A GVSAO-Optimized CNN-BiGRU-Attention Network for Short-Term Wind Power Forecasting

Qing Lv[1,2], Xinyu Zhai[1], Rundong Zhou[1], Qiang Li[1,2], and Xiancong Wu[1,2,*]

[1]*College of Engineering, Hebei Normal University, Shijiazhuang 050024,China*
(e-mail: xinyu@stu.hebtu.edu.cn, zhourundong@stu.hebtu.edu.cn)
[2]*Hebei Provincial Key Laboratory of Information Fusion and Intelligent Control, Shijiazhuang, Hebei 050024,China*
(e-mail: liqiang01@hebtu.edu.cn, lvqing@hebtu.edu.cn, congeve@hebtu.edu.cn)

Abstract: This paper proposes a wind energy forecasting model based on the Good-point-set and Vibration Snow Ablation Optimizer (GVSAO), Convolutional Neural Network (CNN), Bidirectional Gated Recurrent Unit (BiGRU), and attention mechanism. In the proposed framework, CNN is used to extract local feature representations from multivariate wind energy time series, BiGRU is employed to capture bidirectional temporal dependencies, and the attention mechanism is introduced to emphasize critical temporal information. To reduce the influence of manual hyperparameter tuning, GVSAO is developed by incorporating a good-point-set initialization strategy and a periodic oscillatory escape mechanism into the Snow Ablation Optimizer. The improved optimizer is used to adaptively search for key hyperparameters, including the learning rate, convolution kernel size, and number of hidden units. Experiments are conducted on two real-world wind energy datasets, including day-ahead wind power forecasting and short-term wind speed forecasting tasks. The results show that the proposed GVSAO-CNN-BiGRU-Attention model achieves superior prediction performance compared with several baseline models. On Dataset I, the proposed model obtains an RMSE of 0.73206, a MAPE of 0.11725, and an R^2 of 0.92550. On Dataset II, it achieves an RMSE of 0.4228, a MAPE of 0.0868, and an R^2 of 0.8010. In addition, ablation experiments, repeated experiments, standard deviation analysis, and statistical significance tests further verify the effectiveness and robustness of the proposed model.

Keywords: Power prediction; Deep learning; Attention mechanism; Convolutional neural network; Snow ablation optimizer.

1. INTRODUCTION

Recently, the energy crisis has become increasingly severe, while global renewable energy generation has grown rapidly and steadily in the market. Wind power has become one of the major power generation sectors. However, due to its inherent volatility, randomness, and intermittency, wind power presents major challenges for the operation and dispatch of power systems. How to fully extract effective features under complex meteorological conditions and improve the model's stability in different wind power scenarios remains a key issue that current research needs to focus on. Consequently, research on ultra-short-term forecasting of wind power output is particularly crucial (Yang et al., 2024a). Therefore, forecasting models for wind power not only need to have high prediction accuracy but also need to be able to adapt to the strong volatility, non-linearity, and multi-factor coupling characteristics in wind power data.

Currently, scholars both domestically and internationally have conducted extensive research on ultra-short-term wind power forecasting. Current methods for predicting wind power can be broadly classified into physical models, statistical models, traditional machine learning models, and deep learning models. Physical models are usually based on the characteristics of wind farms, meteorological conditions, and the physical mechanisms involved in the formation of wind power to establish predictive relationships (Yang et al., 2024b). These methods have excellent interpretability but often rely on high-quality weather forecasts and detailed physical parameters of the wind farm, resulting in high computational complexity. Statistical models can provide relatively stable prediction results under certain historical data conditions, but their ability to depict highly nonlinear fluctuations and complex changing relationships is limited. Traditional machine learning methods, such as artificial neural networks (ANN) (Karaman, 2023), gray prediction (Ren et al., 2024a), support vector machines (SVM) (Kim and Ko, 2025), and random forests, can handle nonlinear relationships to a certain extent. However, feature engineering, sample size, and hyperparameter settings often influence the predictive performance of these methods.The aforementioned limitations have driven the further application of deep learning models in wind power prediction.

Recently, advanced deep learning networks have been incorporated into models for forecasting wind power to enhance the ability to extract features from historical time series data, thereby further improving the predictive accuracy of traditional models.

As one of the deep learning networks, recurrent neural networks (RNNs) are frequently applied to simple time series forecasting due to their ability to pass historical information through hidden states via recursion. However, RNNs also suffer from poor learning of long-term dependencies, which limits their application scope. Therefore, a specialized form of RNN called the Long Short-Term Memory (LSTM) neural network was proposed, which can partially address the long-term dependency issue in RNNs. Yang et al. (Yang et al., 2023a) and Zhang et al. (Zhang and Wang, 2024) respectively employed the Improved Whale Optimization Algorithm (IWOA) and Convolutional Block Attention Module (CBAM) algorithms to optimize LSTM for wind power forecasting. Experimental results validated that the LSTM deep learning network demonstrated superior performance compared to traditional models. Nevertheless, LSTM-based models still mainly emphasize temporal dependency modeling, and their ability to extract local feature patterns from multivariate data on wind power remains limited.

Convolutional neural networks (CNNs) are also a type of deep learning network. By introducing convolutional operations in the time series dimension, CNNs perform local feature modeling on historical observation data to further extract key information implicit in the time series (Zavala-Mondragón et al., 2023; Yang et al., 2025a). Due to their multi-layer convolutional structure, CNNs can effectively represent the complex nonlinear relationships within time series. Moreover, the local connections and parameter sharing mechanism employed by CNNs effectively enhance feature extraction efficiency and generalization capabilities while reducing model complexity. However, CNNs are not specifically designed to model long-term sequential dependencies, which limits their ability to represent temporal evolution in wind power series when used alone.

Nevertheless, a single deep learning architecture is typically designed to enhance only a specific aspect of representation learning. For instance, convolutional neural networks (CNNs) excel at extracting local patterns, whereas recurrent neural networks (RNNs) are more effective in capturing temporal dependencies. Without an appropriate integration mechanism, it remains challenging for a standalone model to simultaneously characterize local features, long-term dependencies, and critical temporal information. Consequently, the effective combination of complementary deep learning architectures has emerged as a promising strategy for improving wind power forecasting performance. Along this line, existing studies generally pursue two avenues to further enhance prediction accuracy: reducing the complexity of raw time series through signal decomposition techniques and developing hybrid deep learning frameworks that leverage the strengths of different network architectures.

A common approach is to introduce signal processing methods to reduce the complexity of wind power time series, such as Fourier Transform (FT/FFT) (Wang et al., 2024a; Khouili et al., 2025), Short-Time Fourier Transform (STFT) (Xiang et al., 2022; Zhang et al., 2025), Wavelet Transform (WT) (Liu et al., 2024a), Empirical Mode Decomposition (EMD) (Ghanbari and Avar, 2024), Ensemble Empirical Mode Decomposition or Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (EEMD/CEEMDAN) (Yang and Zhao, 2025), Variational Mode Decomposition (VMD) (Lei et al., 2024), and Hilbert-Huang Transform (HHT) (Gao et al., 2025). These methods can decompose the original wind power data into subsequences with different temporal or frequency characteristics, thereby reducing the uncertainty of the original sequence. However, decomposition-based methods usually require additional decomposition and reconstruction steps, making the prediction process more complex. The quality of the decomposition also closely influences the final prediction results. Therefore, developing end-to-end hybrid deep learning models that can directly learn effective representations from raw data remains an important direction in research on predicting wind power.

Another approach builds upon the aforementioned single deep learning model by constructing multiple hybrid models to leverage their complementary strengths. This allows for a more comprehensive utilization of each method's characteristics, thereby further enhancing prediction accuracy and computational efficiency. (Sherry and Rajalakshmi, 2023) proposed a hybrid deep learning forecasting model: the CNN Encoder-Decoder LSTM (CNN-ED-LSTM) model, which employs a one-dimensional convolutional neural network (CNN) as the encoder and an LSTM as the decoder. Experimental results demonstrated that the hybrid model outperformed the standalone LSTM model in predictive performance. Zeni Z et al. (Zeni et al., 2023) constructed a hybrid deep learning-based short-term wind power forecasting model. By integrating VMD, CNN, and GRU for short-term wind power prediction, comparative analysis with multiple traditional single models confirmed the hybrid model's superior performance and practical value in real-world applications.

Meanwhile, as the layer structures of deep learning networks continue to expand, the introduction of optimization algorithms has further facilitated the development of complex hybrid models. Existing research indicates that the synergistic effects of increasing the number of network layers (Wang et al., 2024b), enhancing CNN depth (Chen and Tsou, 2022), and incorporating optimization algorithms (Du et al., 2022; Zhao et al., 2025) contribute to improving model prediction performance to varying degrees. Despite the progress achieved by existing studies, several challenges remain. First, many hybrid forecasting models primarily combine different network architectures, but they often overlook the adaptive optimization of critical hyperparameters. Second, model performance is often evaluated on a single dataset, with limited validation across diverse wind farm scenarios. As a result, the robustness and generalization capability of these models remain insufficiently explored.

The CNN-BiGRU-Attention hybrid structure has gradually been applied to wind power prediction tasks in recent years because this structure can simultaneously combine local feature extraction, bidirectional temporal dependency modeling, and information filtering capabilities based on the attention mechanism. Liu H et al. (Liu et al., 2024b) proposed a short-term wind power prediction model that combines the BKA algorithm and the CNN-BiGRU-Attention hybrid structure. The results show that the optimized hybrid structure has better prediction performance compared to the traditional CNN-BiGRU-Attention model. Karamichailidou D et al. (Karamichailidou et al., 2021) constructed a model for predicting wind power based on radial basis function neural networks and improved the prediction performance by enhancing the generalization ability of the neural network. Han Xiao et al. (Xiao et al., 2023) proposed a GRU neural network model optimized by an improved ant colony algorithm, and the research results indicate that the ACA-GRU model has high prediction accuracy. D Ren et al. (Ren et al., 2024b) proposed the IVMD-CNN-GRU-Attention prediction method, which integrates sample entropy and significantly improves the accuracy of wind power prediction. Wang et al. (Wang et al., 2024c) applied the BiGRU-Attention model to low-frequency wind power prediction and achieved good results. Liu S et al. (Liu et al., 2024c) proposed a hybrid prediction model based on LSTM, providing a reference for the application of deep learning hybrid models in wind power prediction.

Although the aforementioned studies have improved performance of wind power prediction from different perspectives, there are still some shortcomings. First, the performance of hybrid deep learning models is largely influenced by key hyperparameters, such as learning rate, convolution kernel size, and the number of hidden units. Manual tuning of these parameters is inefficient and often leads to suboptimal parameter combinations. Secondly, although existing intelligent optimization algorithms can alleviate the parameter selection problem, there may still be issues such as insufficient initial population diversity and premature convergence during the search process. Third, many models for predicting wind power are still primarily validated based on a single dataset, with insufficient reports on repeated experiments, standard deviation analysis, and statistical significance tests. These issues undermine the reliability and generalizability of the model's predictive results.

To address the issues of low efficiency in manual parameter tuning and premature convergence during the optimization process, this paper combines the improved snow melting optimization algorithm GVSAO with the CNN-BiGRU-Attention prediction framework. Specifically, this paper introduces a good point set initialization strategy to improve the diversity and distribution quality of the initial population; at the same time, a periodic oscillation escape mechanism is introduced to reduce the risk of the algorithm getting trapped in local optima and to enhance global search capabilities. On this basis, GVSAO is used to adaptively optimize the key hyperparameters of the CNN-BiGRU-Attention network, including the learning rate, convolution kernel size, and number of hidden units. By combining intelligent optimization algorithms with deep feature learning, the proposed GVSAO-CNN-BiGRU-Attention model aims to improve the accuracy, robustness, and training stability of wind power prediction.

This paper conducts research in the following areas:

(1) A novel GVSAO optimization algorithm is developed by integrating a good-point-set initialization strategy with a periodic oscillatory escape mechanism. The proposed enhancements improve population diversity during the initialization stage and effectively mitigate premature convergence, thereby strengthening the global optimization capability of the algorithm.

(2) A GVSAO-CNN-BiGRU-Attention forecasting framework is developed for wind power prediction. In this framework, CNN is used to identify local features, BiGRU is used to identify bidirectional temporal dependencies, and the attention mechanism is used to highlight important temporal information. Meanwhile, GVSAO is adopted to optimize key hyperparameters adaptively, including the learning rate, convolution kernel size, and number of hidden units, thereby improving both forecasting accuracy and model stability.

(3) This paper conducts a comprehensive evaluation of the proposed model using two real wind energy datasets. In addition to using the Sotavento wind farm data for day-ahead wind power forecasting, the Sao Paulo public wind turbine dataset is also introduced for short-term wind speed forecasting. In addition, this paper further conducts repeated experiments, standard deviation analysis, statistical significance tests, and ablation experiments to evaluate the robustness, reliability, and specific contributions of each component module of the proposed model.

The rest of this paper is organized as follows: Section 2 introduces the basic neural network modules and optimization algorithms used in this study; Section 3 describes the construction process of the GVSAO-CNN-BiGRU-Attention prediction model; Section 4 presents the data description, experimental setup, prediction results, ablation experiments, and discussion; Section 5 concludes the paper.

2. METHODOLOGY

2.1 CNN

Convolutional networks, also known as convolutional neural networks (CNNs), are a specific type of deep learning model that works well with data that has a structured layout, like grids. CNNs have been extensively applied in various domains, including time series analysis, computer vision, and natural language processing, due to their powerful feature extraction capabilities. Based on the shape of the input data, CNNs can be divided into one-dimensional (1D), two-dimensional (2D), and three-dimensional (3D) convolutional networks, with each type designed for different kinds of data and uses.

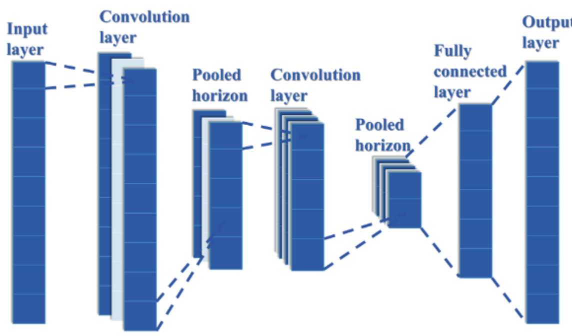


Fig. 1. Schematic of a Convolutional Neural Network.

The architecture of a convolutional neural network (CNN) primarily comprises an input layer and a series of convolutional layers, with its core components including convolutional layers, pooling layers of data (also referred to as subsampling layers), and fully connected layers. The CNN employed in this study adopts a one-dimensional (1D) convolutional structure, which is particularly suited for extracting local patterns and features from time series data. In 1D convolution, convolutional kernels slide over the input sequence, performing element-wise multiplication followed by summation to generate the convolution output. The fundamental computational formula is as follows:

$$y_j = \sum_{i=1}^{k} w_i \cdot x_{j+i-1} + b_j \tag{1}$$

In this formulation, x denotes the input sequence of length n, w represents the convolutional kernel (also referred to as a filter) of length k, b is the bias term introduced to enhance the representational capacity of the model, and y denotes the output sequence of length m.

However, due to the inherent limitation of CNN in capturing long-term temporal dependencies, it is necessary to integrate Bidirectional Gated Recurrent Units (BiGRU), which possess strong capabilities in modeling sequential data, to more effectively handle forecasting tasks for time series.

2.2 BiGRU

The bidirectional gated recurrent unit (BiGRU) is a type of recurrent neural network (RNN) that uses information from both directions in a sequence, allowing it to effectively handle time-related patterns, as shown in Fig.2 .

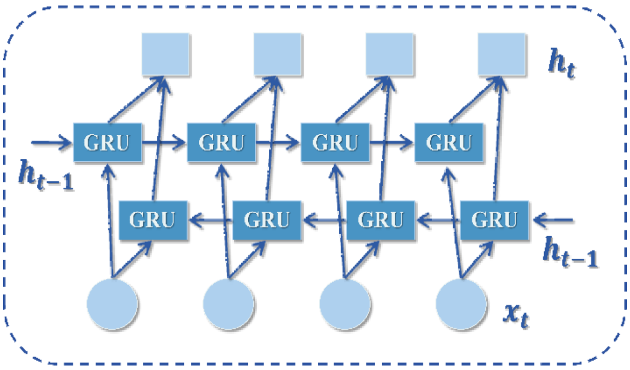


Fig. 2. Structure of BiGRU.

By simultaneously capturing both forward and backward dependencies within a sequence, BiGRU is particularly well-suited for tasks that require comprehensive contextual understanding. This characteristic makes it highly appropriate for the modeling framework proposed in this study.

The core computational equations of the BiGRU are as follows:

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t] + b_r) \tag{2}$$
$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t] + b_z) \tag{3}$$
$$\tilde{h} = \tanh(W_h \cdot [r_t \odot h_{t-1}, x_t] + b_h) \tag{4}$$
$$h_t = z_t \odot h_{t-1} + (1 - z_t \odot \tilde{h}_t) \tag{5}$$

where, h_t denotes the hidden state at the current time step, while x_t represents the input at the corresponding time step, σ denotes the sigmoid activation function, $\odot$ represents the element-wise multiplication.

However, BiGRU lacks the ability to automatically differentiate the importance of individual time steps, treating all temporal inputs uniformly. These limitations may lead to a dilution of critical information and hinder the model's ability to accurately predict power fluctuations. To address this limitation, an attention mechanism is incorporated into the model, enabling it to focus selectively on the most informative time steps and thereby enhancing prediction accuracy.

2.3 Attention

BiGRU can effectively capture bidirectional temporal dependencies, but it treats the encoded hidden representations in a fairly uniform way when making predictions. However, for wind power forecasting, temporal features from different time steps often exhibit varying degrees of relevance to the target variable. To better exploit informative temporal patterns, an attention mechanism is incorporated following the BiGRU layer. By assigning adaptive weights to hidden states according to their contribution to the forecasting objective, the attention mechanism enables the model to focus on the most relevant temporal information while suppressing less informative features, thereby improving the quality of feature representation and forecasting accuracy.

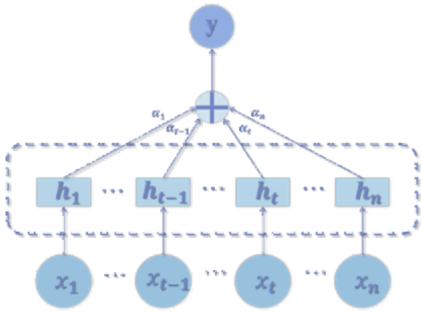


Fig. 3. Structure of the Attention Mechanism.

Given the hidden representations produced by the BiGRU layer, let $H = [h_1, h_2, \ldots, h_t, \ldots, h_T]$ denote the corresponding hidden state sequence, where h_t is the hidden state at time step t and T denotes the sequence length. An attention score is then assigned to each hidden state according to :

$$e_t = v_a^T \tanh(W_a h_t + b_a) \tag{6}$$

Where W_a , b_a and v_a are trainable parameters, and e_t represents the importance score associated with the t -th time step.

The attention weights are then computed using the Softmax function:

$$\alpha_t = \frac{\exp(e_t)}{\sum_{j=1}^{T} \exp(e_j)} \tag{7}$$

where α_t represents the contribution of the t -th hidden state to the final prediction. The context vector is then computed as a weighted sum of the hidden states:

$$C = \sum_{t=1}^{T} \alpha_t h_t \tag{8}$$

The context vector C aggregates the most informative temporal features across the entire sequence and is subsequently fed into the fully connected layer for forecasting. Through adaptive weighting of hidden states, the attention mechanism enables the model to focus on critical temporal information while reducing the influence of redundant or less relevant features.

2.4 SAO

(Lingyun Deng et al., 2023) came up with the Snow Ablation Optimization (SAO) algorithm, which is a population-based metaheuristic algorithm for optimizing data and engineering design. The fundamental principle lies in: achieving a balance between global search and local convergence by simulating the stochastic movement and gradual dissipation processes of snow particles. During each iterative generation, snow particles continuously update their positions within the solution space and seek the optimal solution by controlling contraction and diffusion mechanisms.

The SAO algorithm can generally be divided into three stages:

(1) Initialization Stage

In the SAO algorithm, the iterative process starts with a population randomly sampled from a normal (Gaussian) distribution. The population is represented as an $N \times Dim$ matrix, as shown in formula (9), where each row corresponds to an individual and each column corresponds to a solution dimension.

$$Z = L + \theta \times (U - L) \tag{9}$$

where Z is a matrix of N dimensions and Dim dimensions, N denotes the size of the population, Dim denotes the dimension of the solution space, and L and U denote the lower and upper bounds of the solution space, respectively, θ represents a number randomly generated in the interval [0,1]

(2) Exploratory Stage

When snow or liquid water transformed from snow is converted into steam, due to irregular motion, the exploratory agent displays highly dispersed characteristics. The formula for calculating the position during the exploratory phase is as follows:

$$\begin{aligned} Z_i(t+1) = {} & Elite(t) + BM_i(t) \otimes (\theta_1 \times (G(t) - Z_i(t)) \\ & + (1-\theta_1) \times (\bar{Z}(t) - Z_i(t))) \end{aligned} \tag{10}$$

$$\bar{Z}(t) = \frac{1}{N}\sum_{i=1}^{N} Z_i(t) \tag{11}$$

$$Elite(t) \in [G(t), Z_{second}(t), Z_{third}(t), Z_c(t)] \tag{12}$$

$$Z_c(t) = \frac{1}{N_1}\sum_{i=1}^{N_1} Z_i(t) \tag{13}$$

Where, $Z_i(t)$ represents the i-th individual in the n-th iteration; $BM_i(t)$ denotes a vector containing random numbers based on a Gaussian distribution, representing Brownian motion; $\otimes$ indicates element-wise multiplication; θ_1 is a randomly selected number from the interval [0,1]; $G(t)$ represents the current global best solution; $Elite(t)$ is an individual randomly selected from a set of *Elite* individuals in the population; $\bar{Z}(t)$ denotes the centroid position of the entire population; $Z_{second}(t)$ and $Z_{third}(t)$ represent the *second-best* and *third-best* individuals in the population, respectively; $Z_c(t)$ is the centroid position of the top 50% of individuals based on fitness; N_1 is the number of leaders; and $Z_i(t)$ is the i-th best leader.

(3) Development Stage

This stage simulates the process by which snow transforms into liquid water through melting, encouraging search agents to develop high-quality solutions around the current best solutions. The expression for the snowmelt rate is as follows:

$$M = (0.35 + 0.25 \times \frac{e^{\frac{1}{t_{\max}}} - 1}{e - 1}) \times T(t), T(t) = e^{\frac{-t}{t_{\max}}} \tag{14}$$

where, t_{max} represents the termination criterion.

The position update equation for the development phase is:

$$\begin{aligned} Z_i(t+1) = {} & M \times G(t) + BM_i(t) \otimes (\theta_2 \times (G(t) - Z_i(t)) \\ & + (1-\theta_2) \times (\bar{Z}(t) - Z_i(t))) \end{aligned} \tag{15}$$

Where, M represents the snowmelt rate, and θ_2 denotes a randomly selected number from the interval [0,1].

During application, SAO offers advantages such as straight forward implementation, minimal parameter requirements, and robust global search capabilities, demonstrating strong convergence performance in various nonlinear optimization problems.

However, the original SAO still suffers from limitations such as slow convergence speed, susceptibility to local optima, and insufficient initial diversity when handling high-dimensional complex problems. Based on these limitations, this paper proposes an improved algorithm, GVSAO, to enhance the stability and efficiency of the optimization process.

2.5 GVSAO

In the original SAO algorithm, population individuals progressively cluster around promising solutions during the optimization process, leading to a gradual loss of population diversity. Consequently, the search may become trapped in locally optimal regions, resulting in premature convergence and reduced exploration capability. To overcome these limitations, a Good Point and Vibration Snow Ablation Optimizer (GVSAO) is proposed by integrating a good-point-set initialization strategy and a periodic oscillatory escape mechanism into the SAO framework. Specifically, the good-point-set initialization strategy is employed to improve the uniformity of the initial population and enhance global exploration, while the periodic oscillatory escape mechanism is designed to maintain search diversity and facilitate exploration beyond local optima in the later optimization stages (Fang et al., 2024; Pehlivanoglu, 2013).

During the population initialization phase, GVSAO generates initial individuals using the good-point-set method. Unlike random initialization, this strategy ensures a more uniform coverage of the search space and reduces the likelihood of initial solutions clustering in local regions. Let the population size be N and the optimization dimension be D, the initial population can then be represented as:

$$X = \{X_1, X_2, ..., X_i, ..., X_N\}, X_i \in R^D \tag{16}$$

where X_i denotes the i search individual. The individuals generated by the good point set need to be further mapped into the actual search space. Let l_b and u_b denote the lower and upper bounds of the search space, respectively. Then, the position of the i -th individual in the d -th dimension can be expressed as :

$$X_{i,d} = lb_d + g_{i,d}(ub_d - lb_d), d = 1,2,...,D \tag{17}$$

where $g_{i,d}$ denotes the d -th dimensional component of the i -th point in the good point set within the interval $[0,1]$. In this way, the initialized population can be more uniformly distributed in the search space, thereby improving the global exploration ability in the early stage of the algorithm.

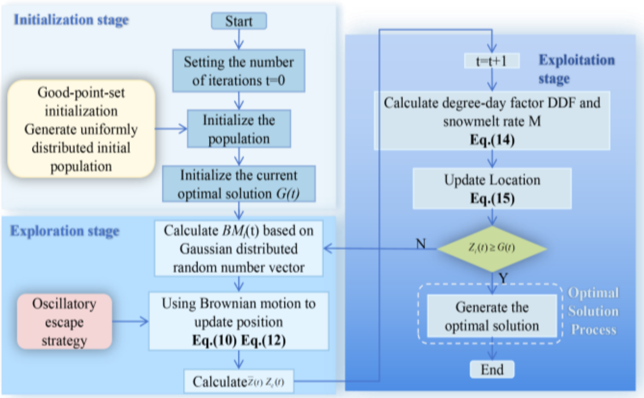


Fig. 4. Flowchart of the GVSAO Algorithm.

During the iterative search stage, GVSAO updates individual positions under the joint guidance of the current best individual, the elite pool, and the population center. Let X_c^t denote the population center at the t -th iteration. Then:

$$X_c^t = \frac{1}{N}\sum_{i=1}^{N} X_i^t \tag{18}$$

When updating an individual, an elite individual E is selected from the *Elite* pool, and the current global best individual X_{best} and the population center X_c are combined to guide the update. The update equation is given as follows:

$$X_i^{t+1} = E + R_B \cdot [r_1(X_{best}^t - X_i^t) + (1-r_1)(X_c^t - X_i^t)] \tag{19}$$

where R_B is a random perturbation term following a normal distribution, and r_1 is a random number within $[0,1]$. This update strategy enables each individual to move toward the current promising region while still referring to the overall population distribution, thereby maintaining a balance between local exploitation and global exploration.

To further reduce the risk of the algorithm falling into local optima, GVSAO introduces a periodic oscillatory escape strategy. When the oscillation-triggering condition is satisfied, the position of an individual is perturbed as follows:

$$X_i^{t+1} = X_i^t \cdot [1 + A_m(0.5 - rand)] \tag{20}$$

where A_m denotes the oscillation amplitude, and $rand$ is a random number within $[0,1]$. This strategy introduces periodic perturbations during the search process, allowing some individuals to escape from the current local region and preventing premature convergence of the population.

After each position update, a boundary-handling strategy is applied to ensure that all candidate solutions remain within the predefined search range:

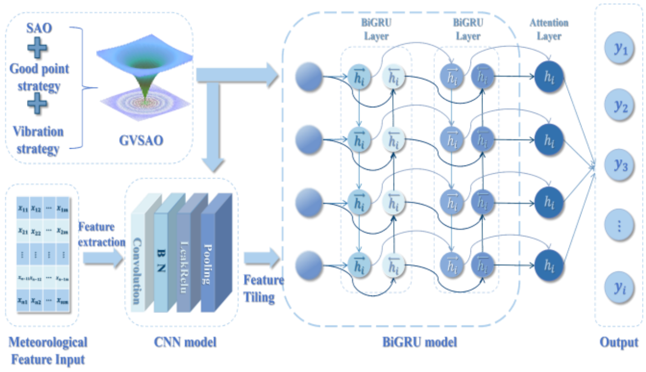


Fig. 5. Schematic Diagram of the GVSAO-CNN-BiGRU-Attention Model.

$$X_{i,d}^{t+1} = \min(\max(X_{i,d}^{t+1}, lb_d), ub_d), d = 1,2,...,D \tag{21}$$

Overall, GVSAO enhances population diversity through good-point-set initialization, achieves a better balance between exploration and exploitation by integrating elite guidance and population-center updating, and strengthens the capability of escaping local optima through the periodic oscillatory escape strategy. The complete optimization procedure of GVSAO is presented in Fig. 4.

3.DEVELOPMENT OF MODEL BASED ON GVSAO-CNN-BiGRU-Attention

3.1 Model Design Rationale

The GVSAO-CNN-BiGRU-Attention model proposed in this paper integrates complementary deep learning and optimization models to address the nonlinear and non-stationary nature of wind power generation time series.

In the CNN-BiGRU-Attention hybrid model, the CNN module captures local spatio-temporal features; the BiGRU obtains more comprehensive contextual information by processing data in both directions of the sequence, enabling the model to simultaneously consider both preceding and subsequent data points within the sequence. Simultaneously, the attention mechanism further amplifies the impact of crucial information, thereby enhancing the accuracy of short-term load forecasting. This integration enables the model to better understand and process complex relationships within sequential data.

However, the predictive efficiency of the CNN-BiGRU-Attention hybrid architecture largely depends on parameter selection, and traditional manual tuning methods cannot guarantee optimal performance. To overcome this limitation, the GVSAO algorithm is introduced for hyperparameter optimization. Leveraging its superior global exploration and convergence capabilities, it autonomously identifies the model's optimal hyperparameters, thereby improving predictive efficiency.

3.2 Model Components

The architecture of the proposed GVSAO-CNN-BiGRU-Attention model is illustrated in Fig. 5. The framework consists of an input layer, a CNN-based feature extraction module, a BiGRU-based temporal modeling module, an attention layer, and an output layer. Specifically, the CNN-BiGRU-Attention network serves as the forecasting backbone, which is designed to learn the nonlinear mapping between historical observations and future wind power outputs. Meanwhile, GVSAO is employed as a hyperparameter optimization module to automatically identify optimal parameter configurations for the forecasting network.

(1) Input Layer
The input layer receives the preprocessed wind power and meteorological data. Each sample is organized as a feature–time matrix, where rows correspond to different input variables and columns represent consecutive time steps. This representation preserves both multivariate feature information and temporal dynamics, providing an effective input structure for subsequent feature extraction and temporal dependency modeling.

(2) CNN Layer
The CNN layer is employed to extract local features from the input feature–time matrix. Through convolution and pooling operations, the network is capable of capturing local correlations among adjacent time steps as well as interactions between different input variables, thereby generating more compact and informative feature representations. The extracted feature maps are subsequently flattened and fed into the BiGRU layer for further temporal dependency modeling.

(3) BiGRU Layer
The BiGRU layer is employed to capture bidirectional temporal dependencies in wind power time series. BiGRU simultaneously learns temporal patterns from both forward and backward directions, thereby generating more comprehensive contextual representations. This capability enables the model to better characterize the dynamic variations and nonstationary behaviors inherent in wind power sequences, leading to improved temporal feature learning.

(4) Attention Layer
Although the hidden states generated by the BiGRU layer encode temporal information from different time steps, their contributions to the forecasting task are not equally important. To better exploit informative temporal features, an attention layer is incorporated following the BiGRU module. By assigning adaptive weights to hidden states according to their relevance to the prediction target, the attention mechanism enables the model to emphasize critical temporal patterns while suppressing redundant or less informative features. Consequently, the quality of feature representation is enhanced, leading to improved forecasting performance.

(5) Output Layer
The output layer employs a fully connected layer to map the weighted feature representation generated by the attention mechanism to the final forecasting output. For Dataset I, the model produces day-ahead wind power forecasts for 24 consecutive time intervals. For Dataset II, the model outputs the wind speed prediction for the next time step. In this way, the forecasting framework can be adapted to different prediction tasks while maintaining a unified model architecture.

(6) GVSAO Optimization Module
The forecasting performance of the CNN-BiGRU-Attention network is highly dependent on several critical hyperparameters, including the learning rate, convolution kernel size, and the number of BiGRU hidden units. To reduce the uncertainty associated with manual parameter tuning, GVSAO is employed to adaptively search for the optimal hyperparameter configuration. By integrating the optimization process with the forecasting network, the proposed framework aims to achieve improved predictive accuracy and training stability. The optimization variables, fitness function, parameter settings, and computational complexity of GVSAO are described in detail in Section 3.3.

3.3 GVSAO-Based Hyperparameter Optimization and Computational Complexity Analysis

In this model, GVSAO is used to optimize the key hyperparameters of CNN-BiGRU-Attention. Considering that the learning rate, convolution kernel size, and the number of BiGRU hidden units directly affect the model's convergence speed, local feature extraction capability, and temporal modeling ability, this paper takes the three as the main optimization targets. The i -th search individual is represented as:

$$X_i = [\eta_i, k_i, n_i] \tag{22}$$

where η_i denotes the learning rate, k_i denotes the CNN convolution kernel size, and n_i denotes the number of BiGRU hidden units.

The learning rate is a continuous variable and is updated directly during optimization; the convolution kernel size and the number of hidden units are discrete variables, so they are rounded before model training. This paper sets the learning rate search range to [0.001, 0.01], the convolution kernel size search range to [2, 5], and the number of BiGRU hidden units search range to [100, 120].

In each fitness evaluation, GVSAO constructs the CNN-BiGRU-Attention model based on the current candidate hyperparameter combination and completes the training and prediction. This paper uses $MAPE$ as the fitness function to measure the prediction error of the model under the current parameter combination:

$$f(X_i) = MAPE(Y, \hat{Y}) = \frac{1}{m}\sum_{j=1}^{m} \left| \frac{Y_j - \hat{Y}_j}{Y_j} \right| \tag{23}$$

where Y_j and $\hat{Y}_j$ denote the true value and predicted value of the j -th sample, respectively, and m denotes the number of test samples.

The optimization goal of GVSAO is to find the parameter combination that minimizes the fitness function:

$$X_{best} = \arg\min_{X_i} f(X_i) \tag{24}$$

In this experiment, the population size of GVSAO is set to 4, the maximum number of iterations is set to 10, and the optimization dimension is 3. The good-point-set initialization and periodic oscillatory strategy are used as the main improvement mechanisms of the algorithm. In the oscillatory strategy, the amplitude A_m is set to 5, and the oscillation frequency f_r is set to 10. The model is trained using the Adam optimizer, with the maximum number of training epochs set to 300, the mini-batch size set to 20, and the gradient threshold set to 1. For reproducibility, the key search settings of GVSAO are summarized in Table 1.

Table 1. Key settings of GVSAO optimization.

Parameter	Setting
Population size	4
Maximum iterations	10
Optimization dimension	3
Learning rate range	[0.0001,0.01]
CNN kernel size range	[2,5]
BiGRU hidden units range	[8,160]

In the optimization process, the computational cost of GVSAO mainly comes from model training and fitness evaluation. Let the population size be P, the maximum number of iterations be T, the optimization dimension be D, and the computational cost of a single CNN-BiGRU-Attention model training and testing process be C_{train}. Then, the total computational complexity of one GVSAO optimization run can be expressed as:

$$O_{GVSAO} = O(PT(D + C_{train})) \tag{25}$$

where PTD denotes the computational cost of population position updating, boundary handling, and fitness sorting operations, while PTC_{train} denotes the computational cost of model training and testing under different candidate parameter combinations. Since the computational cost of deep learning model training is much higher than that of population updating, the overall complexity can be approximately expressed as:

$$O_{GVSAO} \approx O(PTC_{train}) \tag{26}$$

In this study, $P = 4$, $T = 10$, and $D = 3$. Therefore, the complete GVSAO optimization process requires approximately 32–40 candidate parameter evaluations under the settings used in this study. The running time of single model training and the complete GVSAO optimization process is further recorded, and the results are shown in Table 2.

Table 2. Runtime analysis of single model training and GVSAO optimization.

Dataset	Process	Evaluations	Running time (s)
Dataset I	Single	1	24.294
	GVSAO	40	1709.542
Dataset II	Single	1	18.805
	GVSAO	32	6498.125

4. EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Data Description and Preprocessing

To evaluate the effectiveness of the proposed GVSAO-CNN-BiGRU-Attention model under different wind forecasting scenarios, two real-world wind energy datasets are employed in this study.

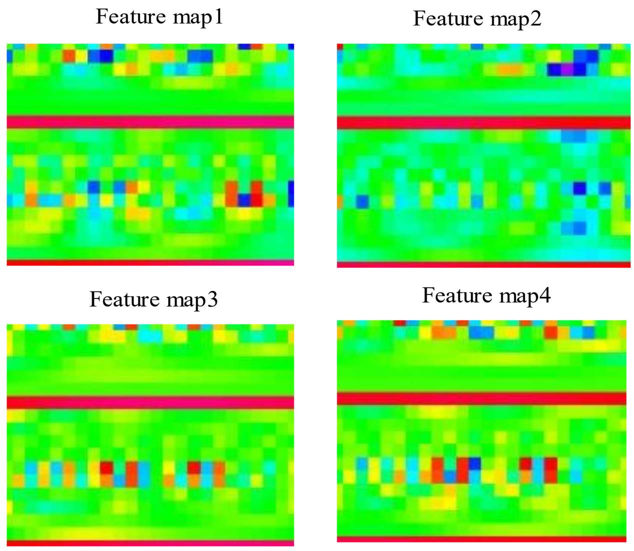


Fig 6. Feature mappings of the first four samples from Dataset I.

Dataset I is collected from the Sotavento wind farm in Spain and contains wind power measurements together with multiple meteorological variables recorded at an hourly sampling interval. Considering the characteristics of the day-ahead wind power forecasting task, 18 meteorological features observed during the previous 24 h are used as input variables to predict wind power generation over the subsequent 24 h. Accordingly, each sample in Dataset I is represented by an input feature matrix of size (18×24), with the corresponding output consisting of a (1×24) wind power sequence.

To further assess the adaptability of the proposed model across different data sources, a publicly available wind turbine dataset from Sao Paulo is adopted as Dataset II. This dataset contains wind speed measurements and related operational variables collected from wind turbine systems. Unlike Dataset I, which focuses on day-ahead wind power forecasting, Dataset II is designed for short-term wind speed prediction. A sliding-window strategy is employed to construct the samples, where observations from the previous 24 h are used to predict wind speed at the next time step. Consequently, the input samples are also organized as (18×24) feature matrices, while the output corresponds to the wind speed value of the subsequent time step.

Since the two datasets differ in data source, forecasting target, and prediction task, they provide a comprehensive basis for evaluating the generalization capability and practical applicability of the proposed framework under diverse wind forecasting scenarios.

Prior to model development, the raw data are subjected to a series of preprocessing procedures. Missing, abnormal, and invalid records are removed, after which the time series are reconstructed at a uniform hourly sampling interval to ensure temporal continuity of the input samples. Subsequently, all input variables and prediction targets are normalized. Owing to the substantial differences in scale and magnitude among wind power, wind speed, and meteorological variables, directly feeding the raw data into the model may lead to unstable training behavior and slow convergence. Therefore, the Min–Max normalization method is adopted to map all variables into a unified numerical range. The normalization process is expressed as follows:

$$x = \frac{X_i - X_{min}}{X_{max} - X_{min}} \tag{27}$$

where, x denotes the normalized value. X_i denotes the i-th value, X_{min} is the minimum value in the dataset, and X_{max} is the maximum value in the dataset. This method can define the original data within a specific range by utilizing the maximum and minimum values of the variables, scaling the feature attributes of the data to a unified dimension, thereby improving the model's accuracy and accelerating its convergence speed.

To provide a visual illustration of the sample construction process, Fig. 6 and Fig. 7 present the feature mappings of the first four samples from Dataset I and Dataset II, respectively. Each sample consists of observations from 18 features across 24 consecutive time steps, which is consistent with the input structure adopted by the proposed forecasting framework.

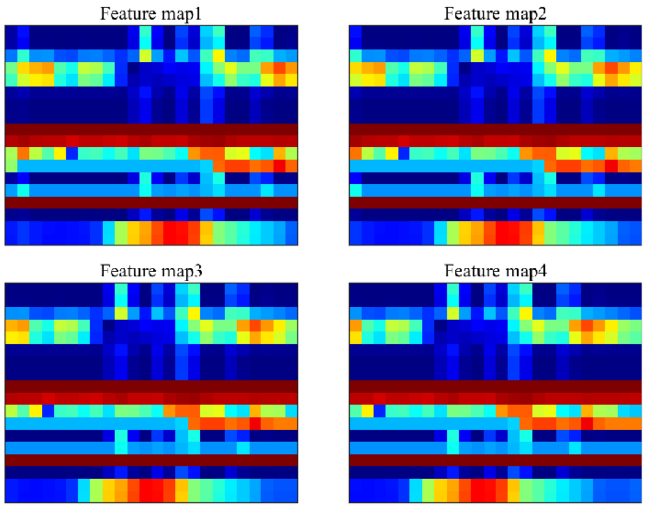


Fig. 7. Feature mappings of the first four samples from Dataset II

4.2 Experimental Setup

All experiments were conducted using the MATLAB platform. To ensure a fair comparison, all forecasting models were evaluated using the same data partitioning strategy and training procedure. The test set was reserved exclusively for final performance evaluation and was not involved in either model training or hyperparameter optimization.

Model training was performed using the Adam optimizer with a batch size of 20 and a gradient threshold of 1. For baseline models without intelligent optimization mechanisms, hyperparameters were assigned fixed values. In contrast, for optimization-based models, key hyperparameters were determined automatically by the corresponding optimization algorithms.

The performance evaluation of predictive models necessitates reliance on specific quantitative metrics. This study conducts systematic error analysis by comparing model predictions with validation datasets to assess model efficacy. Accordingly, the coefficient of determination (R^2), root mean square error (*RMSE*), and mean absolute percentage error (*MAPE*) are selected as core evaluation metrics. The precise mathematical definitions of these indicators are articulated as follows:

Table 3. Error Analysis Functions Used for Model Evaluation.

Error Analysis Metrics	Formula
R^2	$R^2 = 1 - \frac{\sum_{i=1}^{n}(X_{fore}^{i} - X_{true}^{i})^2}{\sum_{i=1}^{n}(\overline{X_{ave}} - X_{true}^{i})^2}$
RMSE	$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(X_{fore}^{i} - X_{true}^{i})^2}{n}}$
MAPE	$MAPE = \frac{\sum_{i=1}^{n} \lvert X_{fore}^{i} - X_{true}^{i} \rvert}{n} \times 100\%$

where, X_{true}^{i} denotes the actual value at time i, X_{fore}^{i} represents the predicted value at time i, $\overline{X_{ave}}$ is the mean value at time i, and n is the number of predicted data points.

4.3 Prediction Results and Analysis

(1) Results on Dataset I

To evaluate the effectiveness of the proposed model for day-ahead wind power forecasting, experiments were first conducted on Dataset I. Table 4 summarizes the mean performance metrics and standard deviations obtained from five repeated trials for all models, while Figs. 8-11 present forecasting curves, the evaluation metric comparisons, error histograms, and regression analysis results, respectively.

Table 4. Comparison of Prediction Results of Different Models Based on Dataset I.

Model	MAPE	RMSE	R^2
GRU	0.52654 ± 0.00701	4.37460 ± 0.04801	0.24880 ± 0.01952
GRU-Attention	0.50690 ± 0.01273	2.75700 ± 0.06339	0.40260 ± 0.02836
CNN-GRU-Attention	0.45874 ± 0.01539	2.19360 ± 0.05286	0.33155 ± 0.02294
CNN-BiGRU-Attention	0.11877 ± 0.01867	4.06680 ± 0.10347	0.90427 ± 0.03072
CPO-CNN-GRU-Attention	0.13645 ± 0.01821	0.81557 ± 0.08353	0.90760 ± 0.03185
GVSAO-CNN-BiGRU-Attention	0.11725 ± 0.01195	0.73206 ± 0.07114	0.92550 ± 0.01404

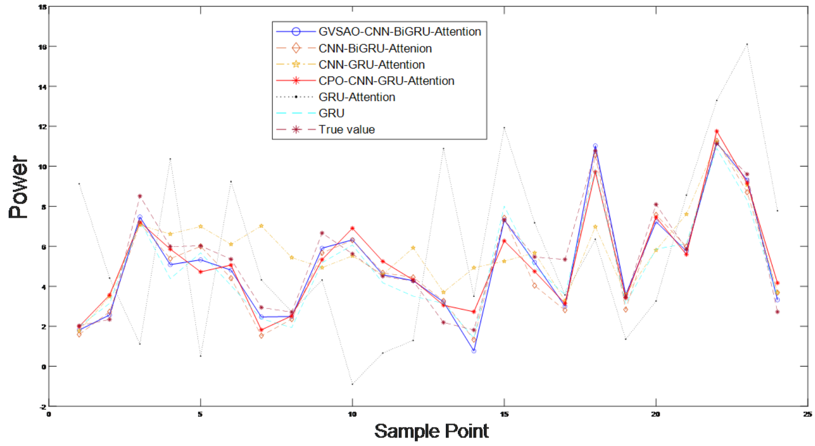
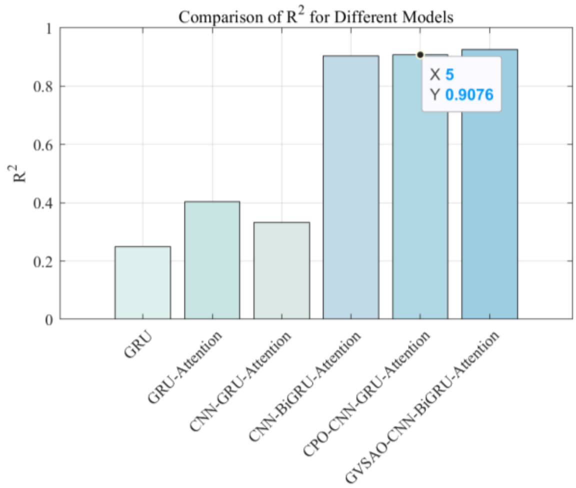


Fig. 8. Comparison of Prediction Results of Different Models for Dataset.

As shown in Table 4, the GVSAO-CNN-BiGRU-Attention model achieved strong overall performance on Dataset I, with $RMSE$, $MAPE$, and R^2 values of 0.73206 ± 0.07114, 0.11725 ± 0.01195, and 0.92550 ± 0.01404, respectively. Compared with the CNN-BiGRU-Attention model without GVSAO optimization, the proposed framework yielded lower prediction errors and improved fitting metrics, indicating that adaptive hyperparameter optimization via GVSAO effectively enhances forecasting performance. The bar charts of evaluation metrics further corroborate this trend, showing that the proposed model achieves a more balanced performance between error reduction and model fitting.



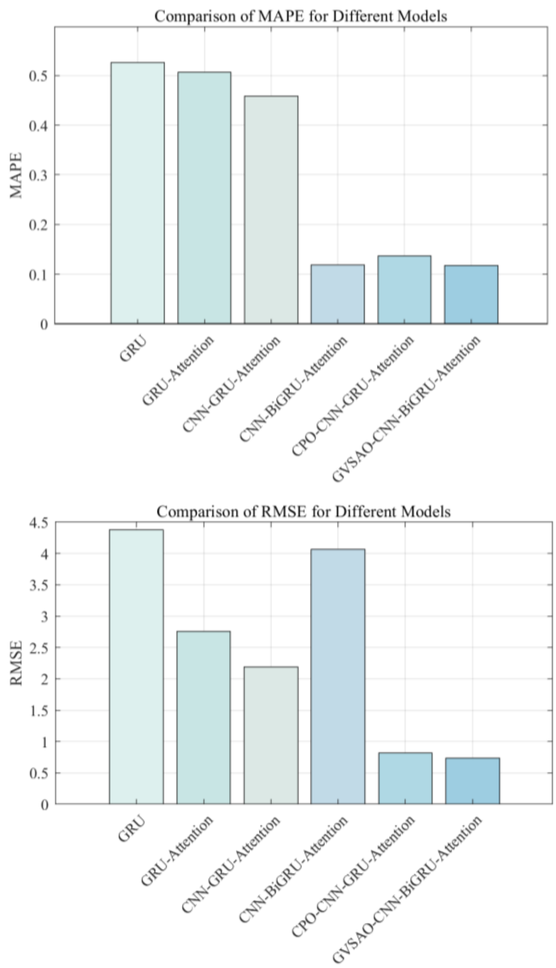
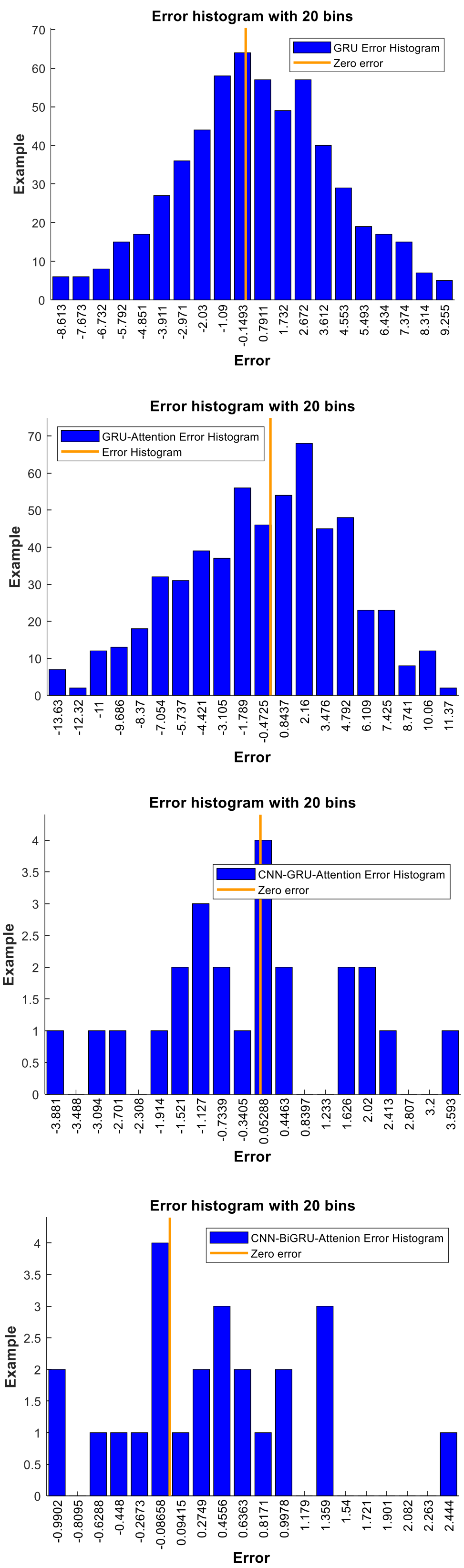


Fig. 9. Experimental Error Analysis for Dataset I.

Fig.10 presents an error histogram with 20 bins, where the methodological errors in this study exhibit pronounced concentration near zero. The error distribution of the GRU model exhibits a relatively wide range, reflecting its limited ability to fit complex load variations. While the GRU-Attention model improves error concentration to some extent, it still exhibits numerous errors of moderate magnitude. By adding convolutional feature extraction, the CNN-GRU-Attention model makes the error distribution even more similar. The CNN-BiGRU-Attention model exhibits a more symmetrical error distribution due to bidirectional temporal modeling. The CPO-CNN-GRU-Attention model enhances parameter optimization through algorithmic refinement, though it retains a small number of high-error samples. In contrast, the proposed GVSAO-CNN-BiGRU-Attention model demonstrates optimal performance in error concentration, distribution symmetry, and extreme error control, exhibiting superior robustness and generalization capabilities.



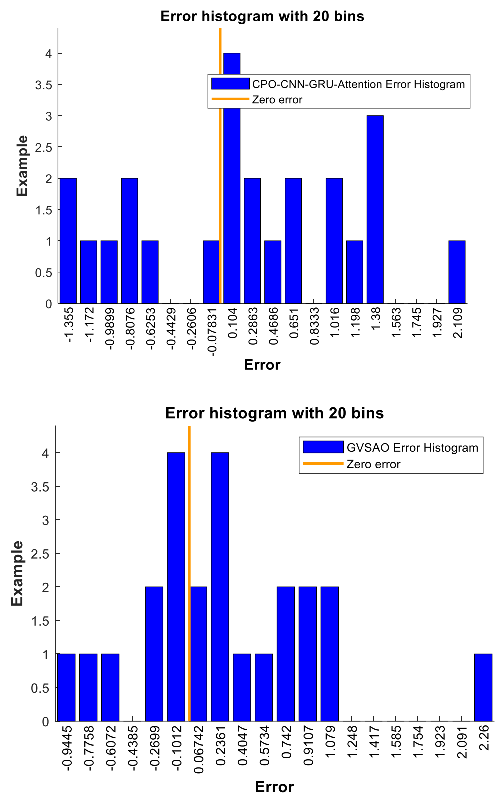
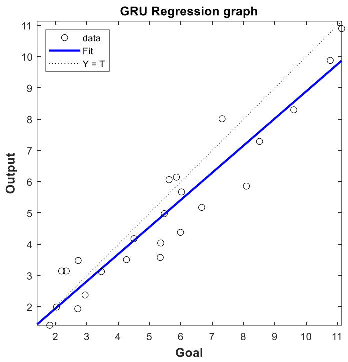
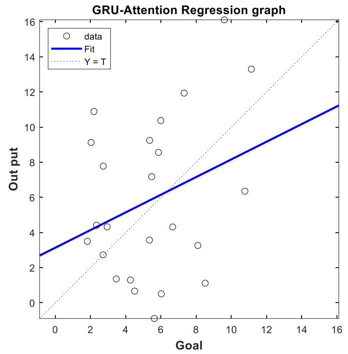
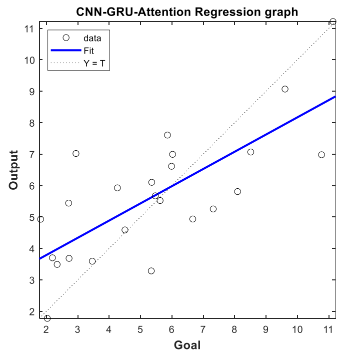


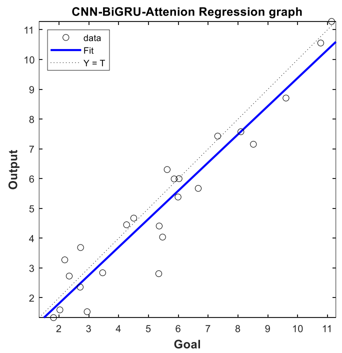
Fig. 10. Comparison of Error Histograms on Dataset I.

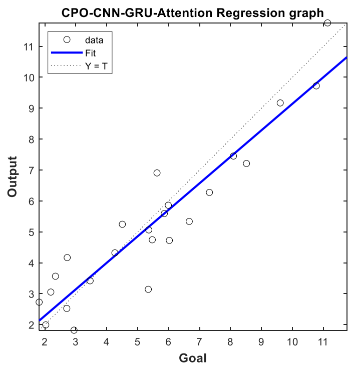
To further analyze the model's fitting performance and generalization capability, this paper plots the prediction values against actual values for the six models on the test set. As shown in Fig.11, the scatter points for each model generally follow the ideal fitting line Y=T, indicating that the models can effectively capture the trend of load changes. The fitting results of the six models reveal significant differences in fitting accuracy and stability among them. The baseline GRU model exhibits a certain deviation from the ideal fit line Y=T, with scattered points distributed relatively widely. After incorporating the Attention mechanism, the GRU-Attention model's fit line approaches the ideal line more closely, though deviations persist in certain intervals. The CNN-GRU-Attention model improves the overall fitting trend through convolutional feature extraction, further enhancing the consistency between the fitted line and Y=T; Under the bidirectional temporal modeling effect of the CNN-BiGRU-Attention model, the fitted line aligns more closely with the ideal line, and the scatter distribution becomes more symmetrical; The CPO-CNN-GRU-Attention model showed improvements in parameter optimization but still exhibited minor deviations from the samples. In contrast, the proposed model's fitted line nearly overlaps with the ideal fitted line, featuring the most concentrated and uniform scatter distribution without noticeable systematic bias. This demonstrates superior fitting accuracy and generalization stability compared to other models.











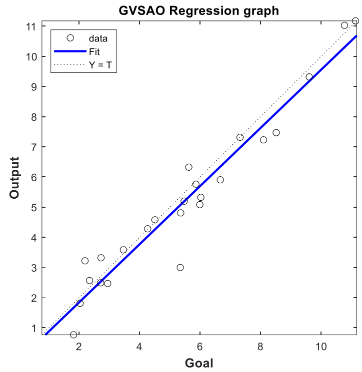


Fig. 11. Regression curve on Dataset I.

(2) Results on Dataset II

To further evaluate the cross-scenario applicability of the proposed framework, short-term wind speed forecasting experiments were conducted on Dataset II. Unlike Dataset I, which focuses on day-ahead wind power prediction, Dataset II uses wind speed as the forecasting target and constructs samples based on a sliding-window strategy, thereby providing a more challenging benchmark for assessing the model's capability in handling short-term fluctuations. Table 5 summarizes the evaluation results of all models, while Figs. 12– 15 present forecasting curves, the corresponding metric comparisons, error histograms, and regression analysis results.

As shown in Table 5 and Fig. 12, the GVSAO-CNN-BiGRU-Attention model maintains strong forecasting performance on Dataset II, achieving overall superior results in terms of $RMSE$, $MAPE$, and (R^2) compared with the benchmark models. The forecasting curves indicate that, despite the pronounced short-term fluctuations in wind speed, the proposed framework is capable of effectively tracking the underlying temporal dynamics. Furthermore, the error histograms and regression plots reveal a relatively concentrated error distribution and a strong agreement between predicted and observed values, further demonstrating the effectiveness of the proposed model.

Since the two datasets differ significantly in data source, forecasting target, and sample construction method, their results should not be directly compared. However, the comparative experiments on each dataset show that the proposed framework is not limited to wind power datasets and still achieves favorable performance compared with other models in the wind speed forecasting scenario. This demonstrates the adaptability and robustness of the proposed GVSAO-CNN-BiGRU-Attention framework across different forecasting tasks.

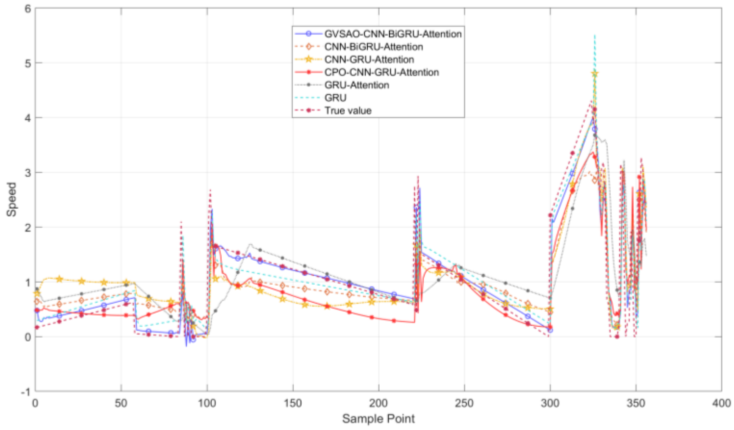
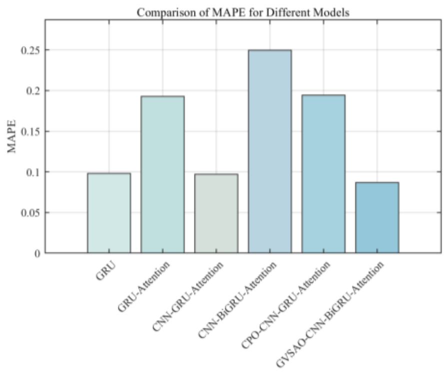
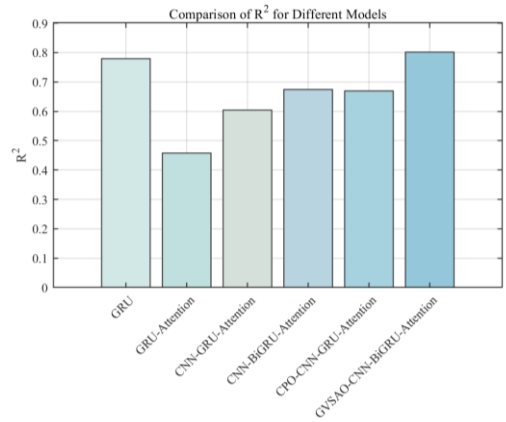


Fig. 12. Comparison of Prediction Results of Different Models for Dataset II.

Table 5. Comparison of Prediction Results of Different Models Based on Dataset II.

Model	MAPE	RMSE	R^2
GRU	0.0980 ± 0.1005	0.4451 ± 0.0867	0.7794 ± 0.1119
GRU-Attention	0.1929 ± 0.0091	0.6982 ± 0.0081	0.4572 ± 0.0080
CNN-GRU-Attention	0.0971 ± 0.0265	0.5967 ± 0.0238	0.6036 ± 0.0238
CNN-BiGRU-Attention	0.2495 ± 0.0182	0.5406 ± 0.0141	0.6747 ± 0.0143
CPO-CNN-GRU-Attention	0.1942 ± 0.0398	0.5455 ± 0.0109	0.6687 ± 0.0105
GVSAO-CNN-BiGRU-Attention	0.0868 ± 0.0335	0.4228 ± 0.0247	0.8010 ± 0.0253





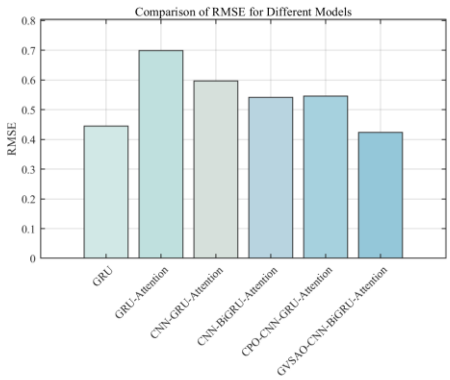
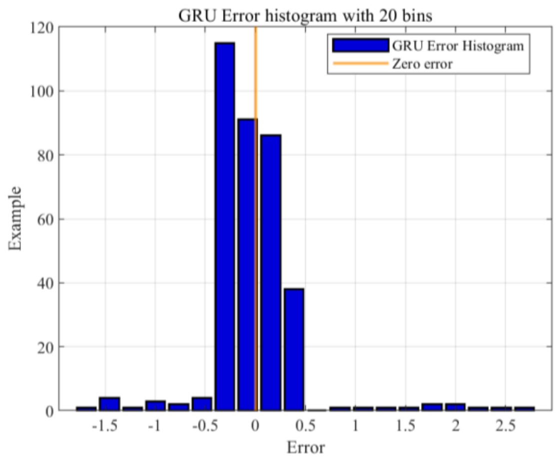
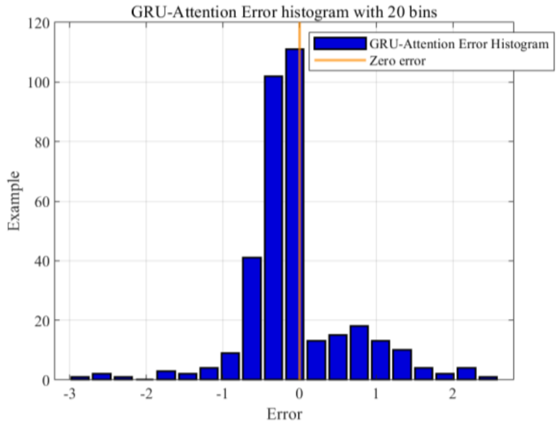
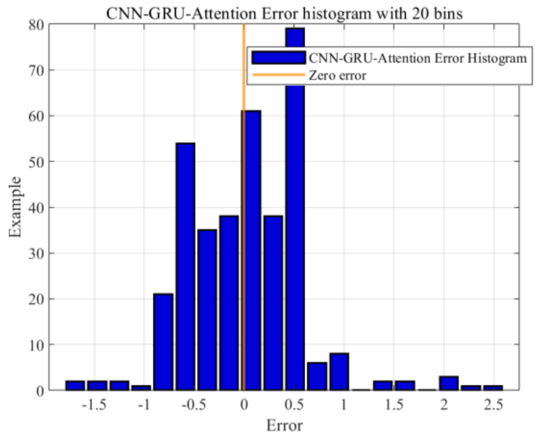
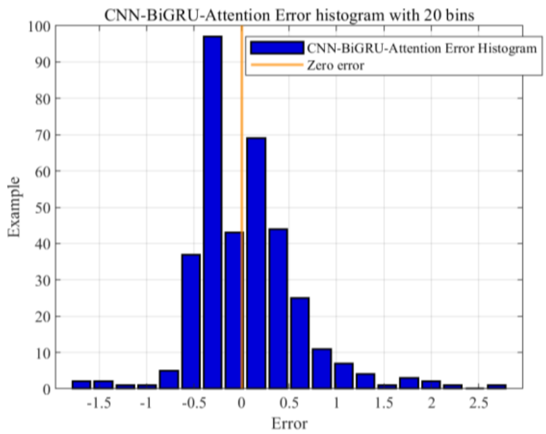


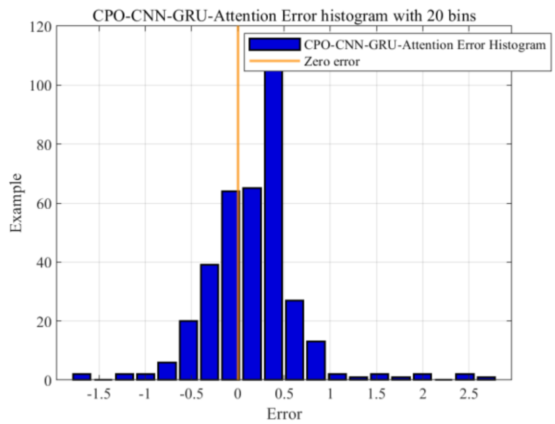
Fig. 13. Experimental Error Analysis for Dataset II.











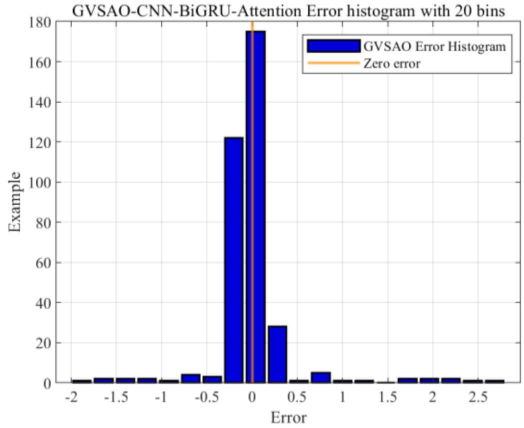
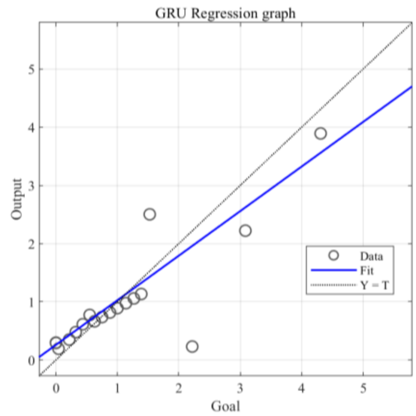
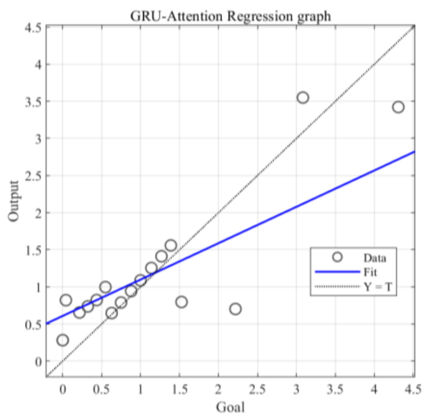
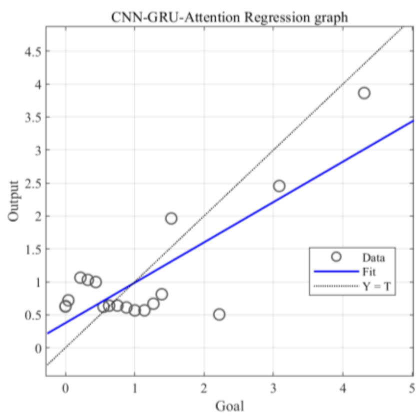


Fig. 14. Comparison of Error Histograms on Dataset II.







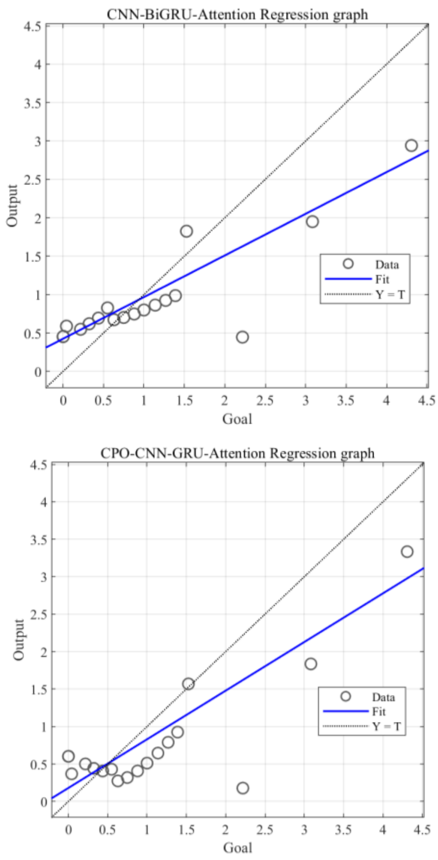


Fig. 15. Regression curve on Dataset II.

4.4 Ablation Study

To evaluate the contribution of individual components to the overall model performance, ablation experiments were conducted on Dataset I, with the results presented in Table 6 and Fig. 16. The experiments were designed to assess both network architecture and optimization strategy effects. For the network architecture, variants included the removal of the attention mechanism, replacing BiGRU with a unidirectional GRU, and constructing a CNN-BiGRU configuration. For the optimization strategy, modifications involved replacing the good-point-set initialization with random initialization and omitting the periodic oscillatory escape mechanism.

As shown in Table 6, the complete GVSAO-CNN-BiGRU-Attention framework achieved the best overall forecasting performance among all model variants. The removal of the attention mechanism resulted in degraded performance, indicating that the model's ability to emphasize informative temporal features was weakened. When BiGRU was replaced with a unidirectional GRU, the forecasting accuracy further declined, suggesting that bidirectional temporal dependency learning plays an important role in capturing the complex dynamics of time series data for wind energy. In addition, the results of the CNN-BiGRU variant demonstrate that the combination of convolutional feature extraction and recurrent sequence modeling alone is insufficient to achieve optimal predictive performance. These findings confirm the effectiveness of both the attention mechanism and the BiGRU architecture in modeling complex wind power sequences.

Table 6. Comparison of Prediction Results of Different Models Based on Dataset I.

Model	MAPE	RMSE	R^2
w/o GVSAO	0.11877 ± 0.01867	4.06680 ± 0.10347	0.90427 ± 0.03072
w/o BiGRU	0.42100 ± 0.01115	2.10970 ± 0.08070	0.38170 ± 0.01646
w/o Attention	0.52280 ± 0.01291	2.38040 ± 0.12783	0.21280 ± 0.02950
w/o Good-point-set Init	0.54340 ± 0.00520	2.52970 ± 0.04272	0.11100 ± 0.00810
w/o Oscillatory Escape	0.52170 ± 0.00742	2.56050 ± 0.04210	0.38921 ± 0.00873
GVSAO-CNN-BiGRU-Attention	0.11725 ± 0.01195	0.73206 ± 0.07114	0.92550 ± 0.01404

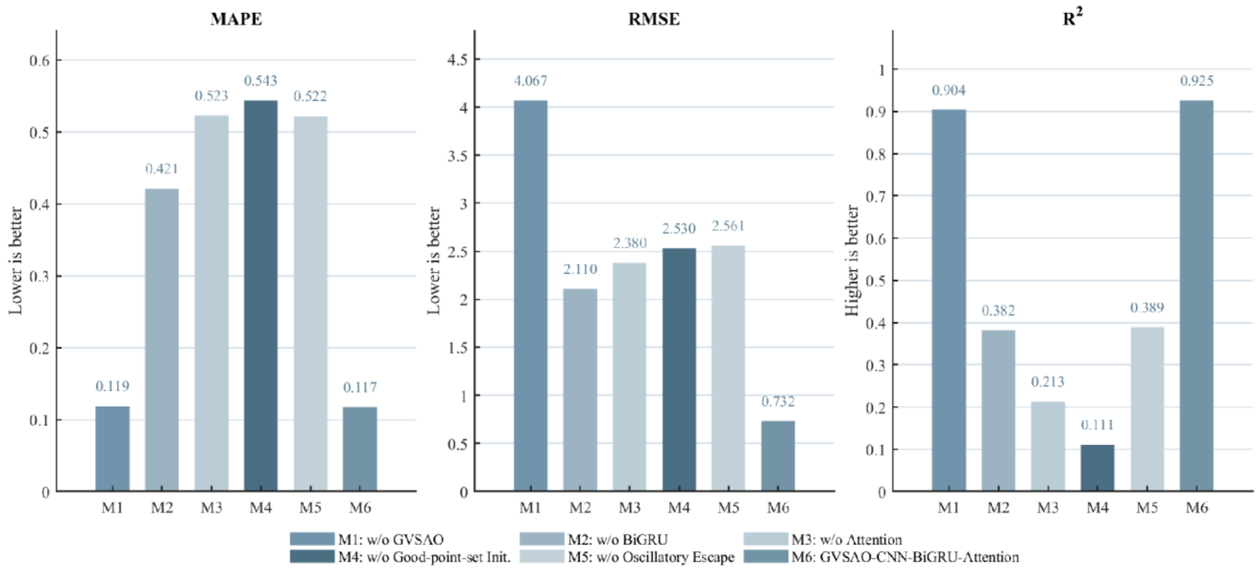


Fig. 16. Comparison of evaluation metrics in the ablation study.

From the perspective of the optimization strategy, both the RandomInit and NoOscillation variants exhibited inferior performance compared with the complete GVSAO framework. Replacing the good-point-set initialization with random initialization reduced the uniformity of population distribution within the search space, thereby weakening the diversity of candidate solutions during the early search stage. Similarly, removing the oscillatory escape mechanism limited the algorithm's ability to escape local optima, increasing the risk of premature convergence. These results indicate that both the good-point-set initialization strategy and the oscillatory escape mechanism contribute to improving the search quality of GVSAO, which in turn leads to enhanced forecasting performance.

The bar chart in Fig. 16 further corroborates the quantitative results reported in Table 6. The complete GVSAO-CNN-BiGRU-Attention model consistently achieves lower error metrics and higher goodness-of-fit indicators than the ablated variants. Overall, the ablation results demonstrate that the CNN, BiGRU, and attention modules, together with the good-point-set initialization and periodic oscillatory escape strategies incorporated in GVSAO, all contribute positively to the overall forecasting performance of the proposed framework.

4.5 Discussion

The experimental results on the two datasets indicate that the GVSAO-CNN-BiGRU-Attention model is effective beyond a single dataset. Dataset I, used for day-ahead wind power forecasting, and Dataset II, employed for short-term wind speed prediction, differ in both data source and forecasting target. Despite these differences, the model demonstrates robust performance across both tasks, reflecting its cross-scenario generalization capability. The lower performance metrics observed on Dataset II are primarily attributed to stronger short-term wind speed fluctuations and the increased difficulty of the prediction task.

Ablation studies reveal that the improvements in forecasting performance arise from the combined effects of the network architecture and the optimization strategy. Specifically, CNN, BiGRU, and the attention mechanism contribute to local feature extraction, bidirectional temporal modeling, and the selection of critical temporal information, respectively, while GVSAO enhances model stability through adaptive hyperparameter optimization. Performance degradation observed when either the good-point-set initialization or the periodic oscillatory escape mechanism is removed further confirms that the optimization process itself significantly influences the final predictive accuracy.

5. CONCLUSIONS

A GVSAO-CNN-BiGRU-Attention forecasting model is proposed to enhance the accuracy and stability of wind power prediction for complex wind power time series.

(1) An enhanced optimization algorithm, termed GVSAO, was developed based on the original SAO. Specifically, a good-point-set initialization strategy was introduced to improve the distribution quality of the initial population and enhance global exploration during the early search stage. In addition, a periodic oscillatory escape mechanism was incorporated to alleviate premature convergence and strengthen the algorithm's ability to escape local optima. Through the integration of these strategies, GVSAO achieved improved global optimization performance.

(2) In terms of the forecasting framework, a hybrid CNN-BiGRU-Attention model was developed. The CNN module extracts local patterns from multivariate inputs, the BiGRU captures bidirectional temporal dependencies, and the attention mechanism emphasizes the most informative temporal features. Building upon this architecture, GVSAO was employed to adaptively optimize critical hyperparameters, including the learning rate, convolution kernel size, and the number of hidden units. This integration reduces the reliance on manual parameter tuning and enhances the stability and effectiveness of the model training process.

(3) The experimental results demonstrate that the proposed GVSAO-CNN-BiGRU-Attention model achieves superior forecasting performance on both real-world wind energy datasets. Dataset I was employed for day-ahead wind power forecasting, while Dataset II was used for short-term wind speed prediction. These datasets differ in source, forecasting target, and sample construction methodology, providing a rigorous basis for evaluating cross-scenario generalization. The results indicate that the proposed model is effective not only on a single dataset but also exhibits adaptability across different forecasting scenarios.

Future research will incorporate wind energy datasets collected from a wider range of geographical regions, seasonal conditions, and operational scenarios to provide a more comprehensive assessment of the model's generalization capability. In addition, although the integration of intelligent optimization algorithms can improve forecasting performance, it also introduces additional computational overhead during the offline hyperparameter search process. Therefore, future studies will investigate more lightweight and efficient optimization strategies to enhance the deployment efficiency of the proposed framework in practical wind forecasting applications.

ACKNOWLEDGEMENTS

This research was funded by the Science and Technology Project of Hebei Education Department (Grant No. CXZX2025039) and sponsored by the Beijing-Tianjin-Hebei Natural Science Foundation Cooperation Project (Grant No. 25JJJJC0049).

REFERENCES

Chen F, Tsou J Y. Assessing the effects of convolutional neural network architectural factors on model performance for remote sensing image classification: An in-depth investigation[J]. *International Journal of Applied Earth Observation and Geoinformation*, 2022, 112: 102865.

Du L, Gao R, Suganthan P N, et al. Bayesian optimization based dynamic ensemble for time series forecasting[J]. *Information Sciences*, 2022, 591: 155-175.

Fang Z, Guoshuai L, Hao T, et al. Dung beetle optimization algorithm based on quantum computing and multistrategy fusion for solving engineering problems[J]. *Expert Systems With Applications*, 2024, 236. DOI:10.1016/J.ESWA.2023.121219.

Gao J, Sun Y, Kim H, et al. Improving wind power prediction with advanced temporal and frequency domain processing combined with error correction[J]. *Scientific Reports*, 2025, 15(1): 44300-44300.

Ghanbari E, Avar A. Short-term wind power forecasting using the hybrid model of multivariate variational mode decomposition (MVMD) and long short-term memory (LSTM) neural networks[J]. *Electrical Engineering*, 2024, 107(3): 1-31. DOI:10.1007/S00202-024-02685-1.

Karaman Ö A. Prediction of Wind Power with Machine Learning Models[J]. *Applied Sciences*, 2023, 13(20).

Karamichailidou D, Kaloutsa V, Alexandridis A. Wind turbine power curve modeling using radial basis function neural networks and tabu search[J]. *Renewable Energy*, 2021, 163: 2137-2152.

Khouili O, Hanine M, Louzazni M, et al. Smart solar power prediction using an FFT-Infused ShuffleNet regressor: A high-accuracy lightweight framework[J]. *International Journal of Electrical Power & Energy Systems*, 2025, 172: 111240.

Kim B, Ko K. Application of multi-source data for improvement of day-ahead wind power forecasting with machine learning models[J]. *Journal of Mechanical Science and Technology*, 2025: 1-10.

Lei P, Ma F, Zhu C, et al. LSTM Short-Term Wind Power Prediction Method Based on Data Preprocessing and Variational Modal Decomposition for Soft Sensors[J]. *Sensors* (Basel), 2024, 24(8).

Lingyun D, Sanyang L. Snow ablation optimizer: A novel metaheuristic technique for numerical optimization and engineering design[J]. *Expert Systems With Applications*, 2023, 225.

Liu H Z, Wang T C, Wei L H, et al. A wavelet-LSTM model for short-term wind power forecasting using wind farm SCADA data[J]. *Expert Systems With Applications*, 2024, 247: 123237.

Liu H, Li P, Wang Q. Short-term prediction of wind power based on BKA-CNN-BiGRU-Attention[J]. Available at SSRN 5260141.

Liu S, Xu T, Du X, et al. A hybrid deep learning model based on parallel architecture TCN-LSTM with Savitzky-Golay filter for wind power prediction[J]. *Energy Conversion and Management*, 2024, 302: 118122.

Pehlivanoglu V Y. A New Particle Swarm Optimization Method Enhanced With a Periodic Mutation Strategy and Neural Networks[J]. *IEEE Transactions on Evolutionary Computation*, 2013, 17(3): 436-452.

Ren Y, Xia L, Wang Y. An improved GM(1,1) forecasting model based on Aquila Optimizer for wind power generation in Sichuan Province[J]. *Soft Computing*, 2024, 28(15): 8785-8805.

Ren D, Ma J, Liu H, et al. The IVMD-CNN-GRU-Attention Model for Wind Power Prediction with Sample Entropy Fusion[J]. *IEEE Access*, 2024.

Sherry G, Rajalakshmi K. A CNN encoder decoder LSTM model for sustainable wind power predictive analytics[J]. *Sustainable Computing: Informatics and Systems*, 2023, 38.

Shuangxin W, Jiarong S, Wei Y, et al. High and low frequency wind power prediction based on Transformer and BiGRU-Attention[J]. *Energy*, 2024, 288: 129753.

Wang H K, Zhang X, Long H, et al. W-FENet: Wavelet-based fourier-enhanced network model decomposition for multivariate long-term time-series forecasting[J]. *Neural Processing Letters*, 2024, 56(2): 43.

Wang H, Ma S, Dong L, et al. Deepnet: Scaling transformers to 1,000 layers[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, 46(10): 6761-6774.

Xiao H, Songsong Z, Zhenwei H, et al. Short-term wind power prediction model based on improved ACA-GRU neural network[J]. *Journal of Physics: Conference Series*, 2023, 2527(1).

Xiang P, Zhongdi L, Rong X, et al. Early warning of damaged wind turbine blades using spatial–temporal spectral analysis of acoustic emission signals[J]. *Journal of Sound and Vibration*, 2022, 537.

Yang G, Zha D, Cao D, et al. Time for a change: Rethinking the global renewable energy transition from the Sustainable Development Goals and the Paris Climate Agreement[J]. *The Innovation*, 2024, 5(2): 100582.

Yang S, Yuan A, Yu Z. A novel model based on CEEMDAN, IWOA, and LSTM for ultra-short-term wind power forecasting[J]. *Environmental Science and Pollution Research*, 2023, 30(5): 11689-11705.

Yang Y, Lou H, Wu J, et al. A survey on wind power forecasting with machine learning approaches[J]. *Neural Computing and Applications*, 2024, 36(21): 12753-12773.

Yang Y, Song W, Han S, et al. Power forecasting method of ultra-short-term wind power cluster based on the convergence cross mapping algorithm[J]. *Global Energy Interconnection*, 2025, 8(1): 28-42.

Yang Y, Zhao Y. Short-Term Forecasting Approach of Wind Power Relying on NWP-CEEMDAN-LSTM[J]. *Processes*, 2025, 13(10): 3276.

Zavala-Mondragón L A, de With P H N, van der Sommen F. A signal processing interpretation of noise-reduction convolutional neural networks: Exploring the mathematical formulation of encoding-decoding CNNs[J]. *IEEE Signal Processing Magazine*, 2023, 40(7): 38-63.

Zeni Z, Sining Y, Lingyun J, et al. Hybrid VMD-CNN-GRU-based model for short-term forecasting of wind power considering spatio-temporal features[J]. *Engineering Applications of Artificial Intelligence*, 2023, 121.

Zhang Z, Wang B. A CBAM-LSTM-based short-term power forecasting method for wind power clusters[J]. *Journal of Northeast Electric Power University*, 2024, 44(1): 1–8.

Zhang H, Li T, Jing A, et al. Sequence–spectrogram fusion network for wind turbine diagnosis through few-shot time-series classification[J]. *Advanced Engineering Informatics*, 2025, 64: 102976.

Zhao W, Zhang Z, Khodadadi N, et al. A deep learning model coupled with metaheuristic optimization for urban rainfall prediction[J]. *Journal of Hydrology*, 2025, 651: 132596.