ITC 3/55
Information Technology and Control
Vol. 55 / No. 3/ 2026
pp. 1383-1406
DOI 10.5755/j01.itc.55.3.45339

An EGWO-XGB-RF Feature Engineering Approach for Short-Term Photovoltaic Power Forecasting

Received 2026/06/22 | Accepted after revision 2026/09/02

HOW TO CITE: Zeng, Y., Huang, P. (2026) An EGWO-XGB-RF Feature Engineering Approach for Short-Term Photovoltaic Power Forecasting. *Information Technology and Control*, 55(3), 1382-1405. https://doi.org/10.5755/j01.itc.55.3.45339

An EGWO-XGB-RF Feature Engineering Approach for Short-Term Photovoltaic Power Forecasting

Yihan Zeng[#,*]

Chongqing University-University of Cincinnati Joint Co-op Institute, Chongqing University, Chongqing, 401331, China; e-mail: 20236346@stu.cqu.edu.cn

Pengjun Huang[#]

School of Microelectronics and Communication Engineering, Chongqing University, Chongqing, 401331, China; e-mail: 20230582@stu.cqu.edu.cn

[#]Yihan Zeng and Pengjun Huang contributed equally to this work and they are both first authors.

Corresponding author: 20236346@stu.cqu.edu.cn (Yihan Zeng)

Accurate photovoltaic (PV) power prediction is essential for efficient PV plant operation and energy management; however, it remains challenging due to the high volatility of meteorological factors. To address this limitation of existing feature engineering methods in extracting effective features from complex nonlinear data, this paper proposes an EGWO-XGB-RF feature engineering method. The approach first employs Pearson correlation analysis to preliminarily screen meteorological features. Subsequently, XGB-RF is applied to generate new features through decision tree structures. Finally, EGWO is utilized to select a high-quality feature subset, which is subsequently used for PV power prediction with LightGBM. Experimental results show that the proposed EGWO-XGB-RF based model achieves an RMSE of 0.1875, an MAE of 0.0962, and R^2 of 0.9854 on the held-out test set, outperforming the eight compared baseline models under the same experimental conditions. The proposed method provides a reliable technical pathway for improving the accuracy of renewable energy forecasting.

KEYWORDS: Machine Learning, Grey Wolf Optimizer, Photovoltaic Power Forecasting, Feature Engineering

1. Introduction

With the accelerated low-carbon transition of the global energy structure, PV power generation, by virtue of its cleanliness and sustainability, has become a core pillar of the renewable energy system [16]. However, PV power output exhibits pronounced intermittency and volatility [11], [18] and is highly susceptible to meteorological conditions such as abrupt irradiance changes and cloud cover. Its inherent intermittency and volatility thus pose severe challenges to reliable PV plant operation and energy management. Accurate PV power forecasting is a critical prerequisite for smoothing fluctuations, optimizing dispatch, and ensuring grid security, and its research value is increasingly prominent [7]. In this context, exploring high-precision, accurate forecasting models has become an urgent scientific problem for improving PV power prediction reliability.

Current mainstream forecasting methods have shifted from physical modeling to data-driven intelligent algorithms, whose powerful nonlinear fitting capability has significantly enhanced prediction potential [2], [13], and hybrid data-driven models such as GRU–XGBoost have achieved accurate long-term solar energy hourly forecasting [29]. Specifically, PV power forecasting algorithms are predominantly based on machine learning. Zhang et al. [35] combined the XGBoost algorithm with a convolutional neural network (CNN) to improve prediction accuracy, but their model structure is complex and imposes high requirements on data quality. Banik and Biswas [4] combined RF and CatBoost into an ensemble for PV power forecasting, improving robustness and accuracy over individual tree-based models. Tercha et al. [27] compared decision tree (DT), RF, support vector machine (SVM), and XGBoost algorithms for forecasting temperature and solar irradiance for photovoltaic systems, noting that machine learning models improve prediction accuracy compared with traditional methods. In a review of data-driven algorithm applications, Gupta et al. [14] noted that deep learning algorithms perform excellently in ultra-short-term forecasting but require substantial data and computational resources. Overall, existing forecasting algorithms generally suffer from insufficient robustness under complex weather conditions and heavy dependence on dataset quality. Moreover, hyperparameter tuning is commonly neglected in these studies, making it difficult to meet the comprehensive demands for strong adaptability and reliability in PV power forecasting.

Hyperparameter tuning of machine learning models is essentially a complex optimization problem, and its performance largely depends on efficient and robust optimization algorithms. Quan et al. [23] employed the dung beetle optimization (DBO) algorithm to optimize the structure and hyperparameters of a self-attention temporal convolutional network–bidirectional long short-term memory (SATCN-BiLSTM) model for ultra-short-term PV power prediction, while Zhao et al. [36] used the whale optimization algorithm (WOA) to address data fluctuations and over-decomposition problems in ultra-short-term photovoltaic power forecasting datasets; however, both methods construct complex hybrid architectures with considerable computational overhead. Al-Dahidi et al. [1] applied the chimp optimization algorithm (ChOA) to automatically search for optimal hyperparameters, such as the number of hidden layers in a multilayer perceptron (MLP), achieving superiority over particle swarm optimization (PSO) and genetic algorithm (GA) in convergence speed and accuracy, but ChOA suffers from high computational complexity and efficiency bottlenecks. Shi et al. [25] applied a Bayesian optimization algorithm to model hyperparameter optimization, yet this method heavily relies on input data quality and preprocessing procedures. Current intelligent optimization algorithms generally exhibit weak adaptability and high complexity, making it urgent to develop novel algorithms with strong robustness and low computational cost for practical deployment in variable environments.

The choice of each component in the proposed framework is guided by the characteristics of short-term PV power forecasting. PV power is driven by nonlinear meteorological interactions that single-model feature extraction cannot fully capture; therefore, XGBoost and RF are combined in parallel, where XGBoost captures fine-grained residual patterns through sequential boosting while RF′s bagging reduces noise sensitivity and discovers diverse feature interactions. The resulting augmented fea-

ture space inevitably contains redundancy. To address this, EGWO incorporates mutual information into its fitness function, rewarding feature-target relevance while penalizing inter-feature redundancy, thereby selecting a compact and discriminative subset. LightGBM is then employed for its efficiency on the augmented features via leaf-wise growth and histogram-based computation. This synergistic design enhances feature quality at the source rather than stacking model complexity, which is a strategy well-suited to data-driven PV forecasting.

Guided by the above design considerations, this study addresses the issues of insufficient feature effectiveness, limited representation capacity of single models, and the tendency of traditional feature selection methods to become trapped in local optima in PV power forecasting. A feature engineering method, EGWO-XGB-RF, is proposed that integrates XGBoost, RF, and the enhanced grey wolf optimizer (EGWO). On the public PVOD v1.0 dataset, the proposed method achieves lower RMSE and MAE and higher R^2 than the eight compared baseline models under identical experimental conditions. The contributions of this work are as follows:

1 **Feature Generation Based on XGBoost-RF:** To overcome the problems of limited representation diversity and sensitivity to anomalous samples in feature construction using a single model XGBoost or RF, a parallel feature generation strategy combining XGBoost and RF is proposed. XGBoost's sequential boosting mechanism is adopted to capture fine-grained residual information, while RF's parallel bagging framework reduces noise and identifies non-mainstream feature interactions. The leaf node outputs from all decision trees are extracted to form a high-dimensional augmented feature matrix, which transforms raw meteorological data into a highly discriminative feature space, thereby enhancing the diversity and information capacity of the feature representations.

2 **Feature Selection using Enhanced Grey Wolf Optimizer:** To overcome the limitations of conventional feature selection methods, such as Pearson ranking and recursive feature elimination, an enhanced grey wolf optimizer (EGWO) is developed in this study. Largely reliant on linear correlation analysis, traditional methods often overlook feature redundancy, thereby yielding suboptimal feature subsets. In the proposed EGWO framework, a mutual-information-based constraint is incorporated into the fitness function. Specifically, feature–target mutual information is used to improve feature relevance, while inter-feature mutual information is employed to reduce redundant information among selected features. In addition, a random extended search strategy is introduced during the later optimization stage, where poorly performing individuals are randomly reinitialized to improve population diversity. This strategy helps alleviate premature convergence and enhances the global search capability of the standard binary grey wolf optimizer (BGWO).

3 **Systematic experimental validation:** Based on the PVOD v1.0 public dataset, the proposed EGWO-XGB-RF method is comprehensively evaluated through ablation experiments comparing ten feature selection strategies, comparative experiments against eight baseline models, and cross-station validation on nine PVOD stations. Experimental results show that EGWO-XGB-RF achieves favorable performance in terms of RMSE, MAE, and R^2, and it accurately captures power variation trends under both stable and volatile irradiance conditions. Although each component builds on established techniques, their joint configuration and systematic evaluation for short-term PV power forecasting provide a practical, reproducible reference for feature engineering in this domain.

2. Literature Review

2.1. Prediction Models

2.1.1. Single-model Prediction

With the development of data-driven algorithms, machine learning techniques have been progressively applied to PV power forecasting. Wang et al. [28] integrated satellite remote sensing imagery data with cloud observation information from adjacent power stations and employed SVM to construct a prediction model, which significantly outperformed traditional forecasting methods in terms of both RMSE and MAE. Tercha et al. [27] systematically compared DT, RF, SVM, and XGBoost for temperature and solar irradiance forecasting, and found that the relative performance of these models depends

on the target variable and data conditions. Jia et al. [17] evaluated SVM, GLMNET, and RF for daily global and diffuse solar radiation prediction in eight Chinese cities under different climate and pollution conditions, and found that SVM consistently outperformed GLMNET and RF across all locations. By comparing LightGBM with KNN, Suanpang et al. [26] verified that LightGBM not only achieves a higher R^2 but also exhibits greater stability and robustness to outliers across various time periods and seasons. The above studies employ single models for prediction; such models attract attention due to their simple structure and ease of deployment, but their performance strongly depends on the compatibility between the dataset and the model. Compared with the high stability and strong robustness exhibited by multi-model ensemble strategies, the application scope of single models is relatively limited, making it difficult to comprehensively cope with the complex and variable environment of PV power generation.

2.1.2. Multi-model Ensemble Prediction

In recent years, given that single-model prediction is constrained by its inherent limitations and struggles to adapt to complex and variable climatic conditions, research has gradually shifted toward multi-model ensemble prediction. Goliatt and Yaseen [12] proposed a hybrid framework that combines the covariance matrix adaptation evolution strategy (CMA-ES) with XGBoost and multivariate adaptive regression splines (MARS) for daily global solar radiation prediction, demonstrating the effectiveness of hybrid evolutionary models in radiation forecasting. Khan et al. [20] developed a stacked ensemble algorithm, DSE-XGB, which integrates the predictions of base models through XGBoost, significantly improving prediction accuracy and demonstrating optimal performance under various weather conditions. Ramsami and Oree [24] constructed a hybrid method combining stepwise regression and artificial neural networks (ANN) to achieve effective 24-hour ahead PV system output power prediction, further confirming the superiority of hybrid models over single models. Hou et al. [15] adopted a strategy that fuses VMD with long short-term memory (LSTM), where the features decomposed and reconstructed by VMD are fed into LSTM for training, and the results showed that the fused model significantly outperforms the single LSTM model. The above studies consistently verify that multi-model ensemble algorithms are superior to single models in terms of prediction accuracy and stability; however, since the setting of model parameters has a significant impact on prediction results, research on optimization algorithms cannot be overlooked.

2.2. Optimization Algorithms

2.2.1. Optimization Algorithms for Feature Selection and Hyperparameter Tuning

Exact algorithms aim to find provably optimal solutions by systematically exploring a finite search space with pruning, relaxation, and bounding strategies. However, their computational cost grows rapidly with problem size in high-dimensional settings, which limits their direct application to feature selection and hyperparameter optimization in PV power forecasting and motivates the increasing use of metaheuristic methods. Eskandari et al. [10] developed an interpretable energy-consumption forecasting framework combining feature selection with transparent machine learning models. Li et al. [21] proposed a hybrid Stacking model for medium-term photovoltaic power forecasting, using ranked feature selection to identify informative meteorological features for an LSTM-Informer predictor. El-kenawy et al. [8] introduced the Swordfish Movement Optimization Algorithm (SMOA) for simultaneous feature selection and hyperparameter tuning in transformer-based deep learning models for photovoltaic power forecasting. Yang and Luo [31] combined random forest screening with symplectic geometry mode decomposition and a grey wolf optimizer-tuned BiLSTM network for short-term photovoltaic power prediction, improving stability and accuracy. Ye et al. [32] constructed a two-stage photovoltaic power forecasting framework in which KNN-based anomaly cleaning, PCC-based linear correlation screening, and SHAP-based nonlinear contribution analysis are used for refined feature engineering, followed by an APO-optimized TCN-LSTM-Attention model. These studies show that metaheuristic optimization has become essential in feature selection and model tuning for PV power forecasting, trading strict optimality guarantees for much lower computational cost in high-dimensional search spaces.

2.2.2. Heuristic Algorithms

Heuristic algorithms are optimization methods based on empirical rules, probabilistic search, or approximate reasoning. Their objective is to obtain a near-optimal solution within acceptable computational time without guaranteeing global optimality. Zafar et al. [33] combined a dynamic grouping strategy with the sine cosine algorithm (SCA) to develop an improved dynamic grouping-based combination (IDGC) algorithm, which dynamically balances global exploration and local exploitation and reduces the risk of local convergence. Shi et al. [25] applied Bayesian optimization to prediction-model hyperparameter tuning, using Gaussian processes to model the objective function and an acquisition function to select subsequent parameter samples. This approach improves search efficiency and reduces repeated model training, making it suitable for computationally expensive and non-differentiable objectives. Karl et al. [19] systematically examined GA, PSO, and Bayesian optimization for multi-objective hyperparameter tuning and showed that these methods can balance prediction accuracy, computational complexity, and generalization. Danach et al. [6] proposed a Q-learning-based adaptive hyper-heuristic framework that dynamically selects search strategies through reinforcement learning and employs a simulated-annealing acceptance criterion to improve stability and convergence in high-dimensional parameter spaces. Overall, heuristic and meta-heuristic algorithms generally provide better search efficiency and stronger global exploration than conventional grid or random search. By adapting search directions according to intermediate solutions, they avoid evaluating every possible parameter combination and allocate more computations to promising regions. This advantage is particularly important for deep-learning and ensemble models with many coupled hyperparameters, where exhaustive search is computationally prohibitive. However, their performance may still depend on initialization quality, search rules, and the balance between exploration and exploitation.

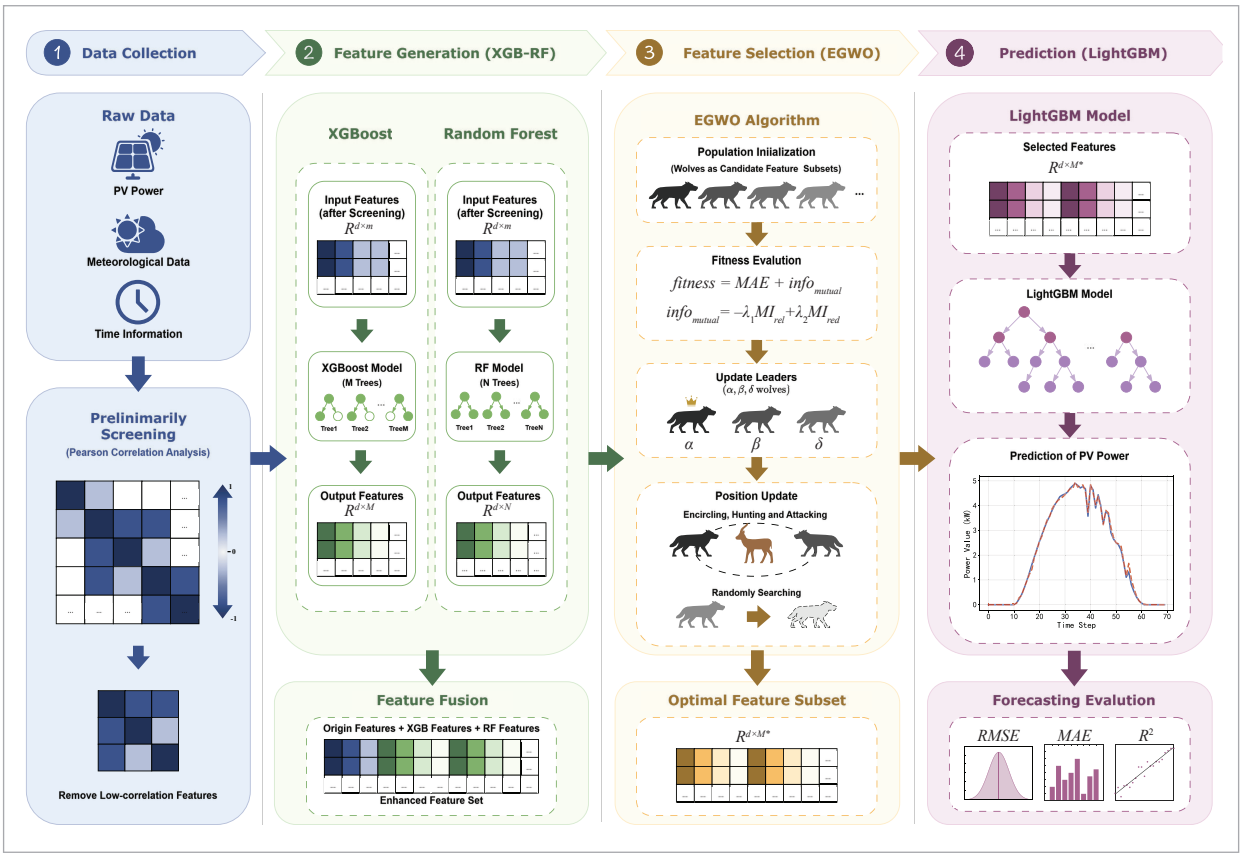
2.2.3. Swarm Intelligence Optimization Algorithms

Inspired by cooperative behaviors observed in biological populations, swarm-intelligence optimization algorithms are stochastic methods that pursue global optimization through information sharing, collaborative search, and group decision-making. Quan et al. [23] employed the DBO algorithm to optimize the structure and hyperparameters of a SATCN-BiLSTM model for ultra-short-term PV power prediction. Zhao et al. [36] used the whale optimization algorithm (WOA) to tune the key hyperparameters of variational mode decomposition (VMD), improving decomposition quality and forecasting performance by simulating prey encircling and spiral updating behaviors. Zhang et al. [34] applied the sparrow search algorithm (SSA) to forecasting-model parameter optimization and compared it with particle swarm optimization (PSO) and genetic algorithm (GA). SSA achieved faster fitness convergence and stronger global optimization capability. Zhou et al. [37] introduced the RIME optimization algorithm for forecasting-model hyperparameter tuning, achieving efficient search by simulating population growth, crossover, and environmental adaptation, which improved prediction accuracy and convergence speed. Alrashidi et al. [3] combined social spider optimization (SSO), PSO, and cuckoo search optimization (CSO) with support vector regression and ANN to improve short-term PV power forecasting. These studies indicate that swarm-intelligence algorithms can reduce the likelihood of becoming trapped in local optima because multiple search agents explore different regions in parallel and exchange information during evolution. This distributed mechanism is valuable when the objective function is nonlinear, non-differentiable, or contains multiple local optima. However, excessive exploitation may cause premature convergence, whereas excessive exploration can slow convergence and increase computational cost. These limitations are particularly important in high-dimensional feature-selection and hyperparameter-optimization problems and motivate improvements to conventional swarm-intelligence frameworks.

3. EGWO-XGB-RF

To address the problems of insufficient feature effectiveness and prediction accuracy in PV power forecasting, we propose the EGWO-XGB-RF photovoltaic power forecasting model, which consists of four major sub-modules: data collection, feature generation (XGBoost-RF), feature selection (EGWO),

Figure 1
Framework of EGWO-XGB-RF.



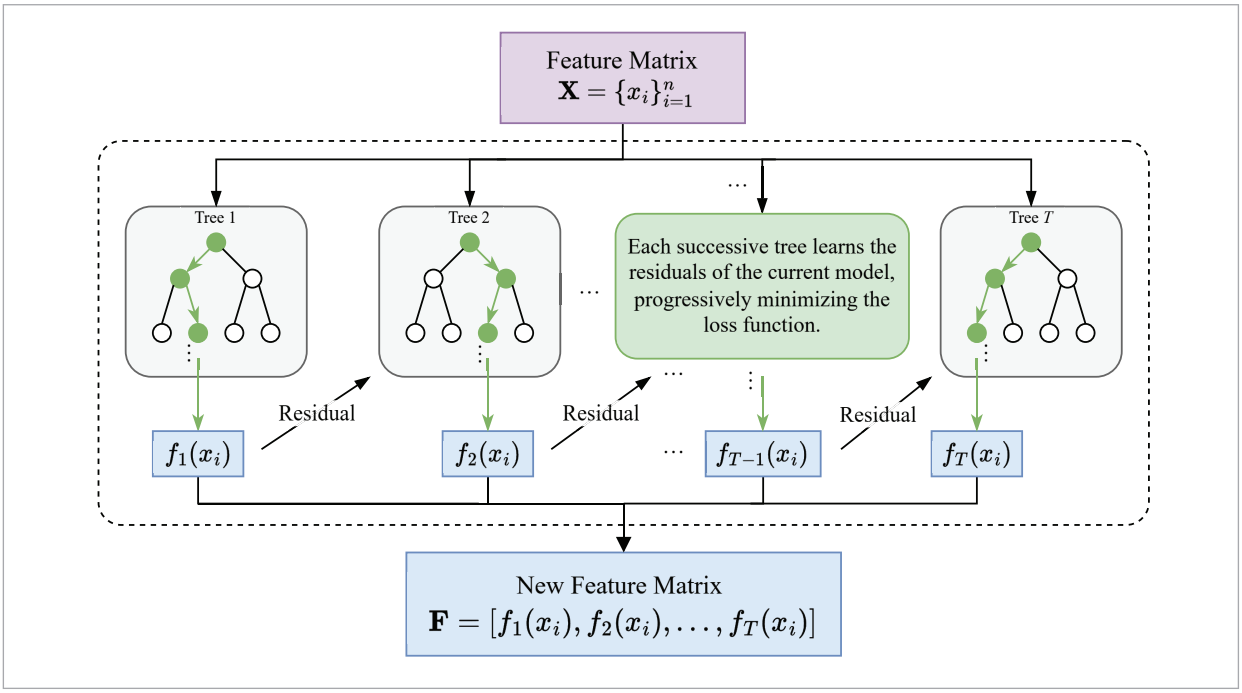
and prediction (LightGBM), as illustrated in Figure 1. The data collection module integrates raw data such as PV power, meteorological data, and time information, and pre-screens low-correlation features through Pearson correlation analysis to reduce noise. The feature generation module trains XGBoost and RF in parallel, uses the screened features to output new features and fuses them to obtain an enhanced feature set, thereby achieving the mapping from raw data to a high-dimensional feature space. The feature selection module employs EGWO, which iteratively refines candidate feature subsets using an objective function that integrates MAE, mutual information, and a redundancy penalty, ultimately producing a compact, discriminative subset that balances dimensionality and predictive performance. The prediction module, based on LightGBM, takes the selected features as inpuwt for training and outputs predicted values, with performance evaluated using RMSE, MAE, and R^2. Through multi-model feature generation for enhanced representation, efficient screening by EGWO for redundancy reduction, and rapid fitting of nonlinear relationships by LightGBM, this model achieves collaborative optimization of feature engineering and forecasting, thereby improving the accuracy and generalization of PV power forecasting.

3.1. XGBoost

XGBoost, introduced by Chen et al. in 2016, is an advanced Gradient Boosting Tree (GBT) algorithm [5]. Compared to traditional Gradient Boosting Decision Trees (GBDT), XGBoost incorporates several enhancements, including second-order Taylor expansion and regularization terms, which significantly improve both prediction accuracy and computational efficiency.

In this study, XGBoost wis employed not for direct target value regression, but rather for feature generation. Unlike conventional applications where the final

Figure 2
Schematic of XGBoost Algorithm.



output of the XGBoost model is used for prediction, we extract the output of each individual tree as new features to be used in subsequent learning stages, as illustrated in Figure 2. The objective of this approach is to leverage the structural information embedded within the ensemble of trees to generate more expressive and informative features, thereby enhancing the performance of the downstream prediction model.

Assuming the t^{th} tree contains L_t leaf nodes, the prediction value of Tree t for a given input sample x_i is denoted as $f_t(x_i)$, and can be formulated as follows:

$$f_t(x_i) = \sum_{j=1}^{L_t} w_j^{(t)} \times I(x_i \in R_j^{(t)}) + f_{t-1}(x_i), \quad (1)$$

where $w_j^{(t)}$ is the weight of the j^{th} leaf node in the t^{th} tree, $R_j^{(t)}$ is the subset of samples that fall into this leaf node. The indicator function $I(x_i \in R_j^{(t)})$ returns 1 if the input sample x_i belongs to node j, and 0 otherwise.

By extracting the output value of each individual tree for every sample, we construct a new feature matrix F, where the i^{th} row contains the prediction values of sample x_i across all decision trees in the XGBoost model:

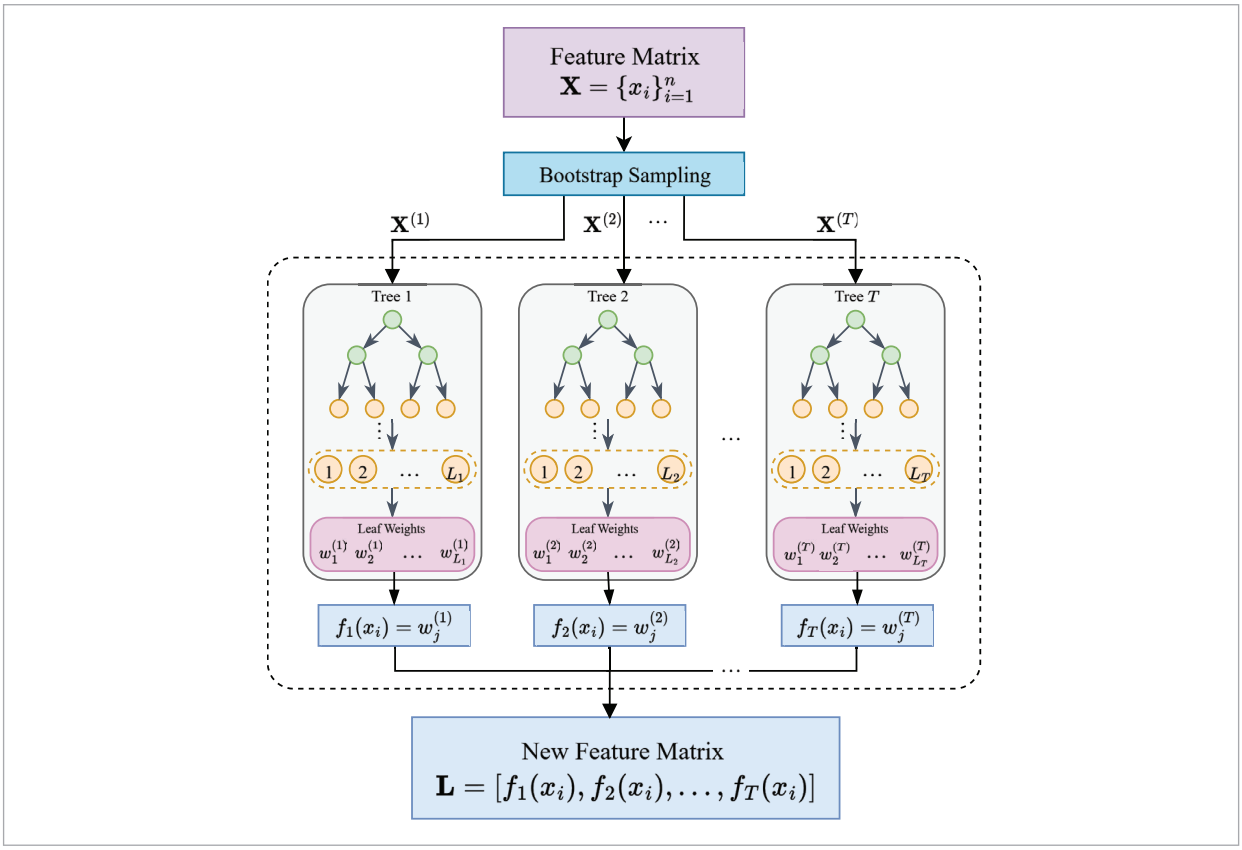
$$F = \left[f_1(x_i), f_2(x_i), f_3(x_i), \cdots, f_T(x_i)\right], \quad (2)$$

where T is the total number of trees in the XGBoost model, and each $f_t(x_i)$ represents the output of the t^{th} tree for the sample x_i. This transformed matrix F is then used as part of the extended feature set for subsequent modeling and learning tasks.

3.2. Random Forest

The leaf node generated by XGBoost through its boosting mechanism can finely characterize residual information, offering strong fitting capability for complex nonlinear relationships. Because XGBoost trains trees sequentially to fit the residuals of the preceding ensemble, its generated tree features may be sensitive to anomalous samples: when raw meteorological data contain noisy outliers, subsequent trees will continually split around these noise points to reduce residuals, causing the generated leaf node encodings to overly rely on a small number of anomalous patterns, thereby weakening the model's generalization ability on unseen data. Due to the greedy splitting strategy inherent in the boosting frame-

Figure 3
Schematic of Random Forest Algorithm.



work, decision trees repeatedly partition around a small set of strongly correlated features, which in turn induces high correlations among the outputs of different trees within the feature matrix constructed by Eq. (2). As the number of trees increases, the expansion of feature dimensionality does not bring a proportional gain in information, and the diversity of the feature space is limited.

To address these limitations, this paper further introduces random forest (RF) on top of XGBoost to construct a complementary feature generation strategy. RF adopts a parallel bagging framework, where individual decision trees are trained independently on different bootstrap sample subsets and random feature subspaces. On one hand, this mechanism ensures that anomalous noise points affect only the growth of a subset of trees, and their extreme outputs can be effectively smoothed by the parallel results of multiple trees. On the other hand, the introduction of random subspaces enables RF to capture, with higher probability, non-mainstream feature interaction combinations that are missed by XGBoost's greedy search, thereby enriching the diversity of feature representations. Similar to the treatment in Section 3.1, this paper does not directly use the ensemble prediction of RF but instead extracts the leaf node output of each decision tree as a new feature representation, as illustrated in Figure 3. The specific construction procedure is as follows.

Assuming the t^{th} tree contains L_t leaf nodes, the prediction value $w_j^{(t)}$ of the leaf node j is defined as the mean of the target values of all samples falling into that node as follows:

$$w_j^{(t)} = \frac{1}{\left|R_j^{(t)}\right|} \sum_{i \in R_j^{(t)}} y_i \,, \tag{3}$$

where y_i is the target value of the sample contained in the leaf node j, $R_j^{(t)}$ is the set of samples assigned to the leaf node j of the t^{th} tree.

Accordingly, for a given sample x_i, the prediction value $f_t(x_i)$ from the t^{th} tree is defined as the weight (i.e., mean target value) of the leaf node j into which the sample falls:

$$f_t(x_i) = w_j^{(t)}, x_i \in R_j^{(t)}, \tag{4}$$

where x_i is the sample contained in the leaf node j, $w_j^{(t)}$ is the weight of the leaf node j.

By extracting the prediction value from each individual tree, we construct a new feature matrix$L$, where each column represents the prediction of a single tree for sample x_i. The matrix is formulated as:

$$L = \left[f_1(x_i), f_2(x_i), f_3(x_i), \cdots, f_T(x_i)\right], \tag{5}$$

where L is the new feature matrix generated by Random Forest, and $f_t(x_i)$ is the prediction value of the t^{th} tree for the sample x_i.

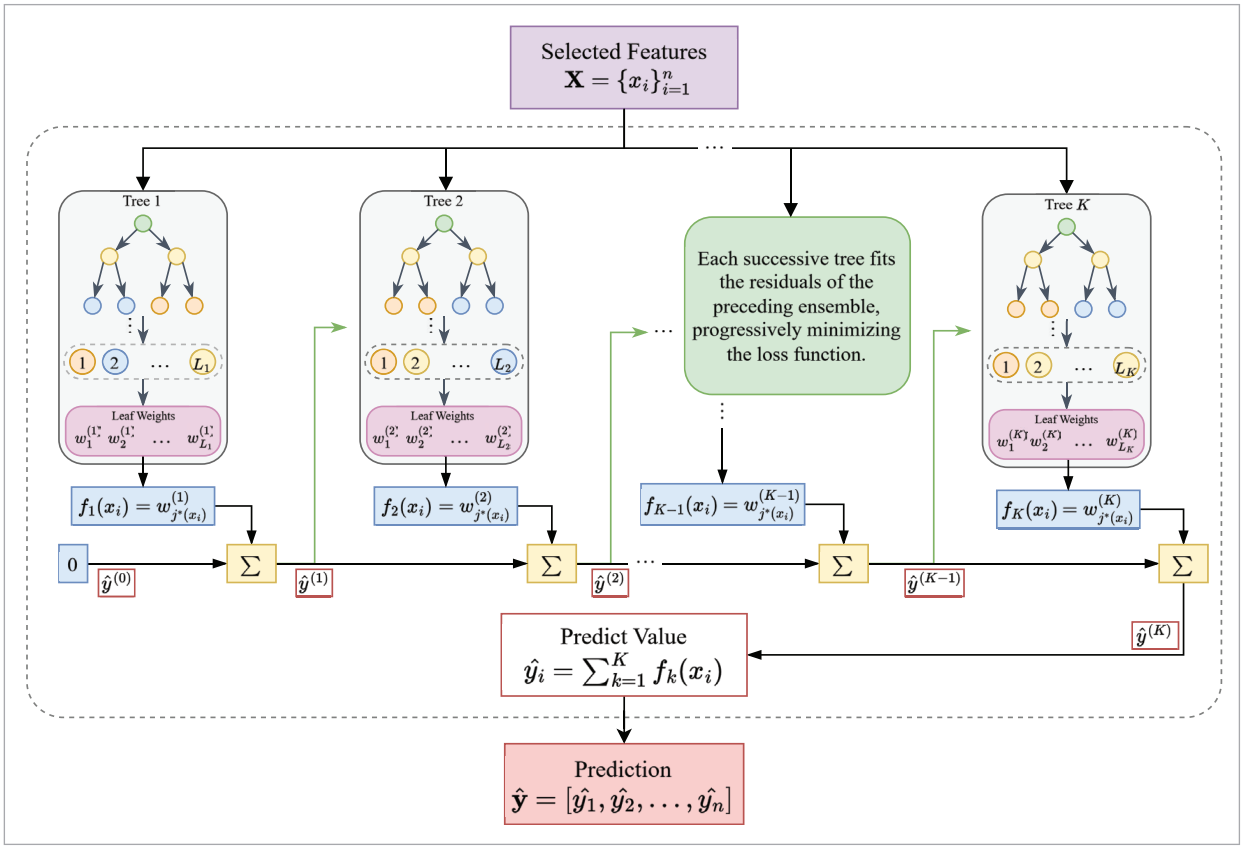
3.3. LightGBM

Light Gradient Boosting Machine (LightGBM) is an efficient gradient boosting decision tree framework. Building on the traditional GBDT, this algorithm incorporates Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) techniques, combined with a histogram algorithm and a leaf-wise growth strategy with depth limitation. It significantly reduces computational overhead while ensuring prediction accuracy, making it particularly suitable for regression and classification tasks involving high-dimensional, large-scale data.

As shown in Figure 4, LightGBM is employed in this study for the final regression prediction task. The model belongs to the typical additive ensemble model. Suppose the model consists of K decision trees. For the k-th tree, each leaf node j corresponds to a weight $w_j^{(k)}$. This weight is obtained by approximately solving the regularized objective function using

Figure 4

Schematic of LightGBM Algorithm.



a second-order Taylor expansion, which can be expressed as:

$$w_j^{(k)} = -\frac{\sum_{i\in D_j^{(k)}} g_i}{\sum_{i\in D_j^{(k)}} h_i+\lambda}, \tag{6}$$

where $D_j^{(k)}$ denotes the set of samples falling into leaf node j of the k-th tree; λ is the regularization coefficient used to suppress overfitting caused by excessively large leaf weights; g_i and h_i are respectively the first order and second order derivatives of the loss function $l(y_i,\hat{y})$ with respect to the ensemble prediction $\hat{y}^{k-1}$ of the preceding k–1 trees, respectively, defined as:

$$g_i = \frac{\partial l(y_i,\hat{y})}{\partial \hat{y}^{k-1}}, h_i = \frac{\partial^2 l(y_i,\hat{y})}{\left(\partial \hat{y}^{k-1}\right)^2}. \tag{7}$$

For an input sample x_i, the output of the k-th tree, denoted as $f_k(x_i)$, is defined as the weight of the leaf node into which the sample falls:

$$f_k(x_i) = w_{j^*x_i}^{(k)}, \tag{8}$$

where j^*x_i is the index of the leaf node that sample x_i falls into in the k-th tree. During training, LightGBM adopts a sequential additive strategy: each new tree fits the residuals based on the ensemble predictions of all preceding trees, thereby progressively reducing the overall loss function value. Finally, the outputs of all K trees are accumulated to obtain the final model prediction:

$$\hat{y}(x_i) = \sum_{k=1}^{K} f_k(x_i). \tag{9}$$

3.4. Enhanced Grey Wolf Optimizer

3.4.1. Basic GWO

The Grey Wolf Optimizer (GWO) is a swarm intelligence optimization algorithm proposed by Mirjalili et al. in 2014 [22]. The algorithm classifies the population individuals into four hierarchical levels—α, β, δ, and ω—and the three leading wolves (α, β, and δ) guide the population search. A dynamic balance between global exploration and local exploitation is achieved through the linear decay of the control parameter a. For the i-th wolf, its position vector is represented as a continuous variable:

$$p_i = \left[p_{i1}, p_{i2}, p_{i3}, \cdots, p_{in_{feature}}\right]. \tag{10}$$

The position of each individual is updated under the guidance of the three leading wolves according to:

$$\begin{cases} X_1 = p_\alpha - A_1 \times D_\alpha \\ X_2 = p_\beta - A_2 \times D_\beta \\ X_3 = p_\delta - A_3 \times D_\delta \\ p_i^{(t+1)} = (X_1 + X_2 + X_3)/3 \end{cases}, \tag{11}$$

where A and D are coefficient vectors used to regulate the search step size. However, th e conventional GWO is designed for continuous optimization problems, whereas feature selection is inherently a discrete combinatorial optimization problem. Therefore, directly applying continuous position vectors cannot produce stable feature subset selection results.

To address this issue, Eman Emary et al. [9] proposed the Binary Grey Wolf Optimizer (BGWO), which introduces a Sigmoid probability mapping mechanism to transform continuous search spaces into binary feature selection representations. For a continuous position vector element p_{ij}, it is converted into a probability value:

$$S(p_{ij}) = \frac{1}{1+e^{-p_{ij}}}. \tag{12}$$

Subsequently, a random number $r_{ij}\in[0,1]$ is generated to complete the binary mapping:

$$b_{ij} = \begin{cases} 0, r_{ij} \in [0,1] < S(p_{ij}) \\ 1, r_{ij} \in [0,1] \geq S(p_{ij}) \end{cases}. \tag{13}$$

Thus, the binary feature selection vector can be obtained as:

$$b_i = \left[b_{i1}, b_{i2}, b_{i3}, \cdots, b_{in_{feature}}\right]. \tag{14}$$

Although BGWO effectively realizes feature selection in discrete domains through the introduction of probabilistic encoding mechanisms, its fitness function still suffers from inherent limitations. Specifically, existing evaluation criteria focus solely on minimizing model prediction error while ignoring the intrinsic correlation structure among variables

in the feature space. Such a single-objective optimization strategy lacks both an incentive mechanism for feature relevance and a penalty mechanism for feature redundancy. Consequently, during the iterative optimization process, the algorithm tends to retain feature subsets that are highly correlated with the target variable but also strongly collinear with one another, thereby introducing substantial redundant information and reducing the generalization capability and interpretability of the model.

Moreover, when viewed through the lens of swarm evolutionary dynamics, the BGWO algorithm experiences a substantial decline in population diversity as the number of iterations increases. Owing to the strong guidance imposed by the grey wolf hierarchy, the population rapidly converges toward the current optimal solution region during the later stages of evolution, resulting in premature convergence of the search behavior. When dealing with high-dimensional and complex feature spaces, this mechanism restricts the algorithm's ability to discover more optimal sparse feature combinations.

3.4.2. Improvement Strategy

To address the above issues, this paper proposes an Enhanced Grey Wolf Optimizer (EGWO), which introduces a mutual information constraint mechanism and a random extended search strategy into the standard grey wolf optimizer framework to enhance the global search capability and subset quality during the feature selection process, as illustrated in Figure 5 and Table 1.

1 Mutual Information Constraint Mechanism

To overcome the limitation that the fitness function of BGWO focuses solely on prediction error while neglecting the correlations among features, this paper introduces a mutual information constraint mechanism into the fitness function to achieve the collaborative optimization of feature relevance enhancement and redundancy suppression. Since the optimization goal is to minimize the fitness value, the improved fitness function is defined as:

$$fitness(b_i) = MAE - \lambda_1 MI_{rel}(b_i) + \lambda_2 MI_{red}(b_i) . \quad (15)$$

In this formulation, MAE is the mean absolute error of the prediction model trained on the selected feature subset, serving as the primary term to be minimized. MI_{rel} denotes the average mutual information

Figure 5
Enhanced Grey Wolf Optimizer Algorithm.

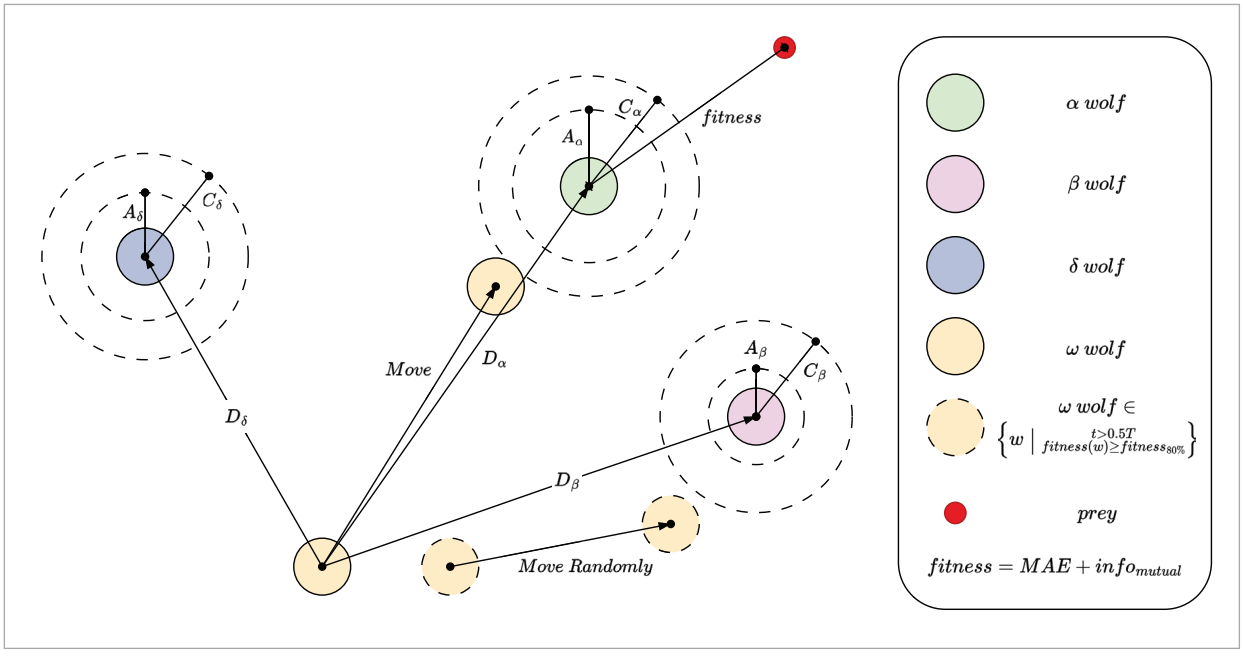


Table 1
Enhanced Grey Wolf Optimizer.

Input: $x_{leaf} \in R^{N \times M}$, leaf weight matrix; $y \in R^{N}$, target value; P, population size; T, maximum number of iterations; K, target number of selected features.

Output: S^*, final feature index set; f^*, corresponding best fitness found.

Randomly initialize continuous wolf positions $x_i \in R^M (i=1,2,3,...P)$, $x_i^j \sim U(-5,5)$
Randomly initialize leading wolves α, β, δ, and set $f_\alpha = +\infty$, $f_\beta = +\infty$, $f_\delta = +\infty$

for $t = 1$ to T do
  for $i = 1$ to P do
  $b_i = BinaryConversion(x_i, K)$ # *convert continuous solution to binary*
  $S_i = \{j | b_i = 1\}$
  if $|S_i| < 2$ then # *avoid selecting too few features*
    $f_i = +\infty$
  else
    Train LightGBM with $x_{leaf}[:, S_i]$ using K-fold cross-validation
    $f_i = mean(MAE - \lambda_1 MI_{rel} + \lambda_2 MI_{red})$ # *fitness with mutual info and redundancy penalty*
  end if
  Update α, β, δ according to ascending f_i
 end for
 $a = 2 - 2t/T$
 for $i = 1$ to P do
  for $j = 1$ to M do
    Compute X_1, X_2, X_3 according to GWO equations
    $x_i^j = \frac{(X_1 + X_2 + X_3)}{3}$
    $\omega \in \left\{ w \middle| \frac{t > 0.5T}{fitness(w) \geq fitness_{80\%}} \right\}$ # *random extended search strategy*
  end for
 end for
end for
$b^* = BinaryConversion(x_\alpha, K)$
$S^* = \{j | b^{*j} = 1\}$
$f^* = f_\alpha$
return S^*, f^*

between the selected features and the target variable, defined as:

$$MI_{rel}(b_i) = \frac{1}{|S|} \sum_{f_j \in S} I(f_j; y), \quad (16)$$

where S denotes the set of features selected by the subset b_i, and $I(f_j;y)$ denotes the mutual information between feature f_j and the target variable y. The negative sign $(-\lambda_1 MI_{rel})$ reduces the fitness value when feature-target correlation is high, thereby rewarding discriminative features.

The degree of redundancy among the selected features is measured by the feature redundancy term $MI_{red}(b_i)$, which is computed as the average mutual information across all pairwise combinations of the selected features:

$$MI_{red}(b_i) = \frac{1}{|S|(|S|-1)} \sum_{f_m, f_n \in S} I(f_m; f_n), \quad (17)$$

where S denotes the set of features selected by the subset b_i, $|S|$ is the number of selected features, and $I(f_m;f_n)$ denotes the mutual information between feature f_m and feature f_n. The positive sign ($+\lambda_2 MI_{red}$) increases the fitness value when inter-feature redundancy is high, thereby penalizing redundant feature pairs.

λ_1 and λ_2 are weight coefficients used to balance the influence of prediction error, feature relevance, and the redundancy penalty. A larger λ_1 strengthens the correlation between the features and the target variable, while a larger λ_2 suppresses feature redundancy, thereby enabling EGWO to obtain a feature subset with low redundancy and high discriminative ability while ensuring prediction accuracy.

2 Random Extended Search Strategy

To counteract the significant decline in population diversity observed in the later stages of the BGWO algorithm, this paper applies a random extended search to individuals with poor fitness during the latter 50% of iterations. The set of wolves that require random search is defined as:

$$\omega \in \left\{ w \middle| \begin{matrix} t>0.5T \\ fitness(w) \geq fitness_{80\%} \end{matrix} \right\}, \quad (18)$$

where T is the maximum number of iterations, $fitness_{80\%}$ denotes the 80th percentile of the population fitness, and ω represents the set of wolves that need to undergo random search. For the individuals in this set, the random extended search strategy is performed as:

$$p_i^{(t+1)} = random\left(p_i^{(t)}\right), i \in \omega, \quad (19)$$

where *random* indicates the random reinitialization of the individual's position within the search space. By reinitializing inferior individuals, this strategy expands the search space, enhances population diversity, and thereby strengthens the algorithm's ability to escape local optima and improves the convergence performance in the later stages.

Finally, the final feature subset is derived from the position of the α wolf and obtained through binary conversion:

$$b_i = binary_Conversion(\alpha_{position}) \quad (20)$$

$$Selected_features = \{j | b_i = 1\}. \quad (21)$$

4. Case Analysis

4.1. Experimental Environment Description

4.1.1. Dataset

This study uses the publicly available Science Data Bank dataset PVOD v1.0 [30], which covers operational data from 10 photovoltaic power plants. The data from station00 is used as a case study for testing and validation. The time span of the station00 dataset is from 16:00:00 on August 15, 2018, to 15:45:00 on June 13, 2019, containing a total of 28,896 data records with a sampling interval of 15 minutes. All records, including night-time intervals with zero PV power, are retained in both training and evaluation to reflect real operating conditions. The specific features are shown in Table 2.

Table 2
Feature Type and Feature Name.

Feature Type	Feature Name
Numerical Weather Prediction	nwp_globalirrad
	nwp_directirrad
	nwp_temperature
	nwp_humidity
	nwp_windspeed
	nwp_winddirection
	nwp_pressure
Local Measured Data	lmd_totalirrad lmd_diffuseirrad
	lmd_temperature
	lmd_pressure
	lmd_winddirection
	lmd_windspeed

In this study, the feature vector x_t, which is available at forecast issuance time t and consists of numerical weather predictions for the target interval, local measurements observed up to time t, and time encoded features, is used to predict the PV power y_{t+1} of the next 15 minute interval. The horizon adopted here is exactly one sampling step.

To ensure the scientific rigor of the experiment and the generalization capability of the model, the dataset is divided chronologically into a training set (first 70%), a validation set (next 15%), and a testing set (final 15%). The validation set is used for hyperparameter tuning and feature selection, while the testing set is held out exclusively for final evaluation.

For all models in this study, the forecasting task is defined as predicting the photovoltaic power for the next 15 minute interval. Feature screening, feature generation, feature selection, and hyperparameter optimization were performed exclusively on the training set. The validation set was used for early stopping and model selection, while the testing set was held out throughout these procedures and used only for final evaluation.

Table 3

Bayesian Hyperparameter Optimization Results.

Model	Parameters	Value
XGBoost	learning_rate	0.28611864
	max_depth	7
	n_estimators	74
	gamma	0.02457175
RF	n_estimators	55
	max_depth	7
	min_samples_split	58
LightGBM	learning_rate	0.02072677
	n_estimators	421
	min_split_gain	0.00128949
EGWO	n_wolves	12
	target_trees	15
	λ_1	0.0674
	λ_2	0.0943

4.1.2. Model Parameter Settings

To enhance prediction performance, hyperparameter optimization was performed on the XGBoost, Random Forest, and LightGBM models. The optimization process employed the Bayesian Optimization (BO) algorithm. The obtained optimal parameters and EGWO parameters are presented in Table 3.

4.1.3. Experimental Environment Configuration

Table 4

Simulation Environment Configuration.

Item	Configuration
Programming Language	Python 3.12.0
Integrated Development Environment	Visual Studio Code
CPU	Intel Core i9-13980HX
GPU	NVIDIA RTX 4060 Laptop
RAM	16 GB
Operating System	Windows 11

4.1.4. Evaluation Indicators

To compare the performance of different models and validate the effectiveness of the proposed EGWO-XGB-RF based feature engineering prediction model, this study introduces three evaluation metrics: MAE (Mean Absolute Error), RMSE (Root Mean Square Error), and R^2. The mean absolute deviation between the predictions and actual power values is captured by MAE, the standard deviation of these prediction errors is measured by RMSE, and the goodness-of-fit of the constructed model is reflected by R^2.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y} - y_i| \tag{22}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y} - y_i)^2} \tag{23}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}-y_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}, \tag{24}$$

where y_i, $\hat{y}$ and $\bar{y}$ represent the actual value, predicted value, and mean value of PV power generation, re-

spectively. All evaluation metrics are computed over the full held-out test set, including night-time intervals, so that the reported accuracy reflects the model's performance under real operating conditions.

4.2. Feature Engineering

4.2.1. Feature Screening

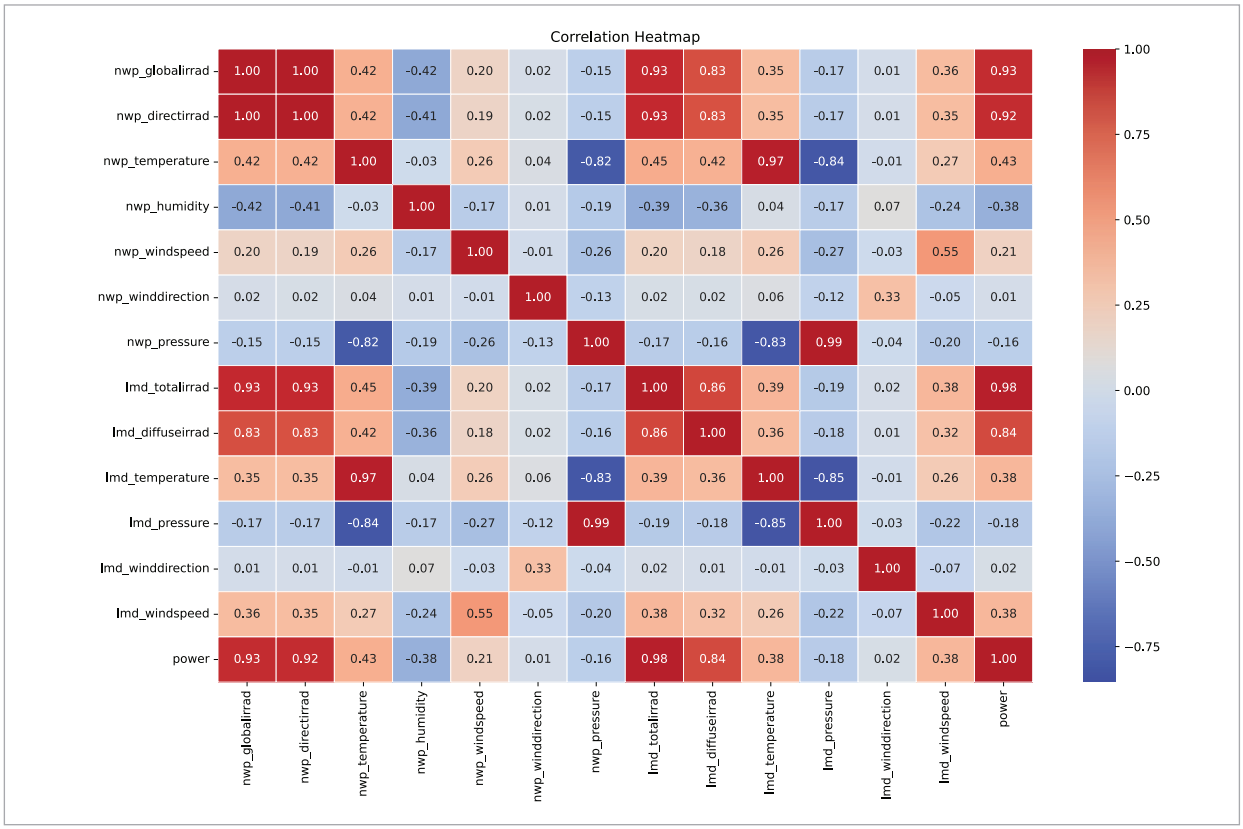
The PVOD v1.0 dataset provides 13 meteorological features, as listed in Table 2. Figure 6 presents the Pearson correlation matrix among these features and PV power. A strong correlation is observed among the four irradiance-related features (nwp_globalirrad, nwp_directirrad, lmd_totalirrad, lmd_diffuseirrad), all of which were retained due to their direct physical relevance to PV power output. Two temperature features (nwp_temperature, lmd_temperature) were also retained based on their moderate correlation with PV power and their known influence on panel efficiency. The two wind direction variables exhibit negligible correlation with PV power and were excluded. The remaining variables (humidity, pressure, wind speed) were omitted to maintain a compact feature set and to avoid station-specific dependencies, which facilitates comparison with other methods and reproduction across different PV stations. In addition, five time-derived features (hour_sin, hour_cos, month_sin, month_cos, and a binary daylight indicator) were constructed to capture diurnal and seasonal cycles. This time encoding scheme is widely adopted in PV forecasting literature, and employing the same scheme ensures consistent temporal representation across all compared models. The resulting 11-dimensional input vector is used uniformly across all models in this study.

4.2.2. EGWO-XGB-RF Feature Engineering

This study employs XGBoost to generate tree-based features from the original inputs. For each sample, the leaf-node weight produced by each individual XGBoost tree is extracted as an independent feature. The

Figure 6

Pearson correlation coefficient matrix of features.



outputs are not accumulated across trees. Similarly, for RF, the leaf-node prediction produced by each independently trained tree is extracted as an individual feature. In contrast, due to structural differences, the RF directly uses the individual prediction results of each tree without the need for accumulation.

Based on this, the Pearson correlation coefficients between the feature trees generated by XGBoost and RF and the actual power generation are calculated, and the coefficients are ranked accordingly. The top 10 trees with the highest correlation in XGBoost and the top 10 trees in RF are selected as the feature sources, with the selection ratio and the difference in the number of feature trees generated by the two specified models. This selection strategy effectively reduces dimensionality during the feature generation phase while retaining feature information that is relevant to the target variable, which helps improve computational efficiency and mitigate the risk of overfitting.

Table 5
Selected Feature Number and Pearson Correlation coefficient.

Feature Number	Feature Name	Pearson Correlation Coefficient
1	RF Tree 1	0.9800
2	RF Tree 2	0.9794
4	RF Tree 4	0.9791
7	RF Tree 7	0.9765
9	RF Tree 9	0.9751
12	XGB Tree 2	0.9801
13	XGB Tree 3	0.9796
15	XGB Tree 5	0.9794
16	XGB Tree 6	0.9792
20	XGB Tree 10	0.9763
23	nwp_temperature	0.3960
24	lmd_totalirrad	0.9834
27	hour_sin	0.6597
30	month_cos	0.1621
31	daylight	0.3019

Subsequently, the selected feature trees are concatenated with the original meteorological features to form the merged feature set, which is then subjected to feature selection using the Enhanced Grey Wolf Optimizer (EGWO) algorithm. Among the 31 features, a 15-dimensional feature subset is selected by EGWO. MAE, augmented with a mutual information term and a redundancy penalty, is adopted as the fitness function, and the feature subset with the lowest fitness value found during the search is selected as the final input. The selected features and corresponding Pearson correlation coefficients are presented in Table 5.

4.3. Results and Analysis

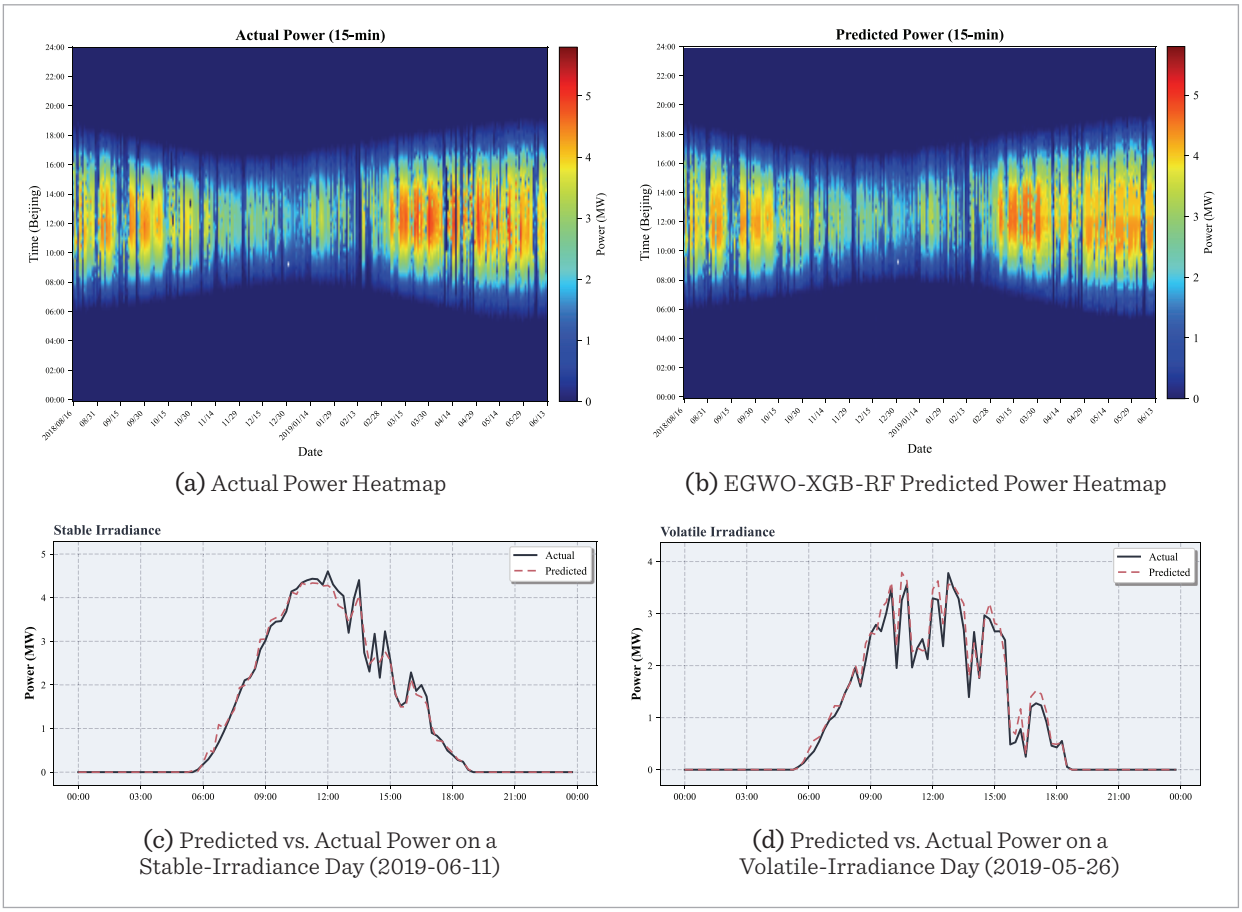
4.3.1. EGWO-XGB-RF Prediction Result

Figure 7 presents the prediction results of the EGWO-XGB-RF model over the full dataset (28,896 records). Figure 7(a) and (b) show heatmaps of the actual and predicted PV power, respectively, where each column corresponds to one day and each row to a 15-minute interval. The predicted heatmap closely reproduces the diurnal activation pattern and seasonal variations evident in the actual heatmap, indicating that the model captures both the daily solar cycle and month-to-month irradiance trends without systematic distortion.

To further examine prediction behavior under different irradiance conditions, two representative days were selected from the test set based on the daily mean and standard deviation of global horizontal irradiance (GHI). A day with clear-sky irradiance and low short-term variability (2019-06-11) was chosen to represent stable irradiance conditions, and a day with elevated GHI variability (2019-05-26) was chosen to represent periods of significant irradiance fluctuation. The predicted and actual power curves for these two days are plotted in Figure 7(c) and (d), respectively.

On the stable-irradiance day (2019-06-11), the predicted curve closely tracks the actual power throughout the daylight hours, with only minor deviations near the morning ramp-up and evening ramp-down transitions. On the volatile-irradiance day (2019-05-26), the actual power exhibits sharp drops and rapid recoveries consistent with intermittent cloud passage. The predicted curve follows these fluctuations with reasonable fidelity, capturing the timing

Figure 7
EGWO-XGB-RF Prediction Result.



(a) Actual Power Heatmap

(b) EGWO-XGB-RF Predicted Power Heatmap

(c) Predicted vs. Actual Power on a Stable-Irradiance Day (2019-06-11)

(d) Predicted vs. Actual Power on a Volatile-Irradiance Day (2019-05-26)

and magnitude of the major power dips, although short-duration extremes are partially smoothed. These results indicate that the EGWO-XGB-RF model can provide reliable predictions under both stable and variable irradiance conditions.

4.3.2. Ablation Experiment

To evaluate the contribution of each component in the proposed pipeline, ablation experiments were conducted comparing 10 strategies. RF leaf weights and XGBoost leaf weights serve as single-source baselines, using only features generated by their respective models. The remaining eight strategies select features from a common 31-dimensional candidate pool consisting of the top-10 RF leaf weights, the top-10 XGBoost incremental leaves, and the 11 original meteorological and time features. All strategies use LightGBM as the final predictor with identical hyperparameter optimization and follow the same chronological split described in Section 4.1.1.

1. **XGBoost:** Only the top-10 XGBoost leaf-weight features, selected by their Pearson correlation with PV power, are input to LightGBM.
2. **Random Forest:** Only the top-10 RF leaf weight features, selected by their Pearson correlation with PV power, are input to LightGBM.
3. **Polynomial Features:** The features are expanded into pairwise interaction and quadratic terms through second-degree polynomial transformation, from which the top 15 features most correlated with PV power are selected for LightGBM prediction.
4. **Pearson-XGB-RF:** The top 15 features are selected as model input based on the Pearson correlation coefficient between each feature and photovoltaic power generation, ranked in descending order.

5 **RFE-XGB-RF:** Linear regression is used as the base learner, and Recursive Feature Elimination is applied to progressively eliminate redundant features, ultimately retaining 15 features.

6 **Voting-XGB-RF:** Logistic regression and Lasso regression are used to model the feature set, with each feature receiving a voting score. The top 15 features with the highest scores are then selected.

7 **BGWO-XGB-RF:** Based on global search optimization in the discrete feature space of BGWO, the 15-feature subset with the best predictive performance is ultimately selected and used as inputs for LightGBM.

8 **BGWO with mutual information:** Based on global search optimization in the discrete feature space of BGWO with mutual information only, the 15-feature subset with the best predictive performance is ultimately selected and used as inputs for LightGBM.

9 **BGWO with random extended search:** Based on global search optimization in the discrete feature space of BGWO with random extended search only, the 15-feature subset with the best predictive performance is ultimately selected and used as inputs for LightGBM.

10 **EGWO-XGB-RF:** Based on global search optimization in the discrete feature space of EGWO, the 15-feature subset with the best predictive performance is ultimately selected and used as inputs for LightGBM.

Quantitative results are summarized in Table 6. Among the single-source methods, XGBoost leaf weights (R^2 = 0.9787) outperform RF leaf weights (R^2 = 0.9773) and polynomial features (R^2 = 0.9717), confirming that boosting-based tree features carry richer predictive information than bagging-based features or simple polynomial expansions under chronological evaluation. The filter and wrapper methods (Pearson, RFE, Voting) achieve R^2 values of 0.9752, 0.9794, and 0.9807 respectively, with Voting outperforming the others due to its dual-perspective ranking mechanism. BGWO (R^2 = 0.9818) surpasses all filter and wrapper methods, demonstrating the advantage of global search over the discrete feature space.

The progressive improvements brought by the two EGWO components are clearly visible. Adding the mutual information constraint alone raises R^2 from 0.9818 to 0.9831, as the feature-target relevance reward guides selection toward more discriminative features. Adding the random extended search alone raises R^2 to 0.9828, indicating that reinitializing poorly performing individuals alleviates premature convergence. The full EGWO combines both mechanisms and achieves the best overall performance, reducing RMSE by 7.0% and improving R^2 by 0.25% compared with standard BGWO. These results confirm that the mutual information constraint and the random extended search strategy both contribute to feature subset quality, and their combination yields the greatest benefit.

Table 6
Performance Comparison of Feature Selection Strategies.

Method	RMSE	MAE	R^2	Time
Polynomial Features	0.2620	0.1460	0.9717	47 s
XGBoost	0.2273	0.1187	0.9787	23 s
Random Forest	0.2347	0.1219	0.9773	28 s
Pearson-XGB-RF	0.2454	0.1280	0.9752	36 s
RFE-XGB-RF	0.2233	0.1163	0.9794	50 s
Voting-XGB-RF	0.2164	0.1170	0.9807	40 s
BGWO-XGB-RF	0.2098	0.1125	0.9818	58 s
BGWO with mutual information	0.2023	0.1070	0.9831	92 s
BGWO with random extended search	0.2041	0.1105	0.9828	59 s
EGWO-XGB-RF	0.1951	0.1015	0.9843	97 s

Table 6 also reports the total runtime of each strategy. The ten methods fall into three tiers. Methods based on deterministic ranking (RF, XGBoost, Pearson, RFE, Voting) complete within 23–50 s, dominated entirely by LightGBM hyperparameter optimization. Polynomial features require 47 s due to the degree-2 expansion. GWO variants without the mutual information term run in 58–59 s, with only a modest increase from the meta-heuristic search. Methods incorporating mutual information require 92–97 s, driven by MI matrix precomputation and repeated lookup during fitness evaluation.

It is worth noting that BGWO with random extended search achieves an R^2 of 0.9828 in 59 s, ap-

proaching EGWO (97 s, R^2 = 0.9843) at 60.8% of the runtime. Nevertheless, EGWO yields the highest accuracy across all metrics, and the additional cost is acceptable for offline model training. The design of EGWO prioritizes prediction accuracy over minimal runtime, which is appropriate for PV forecasting where forecast quality directly impacts operational decisions.

As evidenced in Figure 8, the progressive improvement in goodness-of-fit from basic strategies to the full EGWO method is clearly visible, consistent with the quantitative trends reported in Table 6.

Figure 8 presents the scatter plots of predicted versus actual PV power for each feature selection strategy. The polynomial expansion method exhibits the most pronounced upward bias and the widest dispersion around the diagonal, reflecting its limited ability to extract meaningful features from the candidate pool. RF, XGBoost, Pearson, RFE, and Voting progressively improve the fit but retain noticeable errors in the medium-to-high power range.

The BGWO variants produce tighter distributions than the conventional filter and wrapper methods. Adding mutual information improves alignment around the diagonal, while random extended search further reduces the residual spread. The full EGWO method achieves the tightest clustering around the diagonal overall, with minimal systematic bias and the smallest residual spread, consistent with its highest R^2 value reported in the subplot.

Figure 9 evaluates the sensitivity of EGWO to the iteration and fitness thresholds controlling random reinitialization. The random extended search strategy is governed by the iteration threshold and the fitness threshold. To reduce the computational cost of the 25-parameter grid, LightGBM was con-

Figure 8

Comparison of model fitting performance via scatter distribution plots.

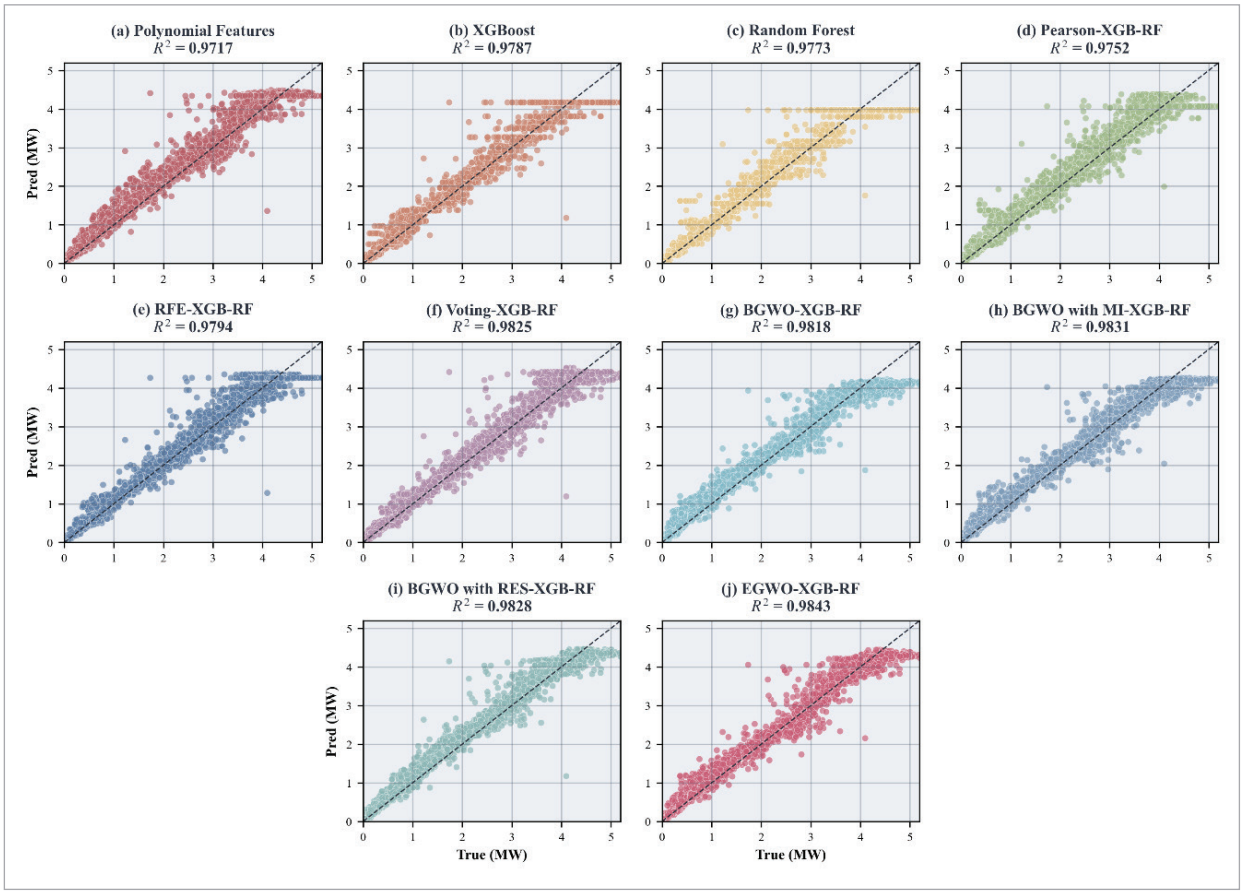
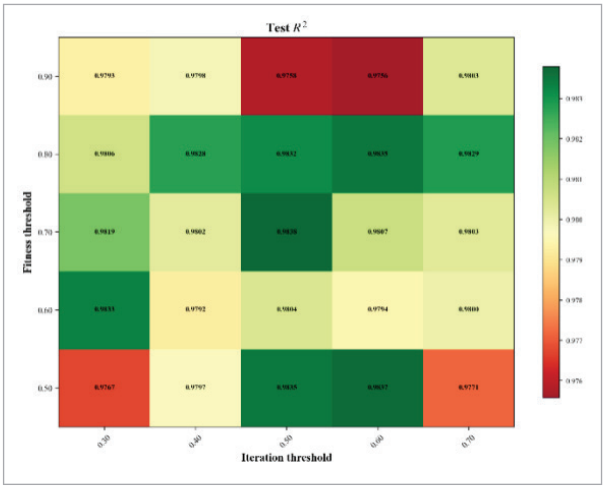


Figure 9
Sensitivity Analysis of Random Extended Search Strategy Parameters.



figured with fixed hyperparameters and fewer Optuna trials; therefore, the absolute R^2 values should be compared only within this analysis. The iteration threshold varied from 0.30 to 0.70 and the fitness threshold from 0.50 to 0.90. Across all combinations, the test R^2 ranges from 0.9756 to 0.9838, indicating that model performance is broadly stable and not critically dependent on precise tuning of either parameter. In terms of marginal averages, the iteration threshold yields the highest mean R^2 at 0.50, while the fitness threshold peaks at 0.80. Lower iteration thresholds may interrupt convergence, whereas overly high fitness thresholds reinitialize too few individuals and provide insufficient diversity. The narrow performance range confirms that EGWO remains stable under moderate parameter changes and does not depend on precise threshold tuning.

4.3.3. Comparative Experiment

To further validate and compare the competitiveness of the proposed EGWO-XGB-RF model, we set up comparative experiments and benchmarked it against mainstream PV power forecasting models from recent literature. Although the ablation experiments have verified the effectiveness of each component, comparison with current state-of-the-art models can better demonstrate the method's comprehensive capability in handling complex nonlinear features and coping with meteorological fluctuations. All models were implemented under identical experimental conditions: the same dataset, the same 70/15/15 chronological training-validation-testing split, the same input variables, and Bayesian optimization for hyperparameter tuning with the same number of trials where applicable. Models with structurally similar components were configured with the same parameter settings unless otherwise constrained by architectural differences. To this end, we selected the following models as benchmarks:

1. **CNN-LSTM:** a PV forecasting model that combines CNN spatial features with LSTM temporal modeling.
2. **CNN-GRU:** a PV power forecasting model that integrates CNN spatial feature extraction with GRU temporal modeling.
3. **LSTM-Transformer:** a multi-resolution collaborative framework for short-term PV prediction synergizing Transformer long-range dependencies with LSTM sequential evolution.
4. **Autoformer:** a Transformer-based time-series forecasting model that replaces self-attention with an auto-correlation mechanism for seasonal-trend decomposition, applied to the same 11-dimensional input features.
5. **CatBoost:** a gradient boosting decision tree model with ordered target encoding and Bayesian bootstrap, trained on standardized meteorological features for PV power prediction.
6. **SHAP-LightGBM:** a two-stage pipeline where SHAP values from a reference LightGBM rank the 11 input features by importance, the top-K features whose cumulative importance reaches 95% are retained, and a second LightGBM is trained on the selected subset for final prediction.
7. **HybrEnNet:** a hybrid ensemble learning model that integrates RF, XGBoost, and CatBoost.
8. **RF-CatBoost Ensemble:** an ensemble prediction model for PV power that integrates Random Forest and CatBoost.

Table 7 reports the mean and standard deviation over three independent runs; under identical experimental conditions, EGWO-XGB-RF achieves the lowest RMSE and MAE and the highest R^2 among all compared models. The results reveal distinct performance tiers across model categories. Among deep learning models, CNN-LSTM and Autoformer perform comparably, while CNN-GRU yields

Table 7
Performance Evaluation of Models in Comparative Experiment.

Model	RMSE	MAE	R^2
CNN-LSTM	0.3962 ± 0.0237	0.2404 ± 0.0186	0.9351 ± 0.0077
CNN-GRU	0.6476 ± 0.0366	0.4384 ± 0.0176	0.8267 ± 0.0193
LSTM-Transformer	0.2270 ± 0.0232	0.1434 ± 0.0097	0.9786 ± 0.0042
Autoformer	0.3918 ± 0.0031	0.2364 ± 0.0117	0.9367 ± 0.0010
CatBoost	0.2785 ± 0.0297	0.1517 ± 0.0105	0.9678 ± 0.0068
HybrEnNet	0.2240 ± 0.0087	0.1198 ± 0.0053	0.9792 ± 0.0016
SHAP-LightGBM	0.2169 ± 0.0091	0.1240 ± 0.0016	0.9806 ± 0.0016
RF-CatBoost Ensemble	0.2575 ± 0.0084	0.1367 ± 0.0033	0.9726 ± 0.0018
EGWO-XGB-RF	0.1875 ± 0.0076	0.0962 ± 0.0055	0.9854 ± 0.0012

an R^2 of 0.8267, indicating difficulty in capturing PV power dynamics under the chronological split. LSTM-Transformer reaches $R^2 = 0.9786$, benefiting from self-attention over long-range dependencies. Among tree-based methods, single-model CatBoost achieves $R^2 = 0.9678$, while SHAP-LightGBM reaches $R^2 = 0.9806$ by retaining features whose cumulative SHAP importance reaches 95%. The ensemble models HybrEnNet and RF-CatBoost Ensemble confirm the value of multi-model fusion. The proposed EGWO-XGB-RF leads across all metrics because it enhances feature quality at the source through parallel XGBoost-RF tree-feature generation and EGWO-based adaptive selection with mutual information constraints, rather than relying solely on model ensembling or greedy feature ranking.

4.3.4. Cross-Station Validation

To assess whether the proposed method generalizes beyond a single station, the full EGWO-XGB-RF pipeline was independently trained and evaluated on each of the remaining nine stations (station01–09) in the PVOD v1.0 dataset. The results are summarized in Table 8.

Eight of the nine stations achieve R^2 values above 0.975, with station04 and station05 exceeding 0.99. The only exception is station09, which yields an R^2 of 0.4825. Inspection of the data reveals that the training set for station09 is dominated by autumn and winter conditions with generally low irradiance levels, whereas the testing set consists of early-summer data with significantly higher irradiance. This severe distribution shift between training and testing distributions constitutes an extrapolation scenario that is beyond the model's learned domain. These results demonstrate that EGWO-XGB-RF generalizes robustly across stations with representative training data, and that its performance is consistent when the training and testing distributions are comparable.

Table 8
Performance of EGWO-XGB-RF on Independent PVOD v1.0 Stations station01–station09.

Station	RMSE	MAE	R^2
01	0.9344	0.4727	0.9755
02	0.6269	0.2513	0.9750
03	0.4950	0.2155	0.9900
04	0.6646	0.2305	0.9901
05	0.8715	0.3939	0.9907
06	0.2696	0.1148	0.9789
07	0.4670	0.1898	0.9767
08	0.5817	0.2433	0.9820
09	2.1942	0.9244	0.4825

5. Conclusion

Addressing the challenge of prediction accuracy under variable meteorological conditions, we propose an EGWO-XGB-RF methodology that couples ensemble feature generation with feature selection. Based on domain knowledge and correlation analysis, six meteorological features are selected from the 13 available in the PVOD v1.0 dataset, and five time-derived features are constructed to capture diurnal and seasonal cycles. Subsequently, XGBoost and RF models are trained in parallel, and the leaf node outputs of the decision trees are extracted to construct high-dimensional tree features, thereby mapping the original feature space to a highly discriminative feature space. An enhanced grey wolf optimizer (EGWO) is then introduced, which uses MAE combined with a mutual information constraint and a redundancy penalty as the fitness function to perform a search in the discrete feature space and select a near-optimal feature subset. Finally, the selected features are input into a LightGBM model for prediction. This framework achieves collaborative optimization among feature generation, feature selection, and the prediction model.

To verify the effectiveness of the proposed method, this study utilizes 28,896 records of 15-minute-interval PV operational data from station00 in the public dataset PVOD v1.0 from Science Data Bank, with the dataset chronologically divided into training, validation, and testing sets at a 70:15:15 ratio. Ablation experiments comparing 10 feature selection strategies confirm that the combination of XGBoost-RF tree-based feature generation with EGWO-based selection yields the most discriminative feature subset, with the mutual information constraint and the random extended search strategy each contributing incrementally to the result. In comparative experiments, EGWO-XGB-RF achieves $R^2 = 0.9854$, RMSE = 0.1875, and MAE = 0.0962 on station00, outperforming eight baseline models including HybrEnNet ($R^2 = 0.9792$), LSTM-Transformer ($R^2 = 0.9786$), and CatBoost ($R^2 = 0.9678$). Cross-station validation on eight additional PVOD stations yields R^2 values above 0.975, demonstrating the generalizability of the proposed method when training and testing distributions are consistent.

Collectively, the proposed EGWO-XGB-RF framework integrates ensemble feature generation, adaptive feature selection, and an efficient gradient-boosting predictor into a single pipeline, improving the accuracy of short-term photovoltaic power generation forecasting.

It should be noted that the current work is a purely data-driven study that has not yet incorporated PV-specific domain knowledge, such as solar altitude angle, clear-sky index, or irradiance decomposition. Additionally, due to the lack of weather-type classification labels in the PVOD v1.0 dataset, quantitative evaluation under different weather conditions could not be conducted. These limitations stem from the characteristics of the available data. Future work should address these gaps by incorporating physical priors into the forecasting framework, and by exploring transfer learning strategies across multiple PV stations to improve model generalizability and practical applicability.

References

1. Al-Dahidi, S., Alrbai, M., Alahmer, H., Rinchi, B., Alahmer, A. Enhancing Solar Photovoltaic Energy Production Prediction Using Diverse Machine Learning Models Tuned with the Chimp Optimization Algorithm. Scientific Reports, 2024, 14(1), 18583. https://doi.org/10.1038/s41598-024-69544-8
2. AlKandari, M., Ahmad, I. Solar Power Generation Forecasting Using Ensemble Approach Based on Deep Learning and Statistical Methods. Applied Computing and Informatics, 2024, 20(3-4), 231-250. https://doi.org/10.1016/j.aci.2019.11.002
3. Alrashidi, M., Rahman, S. Short-Term Photovoltaic Power Production Forecasting Based on Novel Hybrid Data-Driven Models. Journal of Big Data, 2023, 10(1), 26. https://doi.org/10.1186/s40537-023-00706-7
4. Banik, R., Biswas, A. Improving Solar PV Prediction Performance with RF-CatBoost Ensemble: A Robust and Complementary Approach. Renewable Energy Focus, 2023, 46, 207-221. https://doi.org/10.1016/j.ref.2023.06.009
5. Chen, T., Guestrin, C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIG-

KDD International Conference on Knowledge Discovery and Data Mining, 2016, 785-794. https://doi.org/10.1145/2939672.2939785

6. Danach, K., Aly, W. H. F. Adaptive Hyperheuristic Framework for Hyperparameter Tuning: A Q-Learning-Based Heuristic Selection Approach with Simulated Annealing Acceptance Criteria. European Journal of Pure and Applied Mathematics, 2025, 18(3), 6348. https://doi.org/10.29020/nybg.ejpam.v18i3.6348
7. Djeldjeli, Y., Taouaf, L., Alqahtani, S., Mokaddem, A., Alshammari, B. M., Menni, Y., Kolsi, L. Enhancing Solar Power Forecasting with Machine Learning Using Principal Component Analysis and Diverse Statistical Indicators. Case Studies in Thermal Engineering, 2024, 61, 104924. https://doi.org/10.1016/j.csite.2024.104924
8. El-kenawy, E. S. M., Alhussan, A. A., Mattar, E. A., Radwan, M. Feature Selection and Hyperparameter Tuning in Transformer-Based Deep Learning Models for Photovoltaic Power Forecasting Using the Swordfish Movement Optimization Algorithm (SMOA). International Journal of Electrical Power & Energy Systems, 2026, 174, 111509. https://doi.org/10.1016/j.ijepes.2025.111509
9. Emary, E., Zawbaa, H. M., Hassanien, A. E. Binary Grey Wolf Optimization Approaches for Feature Selection. Neurocomputing, 2016, 172, 371-381. https://doi.org/10.1016/j.neucom.2015.06.083
10. Eskandari, H., Saadatmand, H., Ramzan, M., Mousapour, M. Innovative Framework for Accurate and Transparent Forecasting of Energy Consumption: A Fusion of Feature Selection and Interpretable Machine Learning. Applied Energy, 2024, 366, 123314. https://doi.org/10.1016/j.apenergy.2024.123314
11. Gaboitaolelwe, J., Zungeru, A. M., Yahya, A., Lebekwe, C. K., Vinod, D. N., Salau, A. O. Machine Learning-Based Solar Photovoltaic Power Forecasting: A Review and Comparison. IEEE Access, 2023, 11, 40820-40845. https://doi.org/10.1109/ACCESS.2023.3270041
12. Goliatt, L., Yaseen, Z. M. Development of a Hybrid Computational Intelligent Model for Daily Global Solar Radiation Prediction. Expert Systems with Applications, 2023, 212, 118295. https://doi.org/10.1016/j.eswa.2022.118295
13. Guo, F., Yang, C., Xia, D., Xu, J. Short-Term Prediction of Photovoltaic Power Based on Improved CNN-LSTM and Cascading Learning. Energy Engineering, 2025, 122(5), 1975-1999. https://doi.org/10.32604/ee.2025.062035
14. Gupta, M., Arya, A., Varshney, U., Mittal, J., Tomar, A. A Review of PV Power Forecasting Using Machine Learning Techniques. Progress in Engineering Science, 2025, 2(1), 100058. https://doi.org/10.1016/j.pes.2025.100058
15. Hou, Z., Zhang, Y., Liu, Q., Ye, X. A Hybrid Machine Learning Forecasting Model for Photovoltaic Power. Energy Reports, 2024, 11, 5125-5138. https://doi.org/10.1016/j.egyr.2024.04.065
16. International Energy Agency. Renewables 2025. Paris: International Energy Agency, 2025. Available: https://www.iea.org/reports/renewables-2025.
17. Jia, D., Yang, L., Lv, T., Liu, W., Gao, X., Zhou, J. Evaluation of Machine Learning Models for Predicting Daily Global and Diffuse Solar Radiation Under Different Weather/Pollution Conditions. Renewable Energy, 2022, 187, 896-906. https://doi.org/10.1016/j.renene.2022.02.002
18. Jiang, H., Du, E., He, B., Zhang, N., Wang, P., Li, F., Ji, J. Analysis and Modeling of Seasonal Characteristics of Renewable Energy Generation. Renewable Energy, 2023, 219, 119414. https://doi.org/10.1016/j.renene.2023.119414
19. Karl, F., Pielok, T., Moosbauer, J., Pfisterer, F., Coors, S., Binder, M., Schneider, L., Thomas, J., Richter, J., Lang, M., Garrido-Merchán, E. C., Branke, J., Bischl, B. Multi-Objective Hyperparameter Optimization in Machine Learning-An Overview. ACM Transactions on Evolutionary Learning and Optimization, 2023, 3(4), 1-50. https://doi.org/10.1145/3610536
20. Khan, W., Walker, S., Zeiler, W. Improved Solar Photovoltaic Energy Generation Forecast Using Deep Learning-Based Ensemble Stacking Approach. Energy, 2022, 240, 122812. https://doi.org/10.1016/j.energy.2021.122812
21. Li, Y., Liu, G., Cao, Y., Chen, J., Gang, X., Tang, J. WNPS-LSTM-Informer: A Hybrid Stacking Model for Medium-Term Photovoltaic Power Forecasting with Ranked Feature Selection. Renewable Energy, 2025, 244, 122687. https://doi.org/10.1016/j.renene.2025.122687
22. Mirjalili, S., Mirjalili, S. M., Lewis, A. Grey Wolf Optimizer. Advances in Engineering Software, 2014, 69, 46-61. https://doi.org/10.1016/j.advengsoft.2013.12.007

23. Quan, R., Qiu, Z., Wan, H., Yang, Z., Li, X. Dung Beetle Optimization Algorithm-Based Hybrid Deep Learning Model for Ultra-Short-Term PV Power Prediction. iScience, 2024, 27(11), 111126. https://doi.org/10.1016/j.isci.2024.111126

24. Ramsami, P., Oree, V. A Hybrid Method for Forecasting the Energy Output of Photovoltaic Systems. Energy Conversion and Management, 2015, 95, 406-413. https://doi.org/10.1016/j.enconman.2015.02.052

25. Shi, J., Wang, Y., Zhou, Y., Ma, Y., Gao, J., Wang, S., Fu, Z. Bayesian Optimization-LSTM Modeling and Time Frequency Correlation Mapping Based Probabilistic Forecasting of Ultra-Short-Term Photovoltaic Power Outputs. IEEE Transactions on Industry Applications, 2024, 60(2), 2422-2430. https://doi.org/10.1109/TIA.2023.3334700

26. Suanpang, P., Jamjuntr, P. Machine Learning Models for Solar Power Generation Forecasting in Microgrid Application Implications for Smart Cities. Sustainability, 2024, 16(14), 6087. https://doi.org/10.3390/su16146087

27. Tercha, W., Tadjer, S. A., Chekired, F., Canale, L. Machine Learning-Based Forecasting of Temperature and Solar Irradiance for Photovoltaic Systems. Energies, 2024, 17(5), 1124. https://doi.org/10.3390/en17051124

28. Wang, F., Lu, X., Mei, S., Su, Y., Zhen, Z., Zou, Z., Zhang, X., Yin, R., Duić, N., Shafie-khah, M., Catalão, J. P. S. A Satellite Image Data Based Ultra-Short-Term Solar PV Power Forecasting Method Considering Cloud Information from Neighboring Plant. Energy, 2022, 238, 121946. https://doi.org/10.1016/j.energy.2021.121946

29. Xu, Y., Zheng, S., Zhu, Q., Wong, K., Wang, X., Lin, Q. A Complementary Fused Method Using GRU and XGBoost Models for Long-Term Solar Energy Hourly Forecasting. Expert Systems with Applications, 2024, 254, 124286. https://doi.org/10.1016/j.eswa.2024.124286

30. Yao, T. C., Wang, J., Wu, H. Y., Zhang, P., Li, S., Wang, Y., Chi, X., Shi, M. A Photovoltaic Power Output Dataset: Multi-Source Photovoltaic Power Output Dataset with Python Toolkit. Solar Energy, 2021, 230, 122-130. https://doi.org/10.1016/j.solener.2021.09.050

31. Yang, S., Luo, Y. Short-Term Photovoltaic Power Prediction Based on RF-SGMD-GWO-BiLSTM Hybrid Models. Energy, 2025, 316, 134545. https://doi.org/10.1016/j.energy.2025.134545

32. Ye, N., Zhi, C., Yu, Y., Lin, S., Liu, F. TCN-LSTM-AM Short-Term Photovoltaic Power Forecasting Model Based on Improved Feature Selection and APO. Sensors, 2025, 25(24), 7607. https://doi.org/10.3390/s25247607

33. Zafar, M. H., Khan, N. M., Mansoor, M., Mirza, A. F., Moosavi, S. K. R., Sanfilippo, F. Adaptive ML-Based Technique for Renewable Energy System Power Forecasting in Hybrid PV-Wind Farms Power Conversion Systems. Energy Conversion and Management, 2022, 258, 115564. https://doi.org/10.1016/j.enconman.2022.115564

34. Zhang, T., Zhao, W., He, Q., Xu, J. Optimization of Microgrid Dispatching by Integrating Photovoltaic Power Generation Forecast. Sustainability, 2025, 17(2), 648. https://doi.org/10.3390/su17020648

35. Zhang, X., Wu, Y., Wang, Y., Lv, Z., Huang, B., Yuan, J., Yang, J., Ma, X., Li, C., Zhang, L. Prediction of Photovoltaic Power Generation Based on a Hybrid Model. Frontiers in Energy Research, 2024, 12, 1411461. https://doi.org/10.3389/fenrg.2024.1411461

36. Zhao, Y., Peng, X., Tu, T., Li, Z., Yan, P., Li, C. WOA-VMD-SCINet: Hybrid Model for Accurate Prediction of Ultra-Short-Term Photovoltaic Generation Power Considering Seasonal Variations. Energy Reports, 2024, 12, 3470-3487. https://doi.org/10.1016/j.egyr.2024.09.025

37. Zhou, D., Liu, Y., Wang, X., Wang, F., Jia, Y. Combined Ultra-Short-Term Photovoltaic Power Prediction Based on CEEMDAN Decomposition and RIME Optimized AM-TCN-BiLSTM. Energy, 2025, 318, 134847. https://doi.org/10.1016/j.energy.2025.134847

This article is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution 4.0 (CC BY 4.0) License (http://creativecommons.org/licenses/by/4.0/).