Consequently, the credible and accurate forecasting of wind power becomes crucial for effective generation

* Corresponding author.
E-mail address: binli@ustc.edu.cn (B. Li).

https://doi.org/10.1016/j.apenergy.2024.124035
Received 26 January 2024; Received in revised form 8 June 2024; Accepted 23 July 2024

planning, reliability management, risk mitigation and real-time decision-making [3].

Wind power forecasting can be divided into different time scales: ultra-short-term (a few seconds to 30 min ahead), short-term (30 min to 6 h ahead), medium-term (6 h to 1 day ahead), and long-term (1 day or more ahead) [4–6]. Due to the ability of short-term and ultra-short-term wind power forecasting to provide reliable transient power information for power dispatch and grid safety [7], they have received significant attention. Wind power forecasting methods can be roughly classified into four types: physical methods, traditional statistical methods, artificial intelligence methods, and hybrid methods [8–11]. Physical methods use physical modeling to transform weather parameters into predicted wind power curves, and then extrapolate to obtain the trend of the wind power series [6,12]. However, physical modeling is computationally complex, and its effectiveness is limited in short-term and ultra-short-term wind power forecasting due to the lack of trend characteristics in historical data [13,14]. Traditional statistical methods, such as autoregressive moving average (ARMA [15]), autoregressive integrated moving average (ARIMA [16]), and Hammerstein autoregressive models [17], predict by establishing the underlying relationship between historical data and forecasted power. Traditional statistical models struggle to handle nonlinear time series, which also hinders their development in wind power forecasting [8].

With the development of computer science, many artificial intelligence models have been applied to wind power forecasting. Deep learning, as an important branch of artificial intelligence, has powerful feature extraction and non-linear mapping capabilities, making it a research hotspot in the field of wind power forecasting [18]. A study by Yu et al. [19] proposed an improved LSTM network with an enhanced forget gate, which analyzed wind power data using this network and improved the accuracy of wind power forecasting. Kisvari et al. [20] constructed input features using different wind speeds, generator temperatures, and gearbox temperatures at various heights, and utilized a GRU network to learn the non-linear mapping relationship between the input features and wind power, achieving wind power forecasting. Numerous studies have shown that AI-based models outperform traditional models in terms of predictive performance [6].

Due to the volatility and suddenness in wind speed, the aforementioned models often struggle to extract complex feature correlations from nonlinear and non-stationary wind speed/power data. To overcome this issue, many WPF studies have focused on hybrid models [21]. Duan et al. [22] combined feature attention mechanisms with Bayesian neural networks to achieve accurate WPF while also assessing the uncertainty of the predictions. Wang et al. [23] employed CNN for multi-scale information fusion, followed by the use of BiLSTM for time information extraction, and described the uncertainty of wind power predictions using an asymmetric Laplace distribution. Duan et al. [22] combined time attention mechanisms with BiLSTM to improve the accuracy of ultra-short-term WPF. Shahid et al. [24] utilized the global optimization capability of the genetic algorithm (GA) to determine the optimal time window size and number of neurons in the LSTM network, thereby enhancing the predictive capacity of the model in short-term WPF tasks. Abou Houran et al. [25] optimized the CNN-LSTM model using the Coati Optimization Algorithm (COA) [26], resulting in improved accuracy in photovoltaic/wind power forecasting. Niu et al. [27] combined an enhanced variational mode decomposition (VMD), BiLSTM, and attention mechanisms to achieve reliable wind power interval forecasting. Ding et al. [28] proposed a hybrid forecasting model based on complementary ensemble empirical mode decomposition (CEEMD) and kernel extreme learning machine (KELM) with whale optimization algorithm (WOA) for short-term WPF. Lu et al. [29] applied VMD to historical wind power data and used the decomposed signals along with key meteorological factors as inputs to a model that combined CNN and LSTM for future wind power prediction. Zhang et al. [30] proposed a hybrid model for WPF by combining discrete wavelet transform, seasonal autoregressive integrated moving average, and LSTM. Hybrid models can be effectively developed by synergistically combining diverse models to comprehensively capture and characterize the multifaceted fluctuations in wind speed/power [6]. Table 1 provides a comprehensive summary of numerous studies on WPF, offering an

Table 1
The summary of selected WPF studies.

Classification	Timescales	Input variables	Forecasting Methods
Physical models	Medium-term	Physical processes in atmosphere	Computational fluid dynamics (CFD) [31]
	Medium-term	Physical processes in atmosphere	Clustering Pre-Calculated CFD Method [32]
Traditional statistical models	Short-term	Historical wind power	Combined with Pattern-matching and ARMA-model [15]
	Ultra-short-term	Historical wind power	ARIMA [16]
	Ultra-short-term	Historical wind speed and direction	Hammerstein wind power forecasting model [17]
Artificial intelligence models	Ultra-short-term	Historical wind speeds, historical generator temperature	Gated Recurrent Unit (GRU) [20]
	Short-term	Historical wind power	Improved Long Short-Term Memory-enhanced forget-gate network model [19]
	Ultra-short-term	Historical wind power	BPNN [13]
	Ultra-short-term	Historical wind power, historical wind speed, and historical temperature	A novel nonlinear combination forecasting method based on PSO-DBN model is proposed [22]
	Ultra-short-term	Historical wind speed and historical wind power	A hybrid model based on deep Bayesian model and feature attention mechanism [33]
	Short-term	Historical wind power	A hybrid model based on asymmetric Laplace, multi-convolutional neural network, and bidirectional long-short-term memory network [23]
	Short-term	Historical wind power	An integrated method based on the combination of a CEEMD decomposition model and WOA-KELM [28]
Hybrid models	Ultra-short-term	Historical wind speed and historical wind power	A hybrid model based on Discrete Wavelet Transform (DWT), Seasonal Autoregressive Integrated Moving Average (SARIMA), and Deep-learning-based Long Short-Term Memory (LSTM) [30]
	Short-term	Historical wind power	A hybrid model based on VMD and GRU [34]
	Ultra-short-term	Historical wind power	A novel hybrid model based on Bernstein polynomial with mixture of Gaussians [8]
	Short-term	Historical wind power, historical wind speed, and historical wind direction	A novel hybrid model based on LSTM and GA [24]
	Ultra-short-term	Historical wind power and historical wind speed	A novel hybrid model based on PSO and Adaptive neuro-fuzzy inference system (ANFIS) [35]

overview of the types of input variables, classification, and forecasting methods employed.

While extensive research has been conducted on WPF, there are still some existing issues: (1) Existing methods mainly focus on improving the accuracy of WPF while neglecting the interpretability of the models, leading to lower credibility of the predicted results. Recently, Lim et al. [36] proposed an interpretable model for time series modeling called temporal fusion transformer (TFT), which effectively explains the importance of different variables at various time lags. However, the model neglects the influence of information selection and fusion mechanisms in time series modeling, as it calculates the importance of time series after the time series modeling process (LSTM module). This leads to certain inaccuracies in the analysis results of time series information importance. (2) Existing WPF methods often only consider the mapping relationship between historical classical variables (e. g. historical wind power/speed) and wind power, neglecting the highly coupled relationship between the historical/future states of meteorological factors, time-related variables, and wind power in the lookback window and horizon window. There is a lack of research on WPF models to couple historical and known future information.

To address these issues, this study proposes an interpretable feature-temporal self-attention transformer (IFTT) for short-term wind power forecasting with multivariable. The model adopts an encoder-decoder architecture that effectively integrates historical information of classical variables (historical wind speed and historical wind power), meteorological factors, and time-related variables, as well as prior known future information and pseudo-future information. The designed AFN serves a dual purpose in the model. It not only provides pseudo-future wind speed as an input, enhancing the model's predictive capabilities, but it also establishes a multi-task learning framework by combining it with the wind power forecasting task. This multi-task learning approach leverages the high correlation between the two tasks, providing additional supervision and improving overall performance. In the encoder-decoder structure, the DFTA and VAN enable the interpretability of time series information and historical/future information of multivariate input variables. This study provides a basis for optimizing WPF models and assisting decision-making through interpretive analysis of historical and future information from time series and meteorological data. The main contributions of this paper are as follows:

1) For the first time, interpretive analysis is performed on different dimensions in the time series modeling of WPF. This not only identifies the important variables for WPF by focusing on the importance of different variables for the predicted results but also analyzes the impact of different time lags on the predicted results to analyze sustained temporal patterns and obtain optimal memory window size.
2) The designed DFTA module decouples multi-variable feature attention and temporal attention, thereby avoiding the impact of intra-temporal feature attention on the learning of inter-temporal attention and ensuring the effective learning of temporal attention.
3) Multiple meteorological data and time-related information are incorporated into the WPF task, fully utilizing the nonlinear coupling relationship between multiple relevant variables and wind power to improve the performance of WPF. The VAN network achieves an interpretability analysis of each variable.
4) The designed AFN not only provides pseudo-future wind speed to the model but also forms a multi-task learning paradigm with the wind power forecasting task, utilizing highly correlated multi-task learning to provide additional supervision.

The subsequent sections of this study are structured as follows: Section 2 delves into the intricate details of the proposed methodology. Section 3 elucidates the dataset, input selection, and preprocessing techniques employed. Section 4 showcases the principal findings, accompanied by comprehensive discussions. Lastly, Section 5 encapsulates the key conclusions drawn from this study.

2. Method

2.1. IFTT model

Fig. 1 shows the overall architecture of the IFTT model. From the figure, it can be observed that the IFTT model adopts an encoder-decoder architecture. The encoder is responsible for processing historical time series information, while the decoder handles future time series information and output results of WPF. Both the encoder and decoder consist of an input transformation layer, Variable Attention Network (VAN), Decoupled Feature-Temporal Self-Attention Network (DFTA), LSTM for temporal information modeling module, Gate Residual Network (GRN), and Multi-Head Self-Attention (MHA). The input transformation layer maps the input variables to a high-dimensional space to facilitate subsequent feature extraction. The VAN is used for effective selection and fusion of multiple input variables, providing interpretability for the importance of input variables. To prevent the influence of intra-temporal feature attention on the learning of inter-temporal attention and to guarantee the successful learning of temporal attention, the developed DFTA module decouples multi-variable feature attention and temporal attention. This module enables efficient temporal importance interpretation and further improves model performance. The LSTM temporal information modeling module extracts and models temporal features from historical/future information, while the GRN selects an appropriate degree of non-linear processing for feature selection. The encoder and decoder are connected through a hidden MHA to achieve effective fusion of historical and future information. In addition, the decoder includes an auxiliary forecasting network (AFN) for obtaining pseudo-future wind speed and a linear layer for final regression, with the obtained pseudo-future wind speed serving as input to the decoder. This study utilizes the correlation between wind speed and wind power to form a multitask learning framework, where WPF and pseudo-future wind speed forecasting mutually supervise each other, thereby improving the performance of WPF.

The input of the IFTT model consists of three parts: historical information, known future prior information, and pseudo-future wind speed. The historical information and known future prior information include classic variables (wind speed and wind power), various meteorological variables (Wind direction, Air temperature, Pressure), and time-related variables (Hour, Day, Month). The IFTT model not only achieves accurate WPF but also utilizes the VAN and DFTA to provide multiple interpretable aspects in wind power forecasting tasks.

2.1.1. Transformation layer

In wind power forecasting, this study integrates classic variables, meteorological variables, and time-related variables as inputs to the forecasting model. Therefore, the input variables consist of both discrete and continuous variables. For discrete variables, this study employs label encoding and linear transformation to convert each variable into a D_i^d-dimensional representation vector. The value of D_i^d is obtained using the empirical guideline from the literature [37]:

$$D_i^d = min\left(round\left(A^*(n_i)^B\right), \widetilde{D}_i^d\right) \tag{1}$$

where n_i and $\widetilde{D}_i$ represent the number of values and the predefined maximum embedding size of the i-th discrete variable, respectively. The $round(\bullet)$ function returns a rounded integer number. A and B are adjustable parameters.

For continuous variables, a linear mapping layer is used to map the input variables to a D_i^c-dimensional representation vector. To facilitate computational convenience, an additional linear transformation is applied to transform each representation vector into a specific H-dimensional representation space for ease of computation.

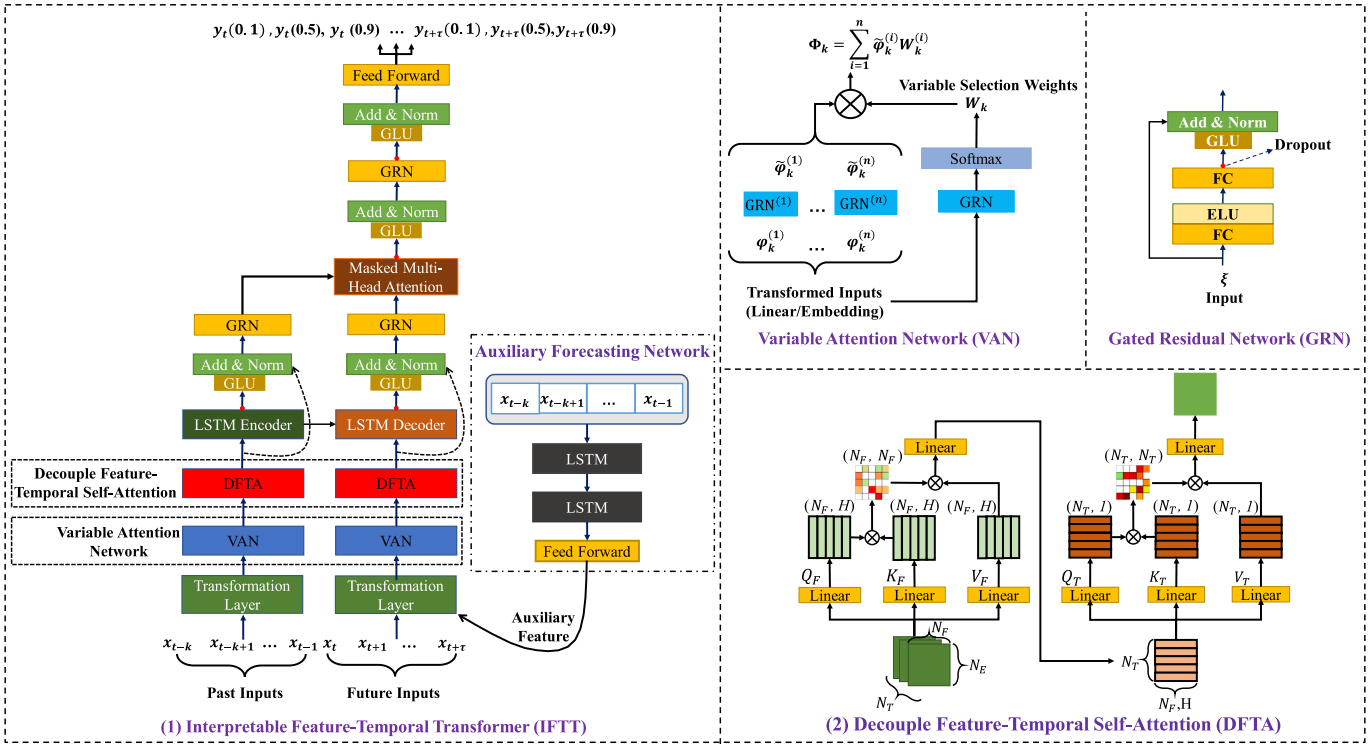


Fig. 1. The network architecture of IFTT.

2.1.2. Gated residual network (GRN)

Due to the inability to determine the exact mapping relationship between feature vectors and the forecasting target, it is difficult to determine the degree of non-linear processing required for each feature vector. In order to give the network flexible non-linear mapping ability, this study constructs a GRN network with a residual structure and introduces a gate mechanism for information selection. The structure of the GRN network is shown in Fig. 1, and for any input vector ξ, we have:

$$\mathrm{GRN}(\xi) = Norm(\xi + GLU(\eta)) \tag{2}$$

$$\eta = \mathrm{ELU}(\xi W_1 + b_1) W_2 + b_2 \tag{3}$$

$$GLU(\eta) = sigmoid(\eta W_3 + b_3) \odot (\eta W_4 + b_4) \tag{4}$$

where ELU refers to the exponential linear activation unit [38], W_i and b_i represent the weight and bias of the corresponding layer, and $\odot$ denotes the element-wise Hadamard product. Gated Linear Unit (GLU) utilizes the sigmoid function to implement the gate mechanism for information selection. When the information is highly responsive, GLU allows all the information to pass through; when the information is not responsive, the output of GLU is almost zero.

2.1.3. Variable attention network (VAN)

The correlation between input variables and the forecasting target cannot be accurately predicted. In order to achieve effective variable selection and suppress possible noise effects, inspired by [36], this study introduces a Variable Attention Network (VAN) to achieve effective selection and interpretability of input variables. The input of the VAN network is the feature representation obtained from the original input variables after the Transformation layer. Let $\varphi_k^{(i)}$ represent the embedding vector of the i-th variable at time step k, and $\Psi_k = \left[\varphi_k^{(1)}, \varphi_k^{(2)}, \ldots, \varphi_k^{(N)}\right]$ represent the set of embedding vectors of all input variables at time step k. All input variables are independently activated with important information using separate GRU networks:

$$\widetilde{\varphi}_k^{(i)} = GRN_i\left(\varphi_k^i\right) \tag{5}$$

At the same time, the importance of each input variable is obtained using the softmax function to achieve information selection of the input variables. The output Φ_k of the VAN network can be expressed as:

$$\Phi_k = \sum_{j=1}^{N} \widetilde{\varphi}_k^{(i)} W_k^{(i)} \tag{6}$$

$$W_k = \mathrm{Softmax}(GRN(\Psi_k)) \tag{7}$$

where W_k is the variable selection weight vector.

2.1.4. Decoupled feature-temporal self-attention (DFTA)

In the standard Self-Attention module (Fig. 2(1)), given the input matrix $X \in \mathbb{R}^{N_T N_F \times N_E}$, three linear projections are applied to obtain the query (Q), key (K), and value (V) representations, where N_T represents the number of time dimensions, N_F represents the number of feature dimensions, and N_E is the embedding dimension of the features. Therefore, the projection matrices for Self-Attention can be obtained using the following equation:

$$Q = X\boldsymbol{W}_Q, K = X\boldsymbol{W}_K, V = X\boldsymbol{W}_V \tag{8}$$

The attention mechanism is then calculated as:

$$Attention(Q, K, V) = \boldsymbol{AV} = softmax\left(\frac{\boldsymbol{QK}^{\mathrm{T}}}{\sqrt{H}}\right)\boldsymbol{V} \tag{9}$$

where $A \in \mathbb{R}^{N_T N_F \times N_T N_F}$, $V \in \mathbb{R}^{N_T N_F \times H}$, and H is the hidden dimension.

The final output Y of the attention is merged with the backbone network using a linear projection $W^O \in \mathbb{R}^{TF \times H}$:

$$\boldsymbol{Y} = \boldsymbol{X} + Attention(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})\boldsymbol{W}^O \tag{10}$$

In this study, we aim to model the temporal information of multivariable features. The standard self-attention module learns attention between temporal information and attention between different variable features at the current time step (Fig. 2(1)). This coupling of attention

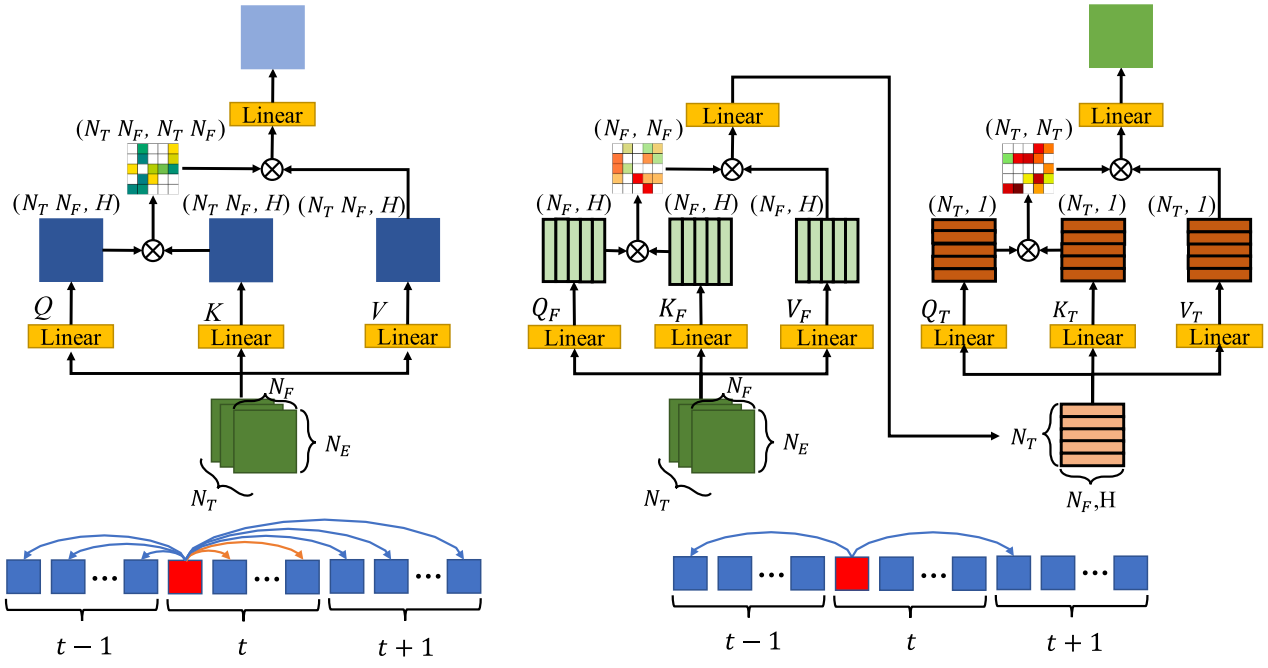


Fig. 2. Standard and decoupled self-attention.

between multi-variable features and temporal information hinders the effective learning of temporal attention. When the query, key, and value share the same transformation matrix, the model cannot distinguish between multi-variable features and temporal context. Interpretable temporal importance is crucial for time series modeling. To effectively learn temporal attention in the modeling of multi-variable feature sequences, we propose a decoupled feature-temporal self-attention (DFTA), as shown in Fig. 2(2). We first decompose the input containing multi-variable features into temporally independent sub-inputs and learn attention between features within each time step. Then, we apply temporal attention to the output of feature attention. The DFTA module effectively avoids the influence of attention between features within a time step on the learning of temporal attention. It enables the model to learn temporal attention only between corresponding features across time steps, further improving its effectiveness.

For a given input matrix $\boldsymbol{X} \in \mathbb{R}^{N_T N_F \times N_E}$, we decompose it into N_T temporally independent input information $X(t) \in \mathbb{R}^{N_F \times N_E}$, and then obtain three projections using the following equation:

$$Q_F(t) = X(t)W_Q^F(t), K_F(t) = X(t)W_K^F(t), V_F(t) = X(t)W_V^F(t) \tag{11}$$

We perform the feature-only attention on all feature positions at each time step:

$$Attention_F(t)(Q,K,V) = A_F(t)V_F(t) = softmax\left(\frac{Q_F(t)K_F(t)^{\mathrm{T}}}{\sqrt{H}}\right)V_F(t) \tag{12}$$

$$Y_F(t) = X(t) + A_F(t)V_F(t)W_F^O \tag{13}$$

where $t \in \{1, 2, \ldots, N_T\}$ represents different time steps, and $A_F(t) \in \mathbb{R}^{N_F \times N_F}$.

The output of the feature attention block is then used as input to the temporal attention block, which performs temporal-only attention operations:

$$Q_T = Y_F W_Q^T, K_T = Y_F W_K^T, V_T = Y_F W_V^T \tag{14}$$

$$Attention_T(f)(Q,K,V) = A_T(f)V_T(f) = softmax\left(\frac{Q_T(f)K_T(f)^{\mathrm{T}}}{\sqrt{H}}\right)V_T(f) \tag{15}$$

$$Y_T(f) = Y_T(f) + A_T(f)V_T(f)W_T^O \tag{16}$$

where $f \in \{1, 2, \ldots, N_F\}$ represents different feature, and $A_T(f) \in \mathbb{R}^{N_T \times N_T}$.

2.1.5. LSTM encoder/decoder module

The long short-term memory (LSTM) network overcomes the long-term memory loss problem (vanishing gradient problem) in the recurrent neural network (RNN). The core of the LSTM module is the cell state, which determines the retention and forgetting of information. LSTM units control the cell state by designing three gates: the input gate, forget gate, and output gate. The forget gate decides which information to discard from the cell state, the input gate determines how much input information to retain in the cell state, and the output gate determines which information to output from the cell state.

The LSTM Encoder/Decoder module is a sequential modeling module constructed by connecting multiple LSTM units, where each LSTM unit receives input information from a time lag/future time step. The LSTM Encoder module processes historical information, while the LSTM Decoder module handles future information. Additionally, the hidden information from the Encoder is used in the Decoder.

By connecting multiple LSTM units in the Encoder and Decoder, the model can capture temporal dependencies and make forecasting based on historical information. The LSTM Encoder/Decoder module, combined with the DFTA module, forms the backbone of our proposed model for effective time series modeling.

2.1.6. Masked multi-head attention

To enhance the learning capability of the standard attention mechanism, the paper by [39] introduced a concept called multi-head attention (MHA), which uses different heads for different representation subspaces. The formulation is as follows:

$$MultiHead(Q,K,V) = Concat(head_1,\ldots,head_H)W^H \tag{17}$$

$$head_h = Attention\left(QW_h^Q, KW_h^K, VW_h^V\right) \tag{18}$$

where $W_h^Q \in \mathbb{R}^{d_{model}\times d_{attn}}$, $W_h^K \in \mathbb{R}^{d_{model}\times d_{attn}}$ and $W_h^V \in \mathbb{R}^{d_{model}\times d_{attn}}$ are head-specific weights for the query, key, and value, respectively. $W^H \in \mathbb{R}^{(H\times d_v)\times d_{model}}$ linearly combines the concatenated outputs of all the heads. Typically, $d_v = d_{attn} = d_{model}/H$.

During the forecasting phase of the model, when decoding, future information is unknown. Therefore, in this study, masked multi-head attention is used to integrate the attention between the encoder and decoder information. This is achieved by using masking to ensure that only the known information up to the current time step is used during the computation of the Multi-Head Attention, thus preventing the model from "peeking" at the answers during decoding.

2.1.7. Auxiliary forecast network

In time-series forecasting tasks, incorporating additional useful variables and their future information can improve the performance of forecasting models. In the traditional encoder-decoder architecture, the encoder's output is typically used as the input for the decoder, enabling feature decoding and forecasting. In the encoder-decoder architecture proposed in this study, besides preserving the encoder's output as the decoder's input, prior knowledge of future information is also added to further enhance the predictive performance of the model. In wind power forecasting tasks, wind speed is a crucial variable highly related to the forecasting target. Therefore, in addition to known meteorological variables, this study introduces an Auxiliary Forecast Network (AFN) (see Fig. 1. for the network structure) that utilizes historical wind speed information to generate pseudo-future wind speed, which is then used as input for the decoder of IFTT model:

$$\hat{x}_{t:t+\tau}^{pf} = AFN\left(x_{t-\tau:t}^{pf}\right) \tag{19}$$

In the AFN network, the features are encoded and decoded using two LSTM layers, and a Feed Forward network [39] is used to obtain the pseudo-future wind speed. The AFN network is a lightweight encoder-decoder structure that balances effective forecasting and model size control.

2.2. Multi-task learning

Wind speed is highly correlated with wind power, and incorporating future wind speed data has significant implications for improving the performance of WPF. However, the volatility and suddenness in wind speed make it impossible to have prior knowledge of future wind speed. Therefore, this paper introduces the Auxiliary Forecasting Network (AFN) to obtain pseudo-future wind speed, which is then used as input for the IFTT model. To ensure the quality of the generated pseudo-future wind speed data and facilitate efficient training of the forecasting model, multi-task learning is introduced into the training of the IFTT model, enabling the two highly correlated forecasting tasks to supervise each other and improve the forecasting performance of the IFTT model.

The WPF task is the primary task, while the pseudo-future wind speed forecasting is the auxiliary task. The primary task is optimized by minimizing the sum of quantile losses [40], L_q over multiple quantiles q and all time points T to forecast:

$$L_{pri} = \sum_{t\in T}\sum_{q\in Q} L_q\left(y_t^q, \hat{y}_t^q\right) \tag{20}$$

where,

$$L_q\left(y_t^q, \hat{y}_t^q\right) = \begin{cases} q\left(y_t^q - \hat{y}_t^q\right), if\ \hat{y}_t^q \le y_t^q \\ (1-q)\left(\hat{y}_t^q - y_t^q\right), if\ \hat{y}_t^q > y_t^q \end{cases} \tag{21}$$

The quantile loss provides accurate numerical forecasting and insights into forecasting uncertainty by providing multiple forecasting intervals. The set of quantiles used in this study is $\Omega = [\,0.1, 0.5, 0.9]$.

The auxiliary task is optimized using the classical regression loss L_{aux}:

$$L_{aux} = \|y_t - \hat{y}_t\|_2^2 \tag{22}$$

Thus, the multi-task joint loss L_{IFTT} of the IFTT model is expressed as:

$$L_{IFTT} = L_{pri} + \lambda L_{aux} \tag{23}$$

where λ is the weight balancing the primary and auxiliary tasks.

2.3. Interpretable wind power forecasting procedure

Intelligent optimization algorithms have been proven to effectively configure the hyperparameters of deep learning models [41]. In this study, the hyperparameters to be optimized in the IFTT model are first determined, including the number of units in DFTA and VAN, the number of heads of MHA, learning rate and batch size, dropout rate, and weight of loss. The COA algorithm is employed to intelligently optimize the aforementioned parameters of the IFTT model, which has been demonstrated to effectively balance exploration capability and convergence speed and be able to converge to the global optimum in an efficient and timely manner [42]. The initial parameters of the population are set as random combinations within the specified parameter range, and the fitness function is defined as 1/loss for selecting the optimal solution. Under the optimization of the COA algorithm, the optimal configuration of the IFTT structure is obtained.

Fig. 3 illustrates the forecasting process of our proposed method, including dataset construction, input sequence partitioning, COA optimization, IFTT model forecasting, and interpretable analysis. The steps of the IFTT model forecasting process are summarized as follows:

Step 1: Construct a dataset using collected variable data, and divide the input sequences based on their time series.

Step 2: Feed the time series sequences of the divided variables into the COA-based IFTT model for WPF. Additionally, input the time series sequence of wind speed into the AFN to predict pseudo-future wind speed, and use the obtained pseudo-future wind speed as input for the COA-based IFTT model to achieve accurate wind power forecasting through multi-task learning.

Step 3: Forecasted results and interpretable analysis. Evaluate the forecasting results using metrics such as NMAE, NRMSE, MAPE, accuracy, and QR on four datasets. Simultaneously, perform interpretable analysis on past inputs, future inputs, and different lag order information to achieve interpretable wind power forecasting.

3. Dataset and input

3.1. Dataset

In this study, we conducted model training and testing on multiple datasets in different geographical locations (the Texas turbine dataset [43] and Fujian turbine dataset [44]). We took the Texas turbine dataset [43] as an example to (default) compare and introduce our method. The Texas Turbine dataset comprises a full year of data from a specific year in Texas, USA. The temporal resolution of the data is 1 h, with 24 points per day, spanning a total of 365 days, resulting in 8760 sample points. It includes the required wind power data and meteorological variables such as wind direction, pressure, and air temperature. According to a report by the Financial Times, wind energy surpassed coal for the first time in the overall energy structure of the state in 2020. Wind energy resources in Texas are abundant and exhibit strong seasonality. Therefore, we divided the entire year's data into seasons: Spring (February to April), Summer (May to July), Autumn (August to October), and Winter (November, December, and January). The training data, validation data, and test data for each season were separated in an 80%:10%:10% ratio (as shown in Fig. 4, wind power changes with time). The three groups

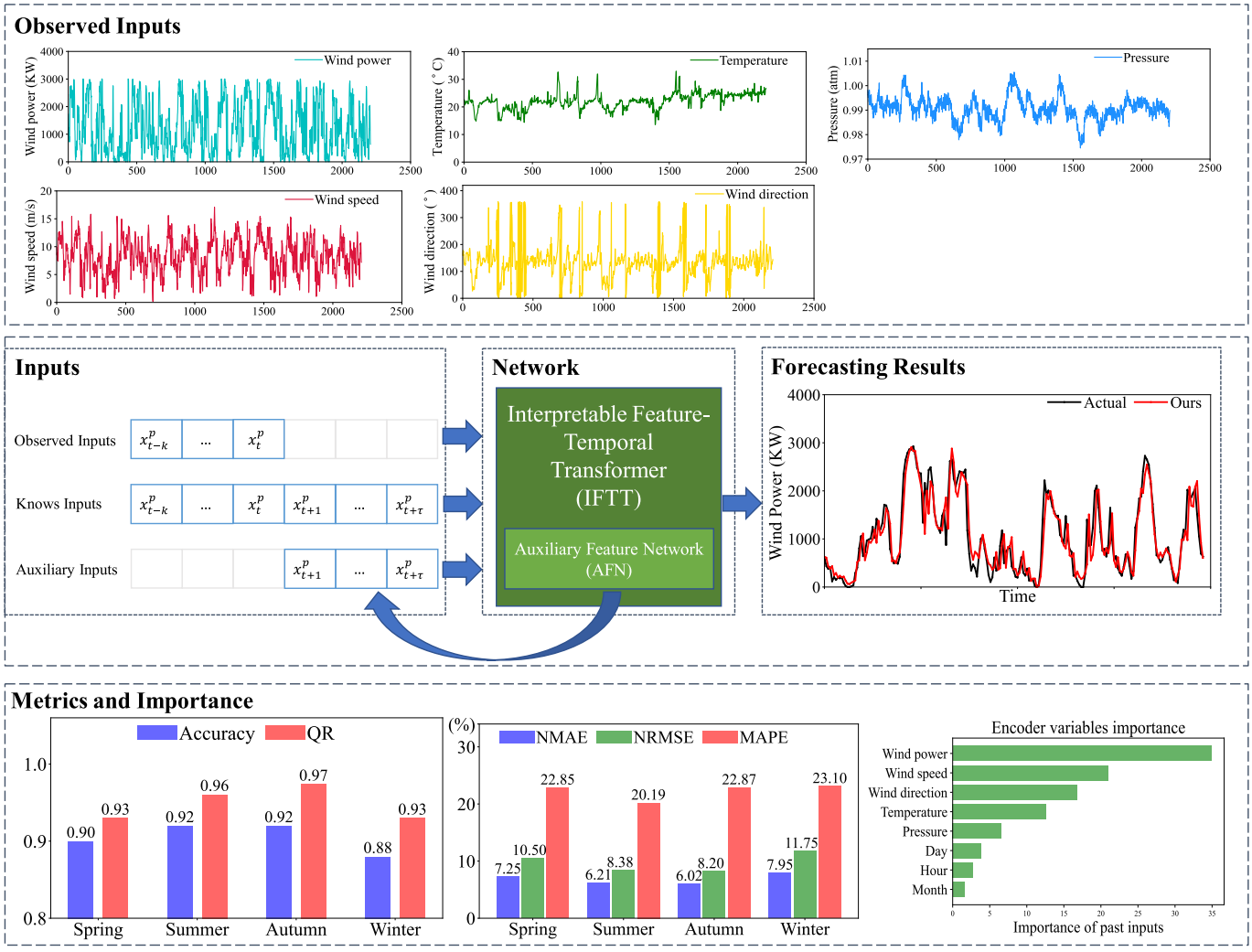


Fig. 3. Flowchart of interpretable WPF procedure.

are used in the following order: model construction, hyperparameter selection, and final model verification.

3.2. Multi-variable inputs

For wind power forecasting, the multi-variable inputs provide different information. In this study, we selected four types of input variables: Classic variables, Meteorological variables, Time-related variables and Pseudo-future variable. The classic variables comprise historical wind power and wind speed. Meteorological variables include Air temperature (Temperature), Wind direction, and Pressure. Time-related variables consist of Hour, Day of month, and Month. The pseudo-future variable refers to the pseudo-future wind speed obtained from AFN. It is noteworthy that classic variables are not included in the future inputs of our model's decoder. The types of input variables and the selection of past and future inputs are shown in Table 2.

To mitigate the impact of varying scales among the data, data standardization becomes imperative. In this study, we employed the Min-Max Normalization technique to individually normalize all input variables, as illustrated in Eq. (24):

$$x^* = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{24}$$

Regarding the input for time series modeling, it entails incorporating variables from previous time steps to get a lookback window. To ensure an adequate period for analysis, our study opted to utilize the historical variables from the preceding 24 h as inputs for the model, while considering the accessible variables from the subsequent 1 h as future inputs.

4. Experimental results and discussions

4.1. Hyperparameter settings and evaluation metrics

Optimizing hyperparameters is a critical factor affecting the performance of Wind Power Forecasting (WPF) models. Intelligent optimization algorithms can obtain optimal network hyperparameter settings, making the network structure setup more rational and efficient, and avoiding the issue of the model getting trapped in local optima due to manual tuning. The COA algorithm [46] has been proven to possess superior capabilities in balancing exploration ability and convergence speed. In this paper, the COA algorithm was chosen to tune the hyperparameters of the ITFF network, and the optimal parameters are shown in Table 3, including the number of units in DFTA and GRN, the number of heads of MHA, learning rate, dropout rate, and weight of loss in multi-task. In addition, the batch size was set to 100 and the optimizer was Adam. All experiments were performed on a workstation with a Dual Intel (R) Xeon (R) E5–2643 v4 CPU @3.4 GHz with 12 cores and 8 NVIDIA TESLA V100.

In this study, normalized mean absolute error (NMAE), normalized root mean square error (NRMSE), and mean absolute percentage error (MAPE) were employed as evaluation metrics to assess the performance

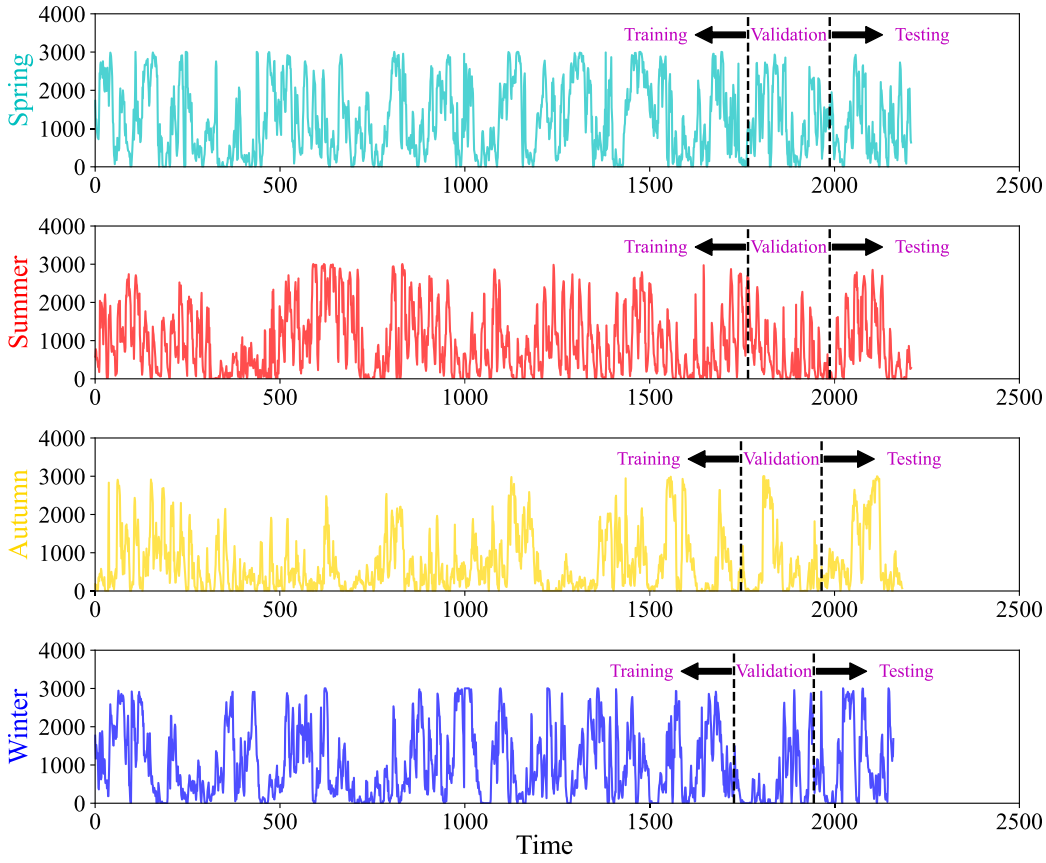


Fig. 4. Division of training, validation, and test sets for the four seasons.

Table 2
Inputs of the IFTT in the dataset.

Input Variable Type	Past Inputs	Future Inputs
Classic variables	Wind power	–
	Wind speed	–
Meteorological variables	Temperature	Temperature
	Wind direction	Wind direction
	Pressure	Pressure
Time-related variables	Hour	Hour
	Day of month	Day of month
	Month	Month
Pseudo-future variable	–	Wind speed (Pseudo)

Table 3
Parameters of the COA-IFTT in the four data sets.

	Hyperparameters	Spring	Summer	Autumn	Winter
IFTT	Learning rate	0.010	0.007	0.009	0.006
	Number of units in DFTA	67	92	82	116
	Number of hidden layers	1	3	1	3
	Number of attention heads	4	1	1	3
	Number of units in GRN	186	161	172	139
	Dropout rate	0.133	0.106	0.076	0.026
	weight	0.6	0.6	0.6	0.6

of the algorithms. Additionally, to ensure the safety of grid connection, the State Grid of China has specified requirements for the accuracy of wind power forecasting (accuracy) and the reporting qualified rate (QR) for wind farms. Wind farms will face severe financial penalties when the QR does not meet the requirements [33]. Therefore, in addition to the aforementioned classical evaluation metrics, this study included the assessment of forecasting accuracy and QR as evaluation metrics. The calculation of the evaluation metrics is given by Eqs. (25)–(30)

$$E_{NMAE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{p_i - y_i}{p_{cap}}\right| \times 100\% \tag{25}$$

$$E_{NRMSE} = \frac{1}{p_{cap}}\sqrt{\frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2} \times 100\% \tag{26}$$

$$E_{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\frac{|p_i - y_i|}{y_i} \times 100\% \tag{27}$$

$$Acc = \left(1 - \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{p_i - y_i}{p_{cap}}\right)}\right) \times 100\% \tag{28}$$

$$QR = \frac{1}{N}\sum_{i=1}^{N} C_i \tag{29}$$

$$C_i = \begin{cases} 1\ , \left(1 - \left|\frac{p_i - y_i}{p_{cap}}\right|\right) \times 100\% \geq r \\ 0\ , \left(1 - \left|\frac{p_i - y_i}{p_{cap}}\right|\right) \times 100\% < r \end{cases} \tag{30}$$

where p_i and y_i represent the forecasting wind power and the actual wind power, respectively, while p_{cap} represents the wind turbine capacity. The parameter r denotes the benchmark of QR, which increases as the requirements for wind power forecasting by the State Grid of China become more stringent. In this study, we selected a benchmark r of 80%.

4.2. Performance comparison and analysis of the models

4.2.1. Ablation experiment of input variables

In this study, an ablation experiment on multivariate inputs was conducted to analyze the effectiveness of each variable. Table 4 presents the results of the ablation experiment on various input variables across four seasons. From the table, it can be observed that adding meteorological variables as network inputs on the basis of classic variables (i.e., historical wind speed and historical power) can effectively improve the predictive performance of the model. On the four seasonal datasets, the NMAE, NRMSE, and MAPE decreased by an average of 0.84%, 1.26%, and 1.77%, respectively. And the ACC and QR metrics also improved by 1.34% and 2.94%, respectively. Furthermore, when time-related variables were added, the prediction results were further improved. For example, in the spring dataset, the NMAE, NRMSE, and MAPE decreased from 8.70%, 12.15%, and 29.49% to 8.26%, 11.92%, and 26.59%, respectively. Significantly, the addition of pseudo-future wind speed as an input greatly improved the model performance. All evaluation metrics exhibited notable improvement, with an average decrease of 0.91%, 1.28%, and 2.56% for NMAE, NRMSE, and MAPE, respectively, the ACC and QR metrics also improved on average by 1.28% and 1.63%. In the autumn season, the correlation coefficient between the forecasted wind power and the actual values reached a high value of 0.97.

From the above experimental results, it can be concluded that classic variables, meteorological variables, time-related variables, and pseudo-future wind speed are all beneficial for enhancing the performance of models during different seasons. The impact of pseudo-future wind speed and meteorological variables on model performance improvement was greater than that of time-related variables. When these variables are combined as inputs, the model achieves optimal performance. Additionally, it was observed that pseudo-future wind speed significantly improved the QR, suggesting that incorporating pseudo-future wind speed can effectively enhance the economic benefits of wind farms.

Fig. 5 displays the forecasting results of the ablation experiment on input variables. It can be observed that when only classic variables are used as network inputs, the model demonstrates good predictive performance during periods of stable wind speed but performs weakly during periods of abrupt wind speed changes. As meteorological variables, time-related variables, and pseudo-future wind speed are introduced, the predictive performance of the model gradually improves. When all these variables are included, the model's forecasting aligns closely with the actual wind power, reaching optimal performance.

4.2.2. Ablation experiment of network structure

In order to verify the effectiveness of the network structure, an experimental analysis was conducted, and the results are shown in Table 5, which illustrates the impact of each module on performance improvement. By adding the DFTA module to the basic TFT model, an effective fusion of different feature information and temporal information was achieved, leading to a significant enhancement in network performance. On the four seasonal datasets, the NMAE, NRMSE, and MAPE indicators decreased on average by 1.11%, 1.57%, and 3.44% respectively, while the Acc and QR indicators increased on average by 1.57% and 4.24% respectively. Furthermore, by incorporating the Auxiliary Forecasting Network (AFN) into the TFT-DFTA model to provide pseudo-future wind speed inputs to the Decoder module, the forecasting accuracy of wind power was further improved. For instance, in the spring season, the addition of the AFN module resulted in a decrease of the NMAE, NRMSE, and MAPE indicators from 8.26%, 11.92%, and 26.59% to 7.25%, 10.50%, and 22.85% respectively. The Accuracy and QR indicators increased from 88.08% and 91.67% to 89.5% and 92.96% respectively. The correlation coefficient between the forecasted results and actual wind power reached a high value of 0.92. The experimental analysis of the network structure demonstrated the effectiveness and necessity of the network structure proposed in this study. Fig. 6 displays the forecasting results of wind power for 192 consecutive time steps using different network structures, from which it can be observed that the addition of the DFTA and AFN modules can effectively enhance the predictive performance of the model.

4.2.3. Impact of intelligent optimization algorithms

Intelligent optimization algorithms can obtain optimal network hyperparameter settings, and avoid the issue of the model getting trapped in local optima due to manual tuning. Several classic intelligent optimization algorithms: Genetic Algorithm (GA) algorithm [45], Particle Swarm Optimization (PSO) algorithm [46], Adaptive Differential Evolution (ADE) algorithm [47] and COA algorithm [26] were been compared in this study. The experimental results are shown in Table 6. From Table 6, it can be observed that applying intelligent optimization algorithms to optimize the IFTT model's network structure has improved the wind power forecasting performance. Among the four seasons, both the ADE and COA algorithms generally outperformed the GA and PSO algorithms in terms of optimizing the model. When comparing ADE and COA algorithms, their performance was comparable in the spring season; In the other three seasons, the COA algorithm significantly outperformed the ADE algorithm in enhancing the wind power forecasting performance of the IFTT model. In this study, the COA algorithm was employed to intelligently optimize the network structure of the IFTT model, thus obtaining the optimal wind power forecasting performance.

4.2.4. Comparison with classic methods

In order to thoroughly validate the predictive performance of the COA-based IFTT model proposed in this study, it is compared with the

Table 4
Ablation experiment of input variables. Cv: Classic variables; Mv: Meteorological variables; Tv: Time-related variables; Pv: Pseudo-future variables.

Seasons	Input	NMAE/%	NRMSE/%	MAPE/%	Acc/%	QR/%	Pearson
Spring	Cv	9.59	13.33	33.12	86.67	86.46	0.89
	+Mv	8.70	12.15	29.49	87.85	88.02	0.89
	+Mv + Tv	8.26	11.92	26.59	88.08	91.67	0.90
	+Mv + Tv + Pv	**7.25**	**10.50**	**22.85**	**89.50**	**92.96**	**0.92**
Summer	Cv	8.54	12.12	26.87	87.88	89.89	0.92
	+Mv	7.79	10.93	26.84	89.07	92.13	0.93
	+Mv + Tv	7.04	9.78	23.74	90.22	93.67	0.94
	+Mv + Tv + Pv	**6.21**	**8.38**	**20.19**	**91.62**	**96.07**	**0.95**
Autumn	Cv	8.07	11.15	25.72	88.85	88.54	0.94
	+Mv	7.15	9.91	24.72	90.40	91.75	0.95
	+Mv + Tv	7.05	9.59	23.25	90.41	95.75	0.96
	+Mv + Tv + Pv	**6.02**	**8.20**	**22.87**	**91.80**	**97.40**	**0.97**
Winter	Cv	10.01	14.57	28.78	85.43	85.21	0.9
	+Mv	9.21	13.14	26.36	86.86	89.94	0.92
	+Mv + Tv	8.72	12.64	25.65	87.36	91.72	0.92
	+Mv + Tv + Pv	**7.95**	**11.75**	**23.10**	**88.25**	**92.90**	**0.93**

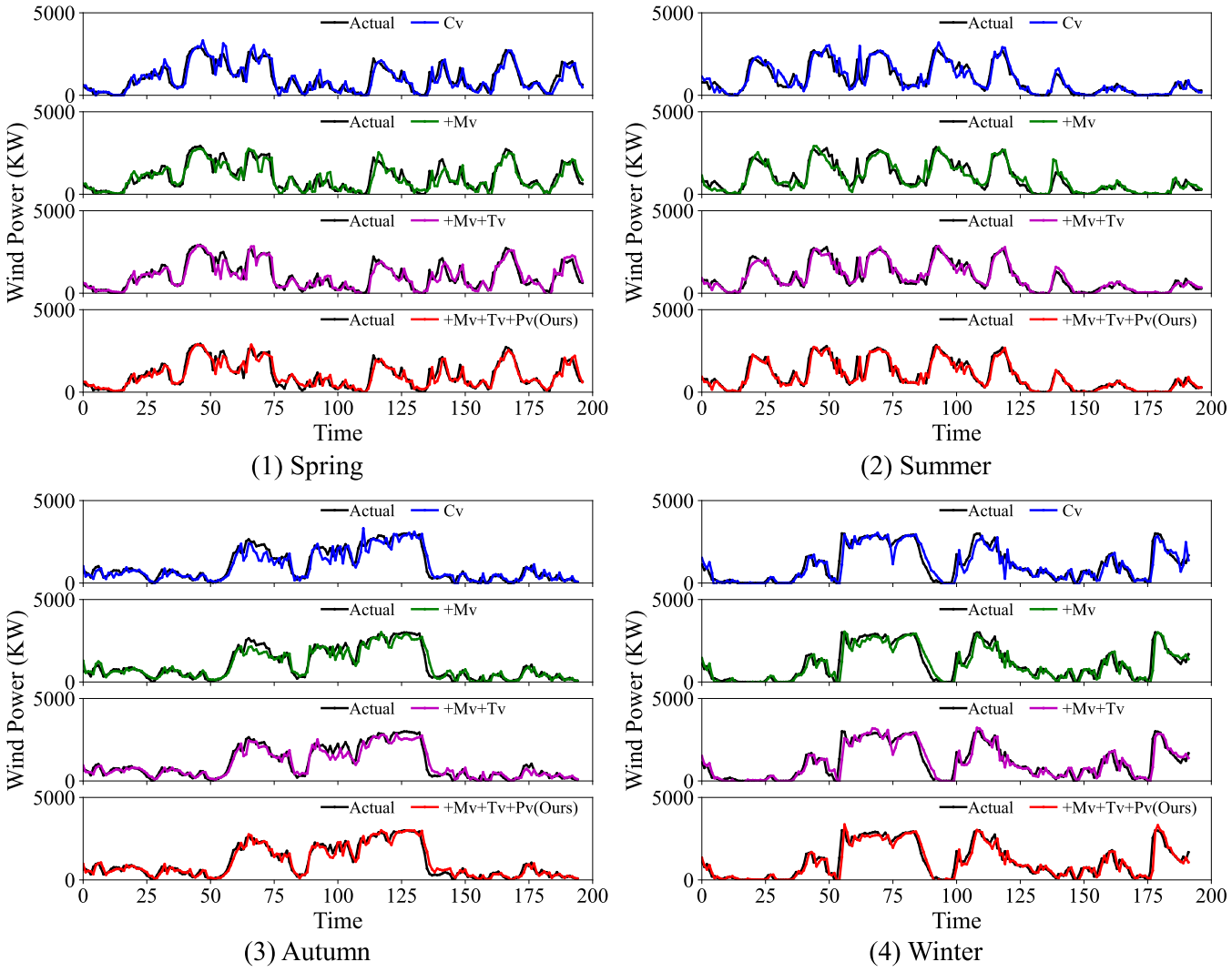


Fig. 5. Ablation experiment results of input variables.

Table 5
Ablation experiments of the model. DFTA represents the decoupled feature-temporal self-attention module, and AFN represents the auxiliary forecasting network.

Seasons	Method	NMAE/%	NRMSE/%	MAPE/%	Acc/%	QR/%	Pearson
Spring	Base Model(TFT)	9.31	13.24	29.12	86.76	85.42	0.87
	+DFTA	8.26	11.92	26.59	88.08	91.67	0.90
	Ours(+DFTA+AFN)	**7.25**	**10.50**	**22.85**	**89.50**	**92.96**	**0.92**
Summer	Base Model(TFT)	7.67	11.41	25.28	88.59	92.13	0.92
	TFT + DFTA	7.04	9.78	23.74	90.22	93.67	0.94
	Ours(+DFTA+AFN)	**6.21**	**8.38**	**20.19**	**91.62**	**96.07**	**0.95**
Autumn	Base Model(TFT)	7.66	10.57	29.60	89.43	94.27	0.94
	TFT + FI	7.05	9.59	23.25	90.41	95.75	0.96
	Ours(+DFTA+AFN)	**6.02**	**8.20**	**22.87**	**91.80**	**97.40**	**0.97**
Winter	Base Model(TFT)	10.87	14.97	28.98	85.03	84.02	0.90
	TFT + DFTA	8.72	12.64	25.65	87.36	91.72	0.92
	Ours(+DFTA+AFN)	**7.95**	**11.75**	**23.10**	**88.25**	**92.90**	**0.93**

DeepAR method [48], the Transformer method [39], the TFT method [36], the PatchTST method [49], the iTransformer method [50], the ModernTCN method [51]. The DeepAR method is a general model proposed by Salinas et al. for time series forecasting, while the Transformer method utilizes attention mechanisms to achieve parallel processing of time series forecasting tasks. Both of these methods have been proven to achieve good predictive performance in wind power forecasting [52,53]. The iTransformer method improves upon the Transformer method by enhancing its capability to capture multivariate correlations. Both the PatchTST method and the ModernTCN method are the latest state-of-the-art (SOTA) prediction models published just this year (ICLR 2024). However, these aforementioned methods lack time series interpretability. In contrast, the TFT model is an interpretable time series model that has quickly gained recognition among researchers in time series tasks. The COA-based IFTT model proposed in this study, inspired by the TFT algorithm, achieves multivariate interpretable forecasting for wind power for the first time.

The experimental comparison results of the various methods are

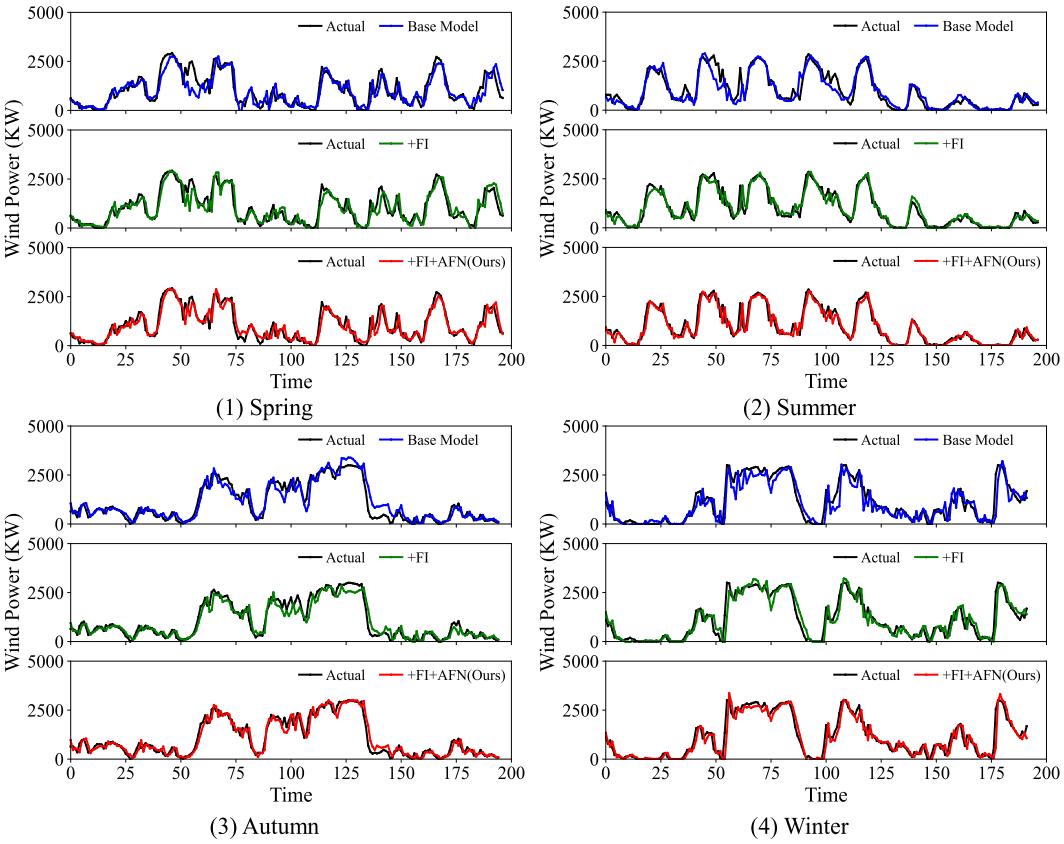


Fig. 6. Display of structural ablation experimental results.

Table 6
Improvement of IFTT model performance by various intelligent optimization algorithms.

Seasons	Method	NMAE/%	NRMSE/%	MAPE/%	Acc/%	QR/%	Pearson
Spring	GA-IFTT	7.46	10.80	24.73	89.20	89.58	0.9142
	PSO-IFTT	7.54	10.83	24.47	89.17	90.10	0.9160
	ADE-IFTT	7.35	**10.42**	24.62	**89.58**	91.15	0.9193
	COA-IFTT	**7.25**	10.50	**22.85**	89.50	**92.96**	**0.92**
Summer	GA-IFTT	6.54	9.49	22.06	90.51	95.51	0.94
	PSO-IFTT	6.53	9.35	21.97	90.65	95.51	0.94
	ADE-IFTT	6.22	9.20	20.93	90.80	95.69	0.94
	COA-IFTT	**6.21**	**8.38**	**20.19**	**91.62**	**96.07**	**0.95**
Autumn	GA-IFTT	6.11	8.45	25.74	91.55	95.83	0.96
	PSO-IFTT	6.07	8.35	23.60	91.65	96.35	0.96
	ADE-IFTT	6.06	8.29	23.98	91.71	95.96	0.9641
	COA-IFTT	**6.02**	**8.20**	**22.87**	**91.80**	**97.40**	**0.97**
Winter	GA-IFTT	8.20	12.32	23.75	87.68	91.79	0.92
	PSO-IFTT	8.18	12.14	23.43	87.86	92.31	0.92
	ADE-IFTT	8.11	12.18	**23.09**	87.82	92.32	0.92
	COA-IFTT	**7.95**	**11.75**	23.10	**88.25**	**92.90**	**0.93**

shown in Table 7. It can be observed that on the data of the four seasons (spring, summer, autumn, and winter), both DeepAR and Transformer achieve comparable predictive performance. In terms of the NMAE, NRMSE, and MAPE, the Transformer method slightly outperforms the DeepAR method. The iTransformer method, building upon the Transformer method, expands the local receptive field, thereby enhancing forecasting performance. Both the TFT and PatchTST methods achieve superior predictive outcomes across various metrics. Compared with the previous methods, the recently published ModernTCN model shows improved performance metrics across all four seasons. However, among these methods, only the TFT method possesses temporal interpretability capabilities. In pursuit of simultaneously achieving model interpretability and accurate wind power forecasting, this paper extends the TFT model framework, focusing on combining interpretability with precision. The COA-based IFTT model proposed in this study not only introduces the VAN network for selecting multivariate information but also proposes the DFTA module to effectively capture the importance of temporal information and integrate information. Additionally, it utilizes multi-task learning to provide additional supervision to the model, further improving the accuracy of wind power forecasting. The average NMAE, NRMSE, and MAPE on the data of the four seasons are 6.86%, 9.71%, and 22.25%, respectively. The average Acc, QR, and Pearson reach 90.29%, 94.83%, and 0.94, respectively. From the experimental results, it can be concluded that the proposed method outperforms the other classical methods. Furthermore, the comparison of the forecasting results in different seasons indicates that the performance of the proposed method is better in summer and autumn compared to spring and winter. This may be due to the potential fixed patterns in wind speed

Table 7
Performance comparisons with SOTA methods in different seasons.

Seasons	Method	NMAE/%	NRMSE/%	MAPE/%	Acc/%	QR/%	Pearson
Spring	DeepAR	16.30	21.21	51.87	78.79	65.62	0.68
	Transformer	16.56	20.25	51.39	79.75	70.31	0.66
	TFT	9.31	13.24	29.12	86.76	85.42	0.87
	PatchTST	8.50	12.01	29.04	87.99	87.50	0.89
	iTransformer	13.06	17.52	42.71	82.48	75.48	0.77
	ModernTCN	7.98	11.22	26.98	88.78	89.90	0.90
	Ours	**7.25**	**10.50**	**22.85**	**89.50**	**92.96**	**0.92**
Summer	DeepAR	15.04	17.38	53.61	82.62	71.91	0.84
	Transformer	11.07	14.50	42.14	85.50	86.52	0.87
	TFT	7.67	11.41	25.28	88.59	92.13	0.92
	PatchTST	7.65	11.52	22.89	88.48	90.86	0.92
	iTransformer	8.82	12.04	29.88	87.96	90.86	0.91
	ModernTCN	7.18	10.04	22.29	89.96	94.62	0.94
	Ours	**6.21**	**8.38**	**20.19**	**91.62**	**96.07**	**0.95**
Autumn	DeepAR	16.24	22.00	70.43	78.00	73.44	0.80
	Transformer	14.27	18.74	57.00	81.26	75.87	0.81
	TFT	7.66	10.57	29.60	89.43	94.27	0.94
	PatchTST	7.17	10.04	28.43	89.96	94.23	0.94
	iTransformer	10.76	16.70	30.37	83.30	85.58	0.84
	ModernTCN	6.49	8.51	22.91	91.49	96.15	0.96
	Ours	**6.02**	**8.20**	**22.87**	**91.80**	**97.40**	**0.97**
Winter	DeepAR	22.29	28.43	59.77	71.57	54.44	0.73
	Transformer	21.64	26.45	48.67	73.55	49.11	0.62
	TFT	10.87	14.97	28.98	85.03	84.02	0.90
	PatchTST	9.61	13.43	28.04	86.57	89.19	0.91
	iTransformer	12.25	17.48	35.11	82.52	77.84	0.85
	ModernTCN	8.51	12.86	25.41	87.14	90.27	0.91
	Ours	**7.95**	**11.75**	**23.10**	**88.25**	**92.90**	**0.93**

during summer and autumn, while wind speed exhibits greater variability in spring and winter, making forecasting more challenging.

Fig. 7 presents the wind power forecasting results of the proposed method and several classical methods in the four seasons datasets. It can be observed that there are significant differences between the forecasting results of DeepAR, Transformer, and iTransformer and the actual wind power. The TFT method performs well in predicting wind power during periods of smooth transition but still exhibits some errors during periods of abrupt changes. The PatchTST and ModernTCN methods achieve favorable forecasting results. The proposed method in this paper yields optimal forecasting results across all four seasons, on the other hand, is able to accurately predict wind power during both smooth and abrupt transition periods.

Comparisons on multiple wind farm datasets with different characteristics across different geographical locations can facilitate a more comprehensive evaluation of model performance. Therefore, in addition to the previously discussed Texas dataset, this paper conducts comparative analyses against various state-of-the-art (SOTA) methods on spring datasets from three wind farms located in Fujian Province, China. These wind farms [44] have different installation capacities: 48 MW (S1), 88 MW (S2), and 280 MW (S3). Similar to the Texas dataset, the Fujian datasets use spring data from the year of 2022, with a temporal resolution of 1 h, 24 points per day for each site. The experimental results are detailed in Table 8. Similar conclusions can be drawn from Table 8 as were from the Texas dataset. On the datasets from the three different wind farms, DeepAR, Transformer, and iTransformer exhibit comparable performances. The TFT and PatchTST methods marginally outperform these three methods, while the ModernTCN method yields results closest to those of our proposed method. Across multiple metrics, our method achieves the best performance, thereby more broadly validating its effectiveness and superiority.

4.3. Interpretable analysis of algorithms

The interpretability of deep neural networks has always been a challenging research topic. The interpretable wind power forecasting model proposed in this study, utilizing the VAN network and DFTA module, not only provides accurate forecasting results but also offers explanations for the importance of various factors, enhancing the credibility of the model. This is advantageous for making informed decisions regarding grid integration and resource allocation. Fig. 8 illustrates the interpretable analysis results of the IFTT model on the Texas dataset, which include three parts: importance ranking of past variables, importance ranking of future variables, and importance ranking of temporal lag information.

Among the past variables, historical wind power is the most important variable in the forecasting of wind power in all four seasons. Historical wind speed ranks second in importance in spring, summer, and winter, but its importance is slightly lower than meteorological elements in autumn. The overall importance ranking of variables in the four seasons indicates that meteorological elements are generally more important for wind power forecasting than time-related variables. However, there are exceptions, such as the Hour variable, which is relatively important in predicting wind power during summer, indicating the sensitivity of summer wind power to the time of day. In spring, autumn, and winter, temperature shows strong importance, possibly due to the lower and more variable temperatures during these seasons, which affect the efficiency of wind turbines and have a high correlation with wind power.

Unlike the past variables, among future variables, the importance contribution of meteorological variables and time-related variables to wind power forecasting changes with season. Pseudo-future wind speed demonstrates strong significance in wind power forecasting performance, while the importance of other variables varies depending on the season. The Hour variable remains important in summer, ranking second after pseudo-future wind speed, but its contribution is relatively small in the other three seasons. Temperature maintains a certain level of importance in all four seasons, and wind direction is most important in predicting wind power during spring, likely due to the variability of wind direction in the research area, which has a significant impact on wind turbine generation.

From the analysis of the importance of temporal lag information, the following observations can be made: (1) as the lag order increases, the importance generally decreases. Specifically, from lag time t-1 to t-8, the

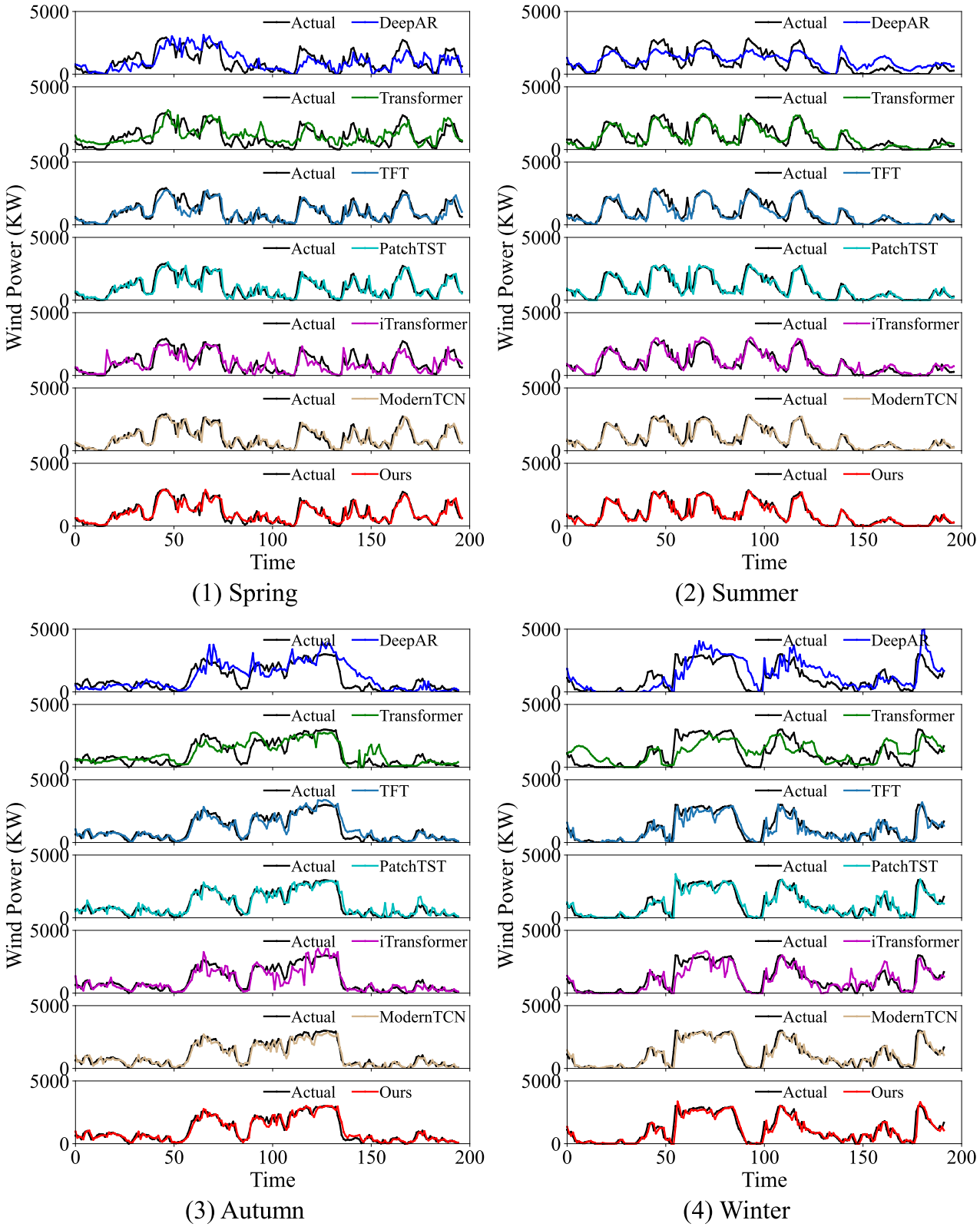


Fig. 7. Performance comparison with classical method in different seasons.

importance rapidly declines. (2) Different seasons require attention to different lag orders. For example, in winter, only the first 7 lag times need to be considered, as the contribution of information with larger lag orders to wind power forecasting is almost negligible. However, in spring, more lag time information needs to be taken into account.

5. Conclusion

To address the challenges of interpretability in wind power forecasting, this study presents interpretable short-term wind power forecasting with multi-variables and a feature-temporal transformer (IFTT). The model effectively learns the nonlinear mapping from input

Table 8
Performance comparisons in multiple wind farm datasets on spring.

Site	Method	NMAE/%	NRMSE/%	MAPE/%	Acc/%	QR/%	Pearson
	DeepAR	9.88	14.63	36.52	77.79	89.97	0.78
	Transformer	9.36	13.71	56.88	65.56	88.08	0.87
	TFT	8.40	13.56	34.07	86.44	86.75	0.89
S1	PatchTST	6.47	10.59	28.37	89.41	92.72	0.92
	iTransformer	9.46	12.54	37.38	87.46	89.40	0.90
	ModernTCN	6.15	9.64	27.69	90.36	94.04	0.94
	Ours	**5.52**	**8.90**	**25.21**	**91.10**	**95.36**	**0.95**
	DeepAR	8.01	12.29	38.74	87.71	88.33	0.89
	Transformer	7.99	12.69	54.46	61.01	89.44	0.86
	TFT	7.16	11.01	34.02	88.99	91.11	0.89
S2	PatchTST	6.48	9.83	30.20	90.17	93.89	0.92
	iTransformer	7.30	10.06	30.22	89.94	93.89	0.92
	ModernTCN	5.77	9.11	28.49	90.89	94.44	0.93
	Ours	**5.08**	**8.26**	**25.14**	**91.74**	**96.39**	**0.94**
	DeepAR	9.26	12.09	37.67	87.91	87.76	0.83
	Transformer	9.35	12.55	49.32	65.76	89.29	0.81
	TFT	8.23	10.73	34.02	89.27	90.82	0.88
S3	PatchTST	7.52	10.32	30.94	89.68	93.37	0.88
	iTransformer	8.48	10.92	35.16	89.08	92.35	0.86
	ModernTCN	6.96	9.68	27.77	90.32	9439	0.89
	Ours	**6.28**	**8.61**	**25.74**	**91.39**	**94.90**	**0.91**

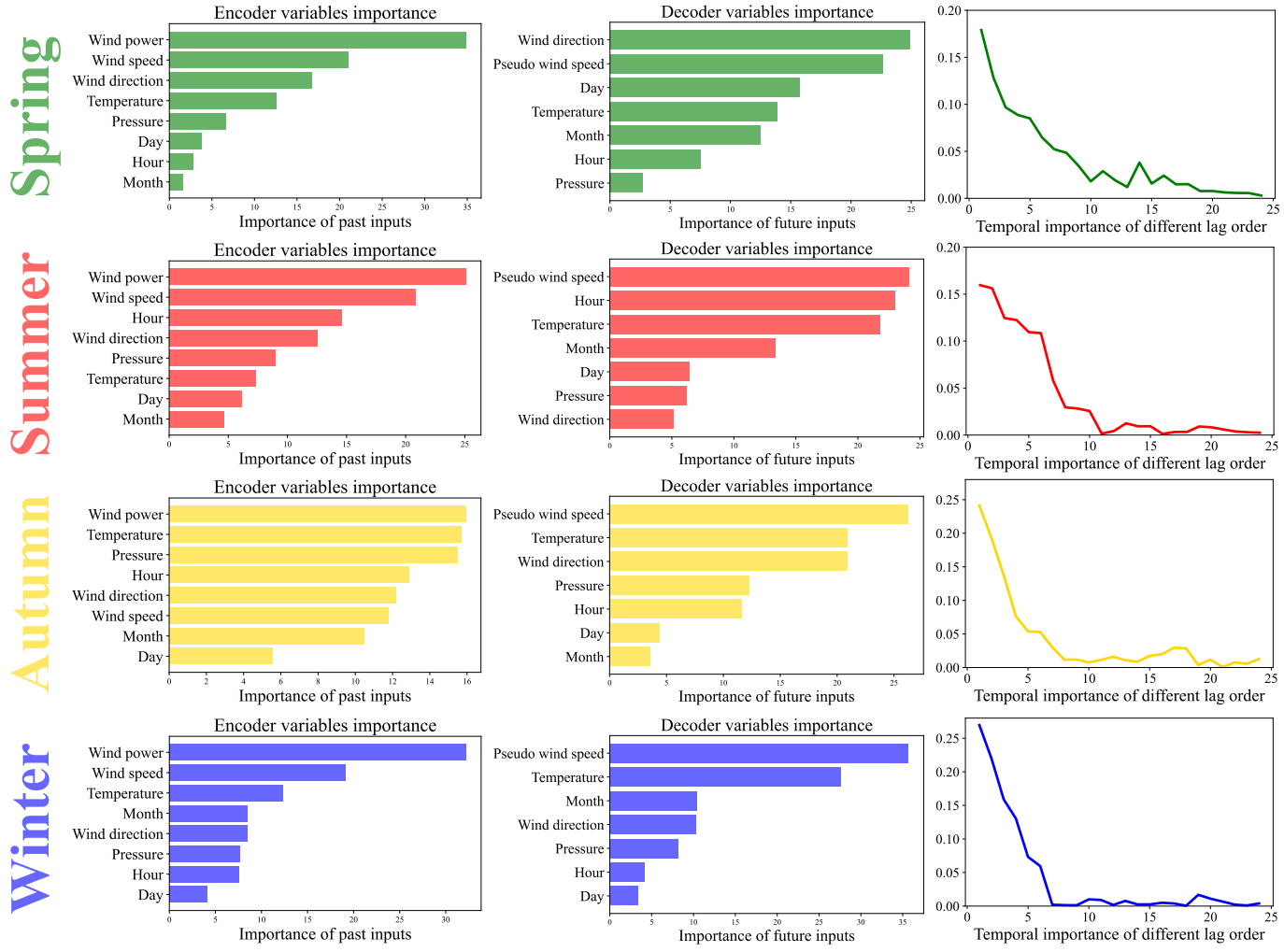


Fig. 8. Interpretable results of the IFTT model for four seasons on the Texas dataset.

information to wind power by combining the encoder-decoder structure and temporal information modeling. It comprehensively considers the historical information of multiple relevant variables, known future information, and pseudo-future information, and constructs the VAN network and DFTA module to perform information selection and importance analysis of temporal information and multivariate inputs.

The designed AFN not only provides pseudo-future wind speed to the decoder but also improves wind power forecasting performance through multi-task learning. Furthermore, the COA algorithm is utilized to optimize the network hyperparameters for further enhancing model performance. Ablation experiments demonstrate the effectiveness and necessity of the selected multivariate inputs and network structure design. Compared with the DeepAR, Transformer, TFT, PatchTST, iTransformer and ModernTCN methods on the multiple public datasets in different geographical locations, our method outperforms in multiple evaluation metrics.

To avoid the impact of information selection and fusion mechanisms in temporal modeling on interpretable temporal information, this study constructs the DFTA module to perform an effective importance analysis of temporal information before temporal modeling. The attention given by the forecasting model to different time lags indicates that the importance of temporal information decreases as the lag order increases, and different lag orders need to be considered for different seasons.

Analysis of the importance of multi-variable inputs shows that historical and future information of classical variables, meteorological variables and time-related variables all play an important role in wind power prediction. Among past variables, wind power and wind speed are the most important, and meteorological variables are generally more important than time-related variables; among future variables, pseudo-future wind speed plays a crucial role, while the importance of meteorological variables and time-related variables changes with the seasons.

The proposed interpretable wind power forecasting model not only achieves accurate forecasting results but also provides directions for optimizing wind power forecasting models and supporting decision-making through interpretable analysis of multivariate inputs and temporal information. Moving forward, we are dedicated to conducting empirical research that bridges the gap between model interpretability and grid management decision-making processes, ensuring our advancements translate seamlessly into practical applications. As part of our future work, we will strive to propose a unified interpretability framework. This framework will facilitate cross-model comparisons and enable more stringent evaluations of interpretability.

CRediT authorship contribution statement

Lei Liu: Writing – original draft, Visualization, Software, Methodology, Funding acquisition, Data curation, Conceptualization. **Xinyu Wang:** Validation, Supervision, Data curation, Conceptualization. **Xue Dong:** Validation, Investigation, Data curation. **Kang Chen:** Writing – review & editing, Validation, Investigation. **Qiuju Chen:** Validation, Supervision. **Bin Li:** Writing – review & editing, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

I have shared the link to my data in the manuscript.

Acknowledgements

This work was funded by the National Natural Science Foundation of China (U19B2044). Supported by the Fundamental Research Funds for the Central Universities (WK2100000045).

References

[1] Putz D, Gumhalter M, Auer H. A novel approach to multi-horizon wind power forecasting based on deep neural architecture. Renew Energy 2021;178:494–505.
[2] Ju Y, Sun G, Chen Q, Zhang M, Zhu H, Rehman MU. A model combining convolutional neural network and LightGBM algorithm for ultra-short-term wind power forecasting. Ieee Access 2019;7:28309–18.
[3] Wang G, Jia R, Liu J, Zhang H. A hybrid wind power forecasting approach based on Bayesian model averaging and ensemble learning. Renew Energy 2020;145:2426–34.
[4] Liu H, Chen C. Data processing strategies in wind energy forecasting models and applications: a comprehensive review. Appl Energy 2019;249:392–408.
[5] Soman SS, Zareipour H, Malik O, Mandal P. A review of wind power and wind speed forecasting methods with different time horizons. In: North American power symposium 2010. IEEE; 2010. p. 1–8.
[6] Wang Y, Zou R, Liu F, Zhang L, Liu Q. A review of wind speed and wind power forecasting with deep neural networks. Appl Energy 2021;304:117766.
[7] Jung J, Broadwater RP. Current status and future advances for wind speed and power forecasting. Renew Sust Energ Rev 2014;31:762–77.
[8] Dong Y, Zhang H, Wang C, Zhou X. A novel hybrid model based on Bernstein polynomial with mixture of Gaussians for wind power forecasting. Appl Energy 2021;286:116545.
[9] Du P, Wang J, Yang W, Niu T. A novel hybrid model for short-term wind power forecasting. Appl Soft Comput 2019;80:93–106.
[10] Hu J, Heng J, Wen J, Zhao W. Deterministic and probabilistic wind speed forecasting with de-noising-reconstruction strategy and quantile regression based algorithm. Renew Energy 2020;162:1208–26.
[11] Lei M, Shiyan L, Chuanwen J, Hongling L, Yan Z. A review on the forecasting of wind speed and generated power. Renew Sust Energ Rev 2009;13:915–20.
[12] Wang C, Zhang H, Ma P. Wind power forecasting based on singular spectrum analysis and a new hybrid Laguerre neural network. Appl Energy 2020;259:114139.
[13] Li L, Li Y, Zhou B, Wu Q, Shen X, Liu H, et al. An adaptive time-resolution method for ultra-short-term wind power prediction. Int J Electr Power Energy Syst 2020;118:105814.
[14] Zhao J, Guo Z-H, Su Z-Y, Zhao Z-Y, Xiao X, Liu F. An improved multi-step forecasting model based on WRF ensembles and creative fuzzy systems for wind speed. Appl Energy 2016;162:808–26.
[15] Cao Y, Liu Y, Zhang D, Wang W, Chen Z. In: Wind power ultra-short-term forecasting method combined with pattern-matching and ARMA-model. IEEE; 2013. p. 1–4.
[16] Tian S, Fu Y, Ling P, Wei S, Liu S, Li K. Wind power forecasting based on Arima-lgarch model. In: 2018 international conference on power system technology. POWERCON: IEEE; 2018. p. 1285–9.
[17] Xu Y, Jia L, Peng D, Yang W. A novel hammerstein wind power forecasting model. In: 2021 6th international conference on power and renewable energy (ICPRE). IEEE; 2021. p. 1019–24.
[18] Cao Y, Gui L. Multi-step wind power forecasting model using LSTM networks, similar time series and LightGBM. In: 2018 5th international conference on systems and informatics (ICSAI). IEEE; 2018. p. 192–7.
[19] Yu R, Gao J, Yu M, Lu W, Xu T, Zhao M, et al. LSTM-EFG for wind power forecasting based on sequential correlation features. Futur Gener Comput Syst 2019;93:33–42.
[20] Kisvari A, Lin Z, Liu X. Wind power forecasting–a data-driven method along with gated recurrent neural network. Renew Energy 2021;163:1895–909.
[21] Shi J, Guo J, Zheng S. Evaluation of hybrid forecasting approaches for wind speed and power generation time series. Renew Sust Energ Rev 2012;16:3471–80.
[22] Duan J, Wang P, Ma W, Fang S, Hou Z. A novel hybrid model based on nonlinear weighted combination for short-term wind power forecasting. Int J Electr Power Energy Syst 2022;134:107452.
[23] Wang Y, Xu H, Zou R, Zhang L, Zhang F. A deep asymmetric Laplace neural network for deterministic and probabilistic wind power forecasting. Renew Energy 2022;196:497–517.
[24] Shahid F, Zameer A, Muneeb M. A novel genetic LSTM model for wind power forecast. Energy 2021;223:120069.
[25] Abou Houran M, Bukhari SMS, Zafar MH, Mansoor M, Chen W. COA-CNN-LSTM: coati optimization algorithm-based hybrid deep learning model for PV/wind power forecasting in smart grid applications. Appl Energy 2023;349:121638.
[26] Dehghani M, Montazeri Z, Trojovská E, Trojovský P. Coati optimization algorithm: a new bio-inspired metaheuristic algorithm for solving optimization problems. Knowl-Based Syst 2023;259:110011.
[27] Niu D, Sun L, Yu M, Wang K. Point and interval forecasting of ultra-short-term wind power based on a data-driven method and hybrid deep learning model. Energy 2022;254:124384.
[28] Ding Y, Chen Z, Zhang H, Wang X, Guo Y. A short-term wind power prediction model based on CEEMD and WOA-KELM. Renew Energy 2022;189:188–98.
[29] Lu P, Ye L, Pei M, Zhao Y, Dai B, Li Z. Short-term wind power forecasting based on meteorological feature extraction and optimization strategy. Renew Energy 2022;184:642–61.
[30] Zhang W, Lin Z, Liu X. Short-term offshore wind power forecasting-a hybrid model based on discrete wavelet transform (DWT), seasonal autoregressive integrated moving average (SARIMA), and deep-learning-based long short-term memory (LSTM). Renew Energy 2022;185:611–28.
[31] Li L, Liu Y, Yang Y, Han S. Short-term wind speed forecasting based on CFD pre-calculated flow fields. Zhongguo Dianji Gongcheng Xuebao(proceedings of the

Chinese Society of Electrical Engineering): Chinese society for. Electr Eng 2013: 27–32.
[32] Wang Y, Liu Y, Li L, Infield D, Han S. Short-term wind power forecasting based on clustering pre-calculated CFD method. Energies 2018;11:854.
[33] Liu L, Liu J, Ye Y, Liu H, Chen K, Li D, et al. Ultra-short-term wind power forecasting based on deep Bayesian model with uncertainty. Renew Energy 2023; 205:598–607.
[34] Chen H, Wu H, Kan T, Zhang J, Li H. Low-carbon economic dispatch of integrated energy system containing electric hydrogen production based on VMD-GRU short-term wind power prediction. Int J Electr Power Energy Syst 2023;154:109420.
[35] Adedeji PA, Akinlabi S, Madushele N, Olatunji OO. Wind turbine power output very short-term forecast: a comparative study of data clustering techniques in a PSO-ANFIS model. J Clean Prod 2020;254:120135.
[36] Lim B, Arık SÖ, Loeff N, Pfister T. Temporal fusion transformers for interpretable multi-horizon time series forecasting. Int J Forecast 2021;37:1748–64.
[37] Howard J, Gugger S. Fastai: a layered API for deep learning. Information 2020;11: 108.
[38] Clevert D-A. Unterthiner T, Hochreiter S. Fast and accurate deep network learning by exponential linear units (elus). arXiv preprint arXiv:151107289. 2015.
[39] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Proces Syst 2017;30:5998–6008.
[40] Wen R, Torkkola K, Narayanaswamy B, Madeka D. A multi-horizon quantile recurrent forecaster. arXiv preprint arXiv:171111053. 2017.
[41] Yan X, Liu Y, Xu Y, Jia M. Multistep forecasting for diurnal wind speed based on hybrid deep learning model with improved singular spectrum decomposition. Energy Convers Manag 2020;225:113456.
[42] Qu J, Qian Z, Pei Y. Day-ahead hourly photovoltaic power forecasting using attention-based CNN-LSTM neural network embedded with multiple relevant and target variables prediction pattern. Energy 2021;232:120996.
[43] https://www.kaggle.com/datasets/pravdomirdobrev/texas-wind-turbine-dataset-simulated/data.
[44] https://www.dcic-china.com/competitions/10098/datasets.
[45] Mirjalili S, Mirjalili S. Genetic algorithm. Evolutionary Algorithms and Neural Networks: Theory and Applications. 2019. p. 43–55.
[46] Kennedy J, Eberhart R. Particle swarm optimization. In: Proceedings of ICNN'95-international conference on neural networks. IEEE; 1995. 1942–8.
[47] Neshat M, Nezhad MM, Abbasnejad E, Mirjalili S, Groppi D, Heydari A, et al. Wind turbine power output prediction using a new hybrid neuro-evolutionary method. Energy 2021;229:120617.
[48] Salinas D, Flunkert V, Gasthaus J, Januschowski T. DeepAR: probabilistic forecasting with autoregressive recurrent networks. Int J Forecast 2020;36: 1181–91.
[49] Nie Y, Nguyen NH, Sinthong P, Kalagnanam J. A time series is worth 64 words: long-term forecasting with transformers. The Eleventh International Conference on Learning Representations. 2023.
[50] Liu Y, Hu T, Zhang H, Wu H, Wang S, Ma L, et al. iTransformer: inverted transformers are effective for time series forecasting. The Twelfth International Conference on Learning Representations. 2024.
[51] Luo D, Wang X. Moderntcn: A modern pure convolution structure for general time series analysis. The Twelfth International Conference on Learning Representations. 2024.
[52] Arora P, Jalali SMJ, Ahmadian S, Panigrahi B, Suganthan P, Khosravi A. Probabilistic wind power forecasting using optimized deep auto-regressive recurrent neural networks. IEEE Trans Industr Inform 2022;19:2814–25.
[53] Wu H, Meng K, Fan D, Zhang Z, Liu Q. Multistep short-term wind speed forecasting using transformer. Energy 2022;261:125231.