Research Article

A Fault and Capacity Loss Prediction Method of Wind Power Station under Extreme Weather

Ling Li,[1] **Yixin Zhuo,**[1] **Wenchuan Meng,**[2] **Ze Chen** (iD),[3] **and Heng Wei**[1]

[1]*Dispatching Control Center of Guangxi Power Grid, Nanning, China*
[2]*Energy Development Research Institute, China Southern Power Grid, Guangzhou, China*
[3]*School of Electric Power, South China University of Technology, Guangzhou, China*

Correspondence should be addressed to Ze Chen; 2729174391@qq.com

Extreme weather events can severely affect the operation and power generation of wind farms and threaten the stability and safety of grids with high penetration of renewable energy. Therefore, it is crucial to forecast the failure and capacity loss of wind farms under extreme weather conditions. To this end, considering the disaster-causing mechanism of severe weather and the operational characteristics of wind farms, this paper first uses the density-based spatial clustering of applications with noise algorithm to cluster the units in the wind farm based on the operating characteristics affected by the weather, and uses correlation analysis methods to extract key disaster-causing factors in extreme weather; then proposes a prediction model based on feature-weighted stacking integration. The model adopts the stacking-integrated learning architecture to support multiple learners and performs feature weighting according to the prediction accuracy of each learner in the base learner, thereby improving the training effect of the meta-learner and improving the prediction accuracy of the model. The prediction model is used to predict each wind turbine group based on the extracted key features and to predict the failure and capacity loss of the wind farm. Finally, an example analysis is performed based on actual data from a wind farm, and the results show that the proposed prediction method can effectively predict the operational reliability of wind farms.

1. Introduction

In recent years, the energy crisis brought about by rapid socioeconomic development and the increasing depletion of fossil energy has become a global concern [1]. Therefore, accelerating the shift of energy consumption from traditional fossil fuels to green and clean energy and building low-carbon and environment-friendly renewable energy systems is the direction of future development [2]. Wind energy, as a clean energy source, is one of the most rapidly developing renewable energy sources. However, wind power is characterized by strong randomness and volatility [3], and is susceptible to weather [4]. Global climate change has led to frequent extreme weather, seriously affecting the normal operation and grid connection of wind power generation. Especially in extreme weather, such as extreme cold and typhoons, wind turbines in wind farms are prone to abnormal operation, such as icing, shutdown, and strong wind

cutouts. Such failures are caused by abnormal weather and have an impact on the normal operation and power supply capacity of the wind farm, resulting in a large deviation in the power forecast of the wind farm, which brings many difficulties to the combination of startup and shutdown, the arrangement and implementation of the grid dispatching plan, and challenges the safe and stable operation of a high proportion of renewable energy power systems [5]. It can be seen that accurately predicting the faults that will occur in the wind farm in extreme weather and analyzing the impact of the fault on the wind farm capacity are of great significance to improving the power prediction accuracy of the wind farm and maintaining the stable and safe operation of the power system.

In recent years, the frequent occurrence of extreme weather has brought potential risks to the operation of wind farms. As a result, scholars have conducted many studies on fault prediction and diagnosis of wind turbines in wind

farms under extreme weather conditions. According to modeling theory, these studies can be classified into methods based on physical models and data-driven statistical methods. Among them, statistical methods based on machine learning (ML) can automatically mine the connections between data features [6, 7], are simple to model, have fast calculation speed and high prediction accuracy, and have been widely used. Literature [8] selected key features according to the icing mechanism in extremely cold weather and used particle swarm optimization (PSO) optimized support vector machine (SVM) to predict the icing fault of wind turbine blades; Zhang et al. [9] studied the use of monitoring and data acquisition system data to detect icing on wind turbine blades, and proposed a prediction model based on the random forest (RF) algorithm; The study fully considers the mixed characteristics of short-term and long-term icing effects based on the physical extraction of bottom icing, and uses these characteristics to establish a Stacked-extreme gradient boosting (XGBoost) model to realize leaf icing diagnosis [10]; Tang et al. [11] proposed a fault detection model for doubly-fed wind turbine pitch system based on IHHO-light gradient boosting machine (LightGBM); literature [12] introduced a modeling method using weather research and forecasting models to predict the failure probability of wind turbines under typhoon weather. In addition, neural network methods have also been applied due to their efficient feature mining capabilities, including artificial neural network [13], long short-term memory [14], recurrent neural network [15], etc. In response to power loss caused by faults, Gao et al. [16] considered the impact of extreme weather events such as extreme cold and proposed a wind farm icing loss prediction model; literature [17] established the statistical correlation between meteorological conditions and icing energy loss based on wind turbine blade icing events, the proposed model can quickly and accurately predict the energy loss of wind turbines with numerical weather forecast data input. Literature [18] evaluated the loss of wind farm shutdowns under typhoon disasters and the risks it brought.

However, the abovementioned studies on fault shutdowns and shutdown losses caused by extreme weather usually take a single wind turbine in a certain extreme weather scenario as the experimental object. In actual operation, wind farms will be affected by different extreme weather conditions, and there are multiple wind turbines with different installed capacities in the wind farm, with differences in distribution area and altitude, different operating state characteristics, and different affected characteristics. Therefore, research on fault and capacity loss prediction of wind farms under different extreme weather scenarios needs to take all wind turbines in the wind farm as research objects and propose more generalizable and accurate prediction model methods. Existing research has shown that by using the stacking idea to integrate different models, the limitations of a single model can be overcome, the applicable scope and advantages of various learners can be integrated, and it has better generalization performance, combining the advantages of various models to obtain better prediction results [19]. It has been used in the multifault detection and classification of wind turbines [20]. Literature [20]

used AdaBoost, K-nearest neighbors, and logistic regression classifiers to adopt a stacking integrated model to achieve effective identification of wind turbine faults.

In summary, in order to overcome the above difficulties and complete the prediction task of different faults and fault loss capacity of wind farms under extreme weather conditions. This paper uses the density-based spatial clustering of applications with noise (DBSCAN) clustering method to divide the wind turbines in the wind farm into different turbine groups according to the reliability characteristics affected by extreme weather and further extracts key disaster-causing factors in extreme weather using correlation analysis technology. On this basis, a prediction model that integrates multiple models applied to wind turbine fault prediction and diagnosis: SVM, RF, XgBoost, and LightGBM, using the Stacking integration idea is proposed and combined with the feature weighting method to improve it to improve prediction accuracy, to predict faults and capacity losses of wind farms under extreme weather conditions. The main contributions are as follows:

(1) In order to avoid modeling each turbine in the wind farm separately, this paper uses the DBSCAN clustering method to divide the turbines in the wind farm into different turbine groups and perform fault prediction separately, which not only simplifies the prediction task but also facilitates the analysis of the fault conditions of different turbines in the wind farm and calculates the capacity loss caused by the fault.

(2) In order to study the disaster mechanism of extreme weather and solve the problem of model input redundancy, the correlation analysis method is used to extract the key weather characteristics that affect the normal operation of each unit group in the wind farm, which improves the training speed and accuracy of the model.

(3) A stacking ensemble that integrates multiple mature models is proposed to overcome the disadvantage of a single prediction model. And on the basis of the traditional stacking ensemble method, an improved method based on feature weighting is proposed, in which the outputs of different primary learners are set with weights based on errors in the model training process, and the output of the algorithm with the better prediction effect is amplified, to improve the overall prediction accuracy.

(4) Using the numerical weather prediction (NWP) forecast data of a wind farm through experimental analysis, the experimental structure verifies the superiority of the proposed prediction model and the effectiveness of the prediction method.

The rest part of this study is as follows: Section 2 describes the stacking integration idea and the principle of the DBSCAN clustering method. Section 3 introduces the whole modeling process of the prediction method. In Section 4, the arithmetic analysis and experimental validation are performed on real data. Section 5 summarizes the conclusions of the study.
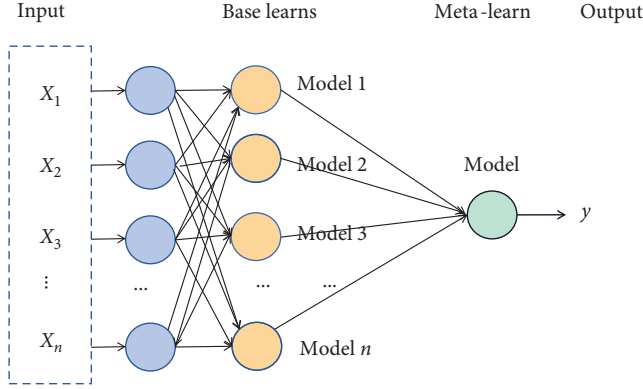
FIGURE 1: Framework diagram of stacking ensemble.

2. Algorithm and Model Principles

2.1. Stacking Ensemble. Stacking ensemble learning is a method that can combine the advantages of multiple models to obtain better and more robust prediction results than a single model. Stacking is a hierarchical model integration framework that is usually designed as two layers, where the first layer consists of several algorithms called base learners, while the second layer is called meta-learner for stacking ensemble [21]. The benefit of this ensemble is the estimation of base learners can be used to teach the meta-learner, resulting in more accurate outcome prediction [22]. It uses the original dataset to train different learners in the base learner, then combines the output of each learner into a new dataset, which is finally used by the second-stage meta-learner to obtain the final prediction [23]. The framework of the stacking ensemble is shown in Figure 1.

The four models, including SVM, RF, XGBoost, and LightGBM, which have achieved good results in different fault prediction of wind turbines, are used as the base learners in this study; LightGBM is selected as a meta-learner because of its better training performance. Different types of model integration are adopted to obtain better generalization performance so that it can cope with the prediction of multiple faults. And these learning models are briefly described below:

(a) SVM is a supervised learning model used for classification and regression, which uses both classification and decremental challenges [24]. SVM maps training data to a multidimensional space called decision space, creating a separating surface called the decision boundary, as shown in Equation (1), dividing the decision space into different regions, and has been widely used in classification problems;

$$f(x) = \text{sgn}\left[\sum_{i=1}^{n}\alpha_i y_i K(x_i \cdot y_i) + b\right], \tag{1}$$

where α is the Lagrange multiplier, $K(x_i \cdot x_j)$ is the kernel function, and b is the classification threshold.

(b) RF is a typical bagging integration algorithm, which uses a decision tree as a base learner and is built by bagging thought. By adopting the resampling method, n samples are randomly selected from the original training set (N) repeatedly and replaced to generate a new training set. K-classification trees are generated to form an RF that is determined by the self-service sample set, and the classification of the new data is determined by the score formed by how many classification trees vote [25].

(c) XGBoost is a learning algorithm based on tree integration., the basic idea of which is to compute the final classification result by integrating multiple basic trees. Meanwhile, based on the gradient-augmented decision tree, the second derivative of the loss function on the prediction result is introduced, and the complexity of a tree model is added as a regularization term in the objective function to prevent overfitting and improve the generalization performance of the model [26]. Its objective function is shown in Equation (2).

$$\text{OBJ} = \sum_{i=1}^{n} L(y_i, \hat{y}_i(t-1) + f_t(x_i) + \Omega(f_t)), \tag{2}$$

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda\|w\|^2. \tag{3}$$

In Equation (2), the first term represents the loss function, y_i is the true value, $\hat{y}_i(t-1)$ is the predicted result of the first $t-1$ tree, $f_t(x_i)$ is the model prediction result of the tth tree, and the sum of the two is the prediction result of sample i after the tth iteration, with Ω as the regularization term. In Equation (3), the first term γT controls the complexity of the tree by the number of leaf nodes and their coefficients; the second item is the L2 norm of the leaf node value, which is used to control the weight score of the leaf node.

(d) LightGBM is a distributed gradient boosting algorithm based on decision tree algorithms, which adopts one-sided gradient sampling and exclusive feature bundling methods and uses histogram algorithms for optimization. It selects the leaf-wise decision tree growth strategy with depth limitation, which has the characteristics of fast running speed, low memory consumption, and high accuracy. It is widely used in classification and regression problems, so it is used as a meta-learner in model building [27].

2.2. DBSCAN. DBSCAN is a classical clustering algorithm based on density, and the algorithm considers a class as a high-density target region, which is divided by the low-density region in data space [28]. The basic idea of the DBSCAN algorithm is to find any core point and then find the sample set so that all core objects can reach the density, which is a cluster. The specific implementation method is as follows: First, determine the parameters including the

neighborhood radius (epsilon (Eps)) and the threshold value of the number of data objects in the neighborhood (MinPt); scan the whole dataset D, the DBSCAN method starts from a random point and retrieves its neighborhood, and if it is a core point, find all density connected data points from the core point to form a cluster; next, a new unvisited point is retrieved to find a new cluster by repeating the above process, until there is no new core point in the dataset, and the data points not included in any cluster are noise points [29]. At the same time, in order to evaluate the rationality of clustering results, the widely used silhouette coefficient (SC) in the evaluation of clustering rationality is used as an indicator to measure the intragroup aggregation and intergroup separation of clustering results [30]. Among them, the definition of contour value $s(x_i)$ and SC are shown in Equations (4) and (5), respectively, the contour value $s(x_i)$ is in the range $[−1,1]$, and the grouping result is optimal when the SC is 1, and the closer to 1, the more reasonable the grouping result.

$$s(x_i) = \frac{b(x_i) - a(x_i)}{\max(a(x_i),\ b(x_i))}, \quad (4)$$

$$SC = \frac{1}{m}\sum_{i=1}^{m} s(x_i), \quad (5)$$

where $b(x_i)$ is the minimum of the average distance between sample point x_i and other cluster units, and $a(x_i)$ is the average of the distance between sample point x_i and all sample points within the same cluster.

DBSCN algorithm, as an unsupervised ML clustering algorithm, does not need to cluster data points with prelabeled targets, which facilitates the division of wind turbines with different extreme weather conditions. In the experiment, we found that there are often some outliers in the monitoring data of the wind farm, and the records of the operation status of each wind turbine in the wind farm need to be included. The DBSCAN algorithm can divide regions with sufficient density into clusters and discover clusters of arbitrary shapes in a spatial database with noise. Thus, it has a unique advantage over popular clustering algorithms such as K-means and hierarchical clustering.

3. Materials and Methods

The method first uses the DBSCAN algorithm to divide the wind turbine group based on the data recorded during the prediction cycle of wind farm failures and capacity losses. Then, the Spearman correlation analysis method is used to analyze the most critical factors affecting the operating status of different wind turbine clusters in extreme weather to lay the foundation for the input data of the prediction model. Finally, feature weighting is used to improve the traditional Stacking integrated model to predict wind farm faults and capacity loss. The flowchart of fault and capacity loss prediction method based on feature-weighted improved stacking ensemble is shown in Figure 2.
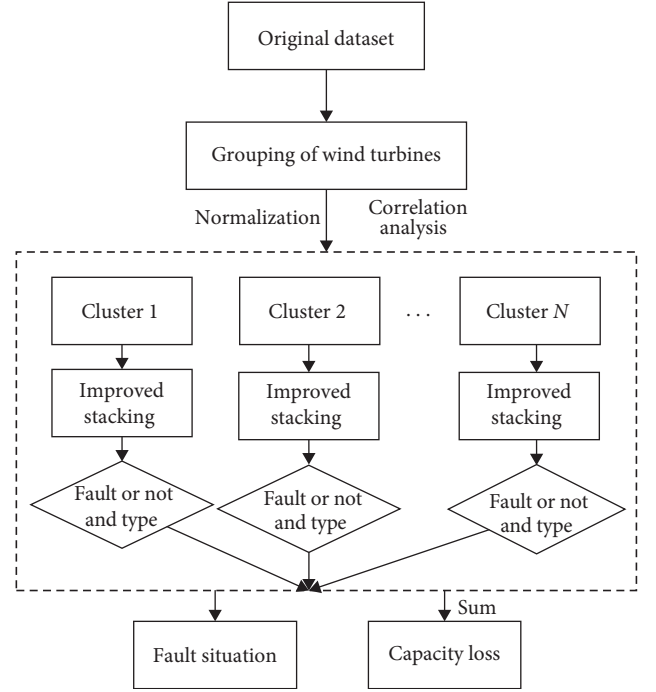

FIGURE 2: Flowchart of the fault and capacity loss prediction method based on feature-weighted improved stacking ensemble.

3.1. Cluster Model of Wind Turbines. There are multiple wind turbines in a wind farm. In order to accurately judge the wind farm failure scenario and further analyze the specific operating conditions of wind turbines arranged in the wind farm and the capacity loss caused by failures, a cluster model of wind turbines based on the DBSCAN clustering algorithm is proposed in this paper. The wind turbines were grouped by identifying the operating state of each wind turbine and the similar characteristics affected by extreme weather, and the operating state of each wind turbine group was predicted by modeling to analyze the lost capacity.

The specific steps of using the DBSCAN algorithm to cluster wind farms are as follows: (1) select the running state and abnormal state reasons at the same time as the characteristic characterization of a single unit to form dataset D; (2) the grouping algorithm DBSCAN model was established, and the characteristic dataset D of each unit was taken as the input for clustering; (3) the clustering contour coefficient was used to measure the rationality of the clustering results; and (4) adjust the model parameters to obtain the optimal grouping results, and output the group results.

3.2. Correlation Analysis and Data Processing. The influence mechanism of extreme weather on wind farms is complicated and difficult to analyze. The disaster factors include wind speed, wind direction, temperature, humidity, and other multidimensional weather characteristics. Moreover, the wind tower in wind farms monitors these weather characteristics at different heights, and the redundancy of characteristic information will affect the accuracy of the prediction model. Therefore, in order to reduce the dimension of data and analyze the weather features that are more critical to the
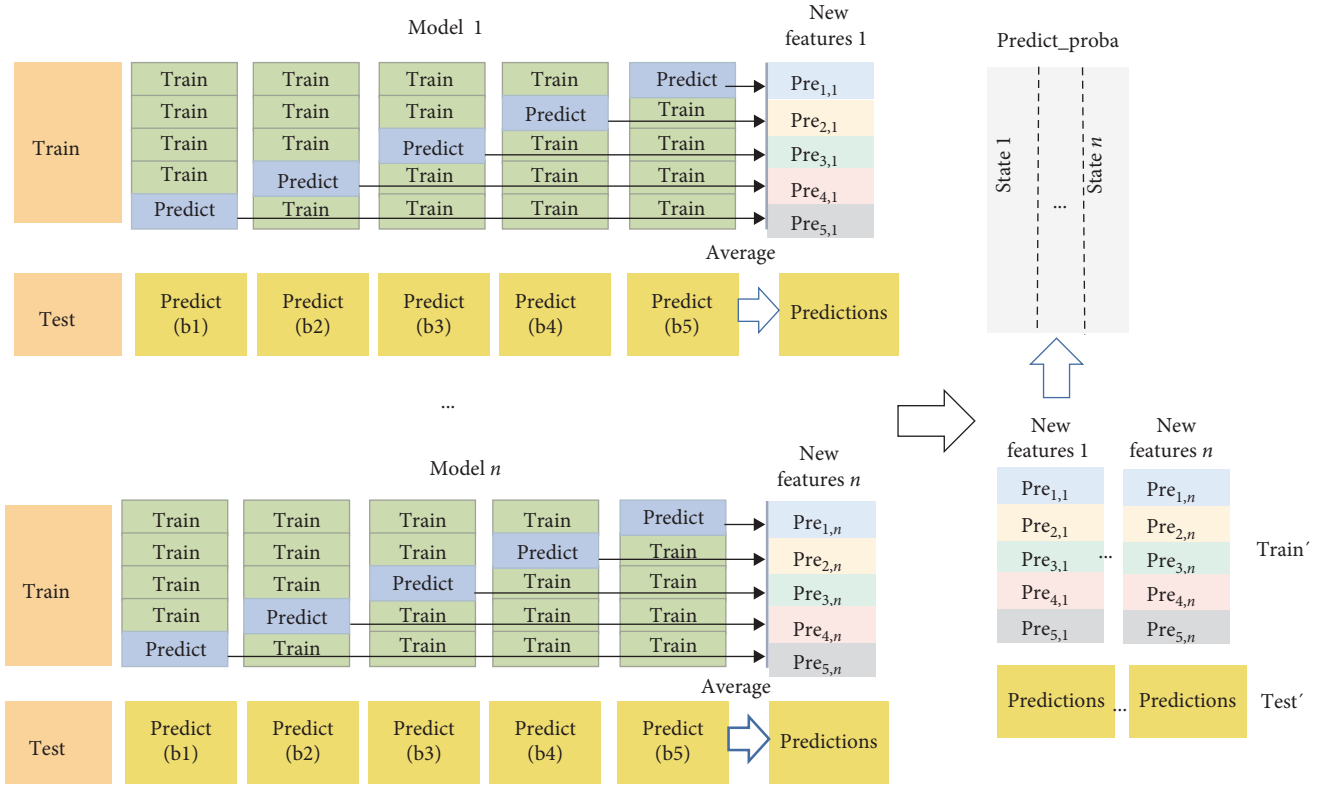
FIGURE 3: K-fold cross-check flowchart.

operating status of wind farms in extreme weather, the correlation analysis method is adopted to analyze the correlation between various weather factors and the operating status of wind farms in different extreme weather conditions on the basis of wind farm unit groups, and the high correlation data are selected to reduce the dimension of input characteristics.

The Spearman correlation coefficient can judge the correlation between nonlinear variables. It uses the difference of rank order to evaluate the nonlinear correlation between two variables. Spearman's correlation coefficient ranges from -1 to 1. The larger the absolute value of the correlation coefficient, the stronger the correlation between variables, and the smaller the absolute value, the weaker the correlation between variables. Its expression is as follows:

$$\rho(x, y) = 1 - \frac{6\sum_{i=1}^{n}|R(x_i) - R(y_i)|^2}{n(n^2 - 1)}, \tag{6}$$

where n is the number of samples, $R(x_i)$ and $R(y_i)$ are the data of x and y variables sorted by size, respectively, and then the calculation result of the sorted rank is recorded.

The feature variable data after extraction are cleaned, and the abnormal and missing sequences are deleted. At the same time, in order to unify dimensions, the data after cleaning are normalized, and Max–Min is set as the original continuous data normalization method, whose expression is as follows:

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}}, \tag{7}$$

where $x_{\max}$ and $x_{\min}$ are the maximum and minimum values of the data sequence, respectively.

3.3. K-Fold Cross Check.
After the data processing is completed, it is input into the stacking integrated model. In order to avoid overfitting the model, the data are usually divided by K-fold cross-validation, and then the primary learner in the Stacking integrated model is trained. They are taking $K = 5$ as an example. In combination with Figure 3, the K-fold cross-check process is described in detail. The specific steps are as follows:

Step 1: The dataset after feature selection and data processing is divided into a training set and a test set.

Step 2: K-fold cross-verification for primary learners: The original training set was randomly divided into K equal parts; each primary learner took 1 part of it as the test set and the remaining $K - 1$ part as the training set. The training set was used to train each primary learner, and the test set was predicted to generate five predictions. The predicted results of each primary learner were combined as the training set of the meta-learner.

Step 3: Each primary learner predicts the original test set, respectively, and averages the predicted results as the verification set of the meta-learner.

Step 4: The new dataset generated by the primary learner is input as a meta-learner to obtain the prediction results.

3.4. Feature-Weighted Improved Stacking Ensemble Model. The stacking integrated model described above is structured in the same proportion of the output of the predicted results of each base learner when the primary learner generates the training data of the meta-learner through *K*-fold cross-check. However, the attribute characteristics of the meta-learner training dataset have a great impact on the performance of stacking integrated prediction [31]. When the input samples of the meta-learner have more characteristic information and fewer errors, it is obvious that the stacking integrated algorithm has better performance. Therefore, this article improves the stacking integration based on this. The idea is to assign a weight value to the prediction result of each base learner according to the training situation and the error of each base learner itself so that the input features of the constructed meta-learner are more affected by the output part of the algorithm with better prediction effect so that the meta-learner can obtain better training effect. This method adjusts the input features of the meta-learner by means of weighting, which is called feature weighting. The specific steps to improve the method are as follows:

Step 1: Evaluate the training results of each base learner: the predicted results of each learner on the training set after the completion of *K*-fold cross-training are taken as the evaluation object of the training results. For the classifier, the accuracy score is taken as the evaluation standard expression as follows:

$$\text{accucancy\_score} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \quad (8)$$

where TP stands for correct, TN stands for correct, FP stands for unexpected, and FN stands for missing.

Step 2: Calculation of feature weights: In order to make the input features of the meta-learner obey the probability distribution, the construction of feature weights should be standardized based on the evaluation results of each base learner. The processing method is shown in Equation (9):

$$\omega_i = n \cdot \text{accuracy\_score}_i / \sum_{j=1}^{n} \text{accuracy\_score}_j, \quad (9)$$

where ω_i is the weight of the *i*th learner in the base learner, n is the number of learners.

Step 3: In *K*-fold cross-validation, we set the output of the base learner to predict. The characteristics of a probe, including various operating states and prediction probabilities, as shown in Figure 3. Based on the output of the base learner and the weights of each learner of the base learner, a feature-weighted meta-learner training dataset is constructed. The specific construction method is as follows: For the n models of the test set on the *i*th fold in the *K*-fold cross-validation, there will be n corresponding prediction results. By multiplying the weight of the corresponding learner, the prediction result train' can be written as follows:

$$\text{Train}_i' = \left[\omega_1 \text{Pre}_{i,1}, \omega_2 \text{Pre}_{i,2}, \ldots, \omega_n \text{Pre}_{i,n}\right], \quad (10)$$

where Train' represents the training data of the meta-learner, $\text{Pre}_{i,n}$ is the predicted value of the *i*th fold check of the *n*th learner, ω_n is the feature weight of the *n*th learner above.

Step 4: The datasets generated after feature weighting are input into the meta-classifier of the stacking integrated model for training, and the numerical weather forecasting NWP data of the trained model is used to predict wind farm faults.

The working principle flowchart is shown in Figure 4.

The specific process description of the improved stacking integrated forecasting model with feature weights is shown in Figure 5.

4. Experiments and Analysis

The experimental data were obtained from multidimensional NWP information, including wind speed, wind direction, temperature, humidity, etc., at different heights of a wind tower in a wind power plant in Guangxi Province, China, as well as the operation status and failure of the whole wind power plant and all fans in the plant. The time resolution of the data was 15 min. The test environment used in the example experiment was deployed in a local high-performance computing cluster with the Linux operating system, using an eight-core Intel i7 processor, 64 GB memory, and a working frequency of 2.3 GHz. Models were trained and tested in Python 3.8.5.

4.1. Wind Turbine Cluster Scheme. According to the established wind turbine cluster model, 35 units of the wind farm were grouped as research objects. The operating states of different units within the same time interval were selected as characteristic characterization, and the different operating states were coded to form characteristic datasets. By adjusting the parameters to obtain the optimal classification model with a higher contour coefficient, a clustering model with an Eps of 3 and the MinPt of 20 was ultimately obtained. The model contour coefficient was 0.89, and the clustering results
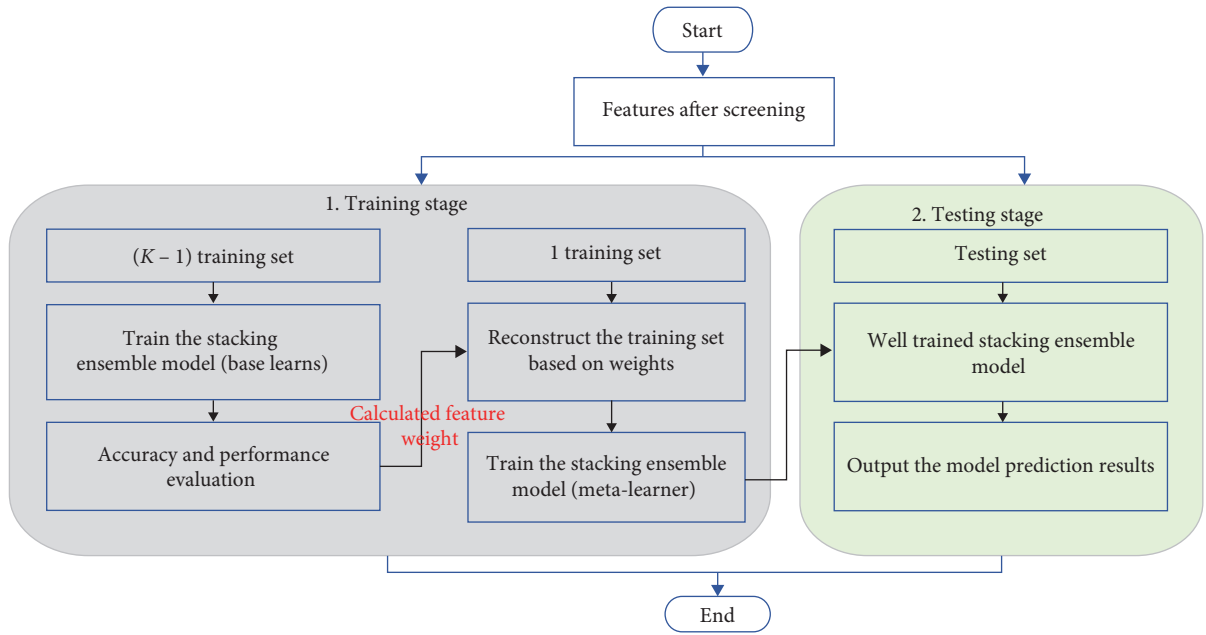
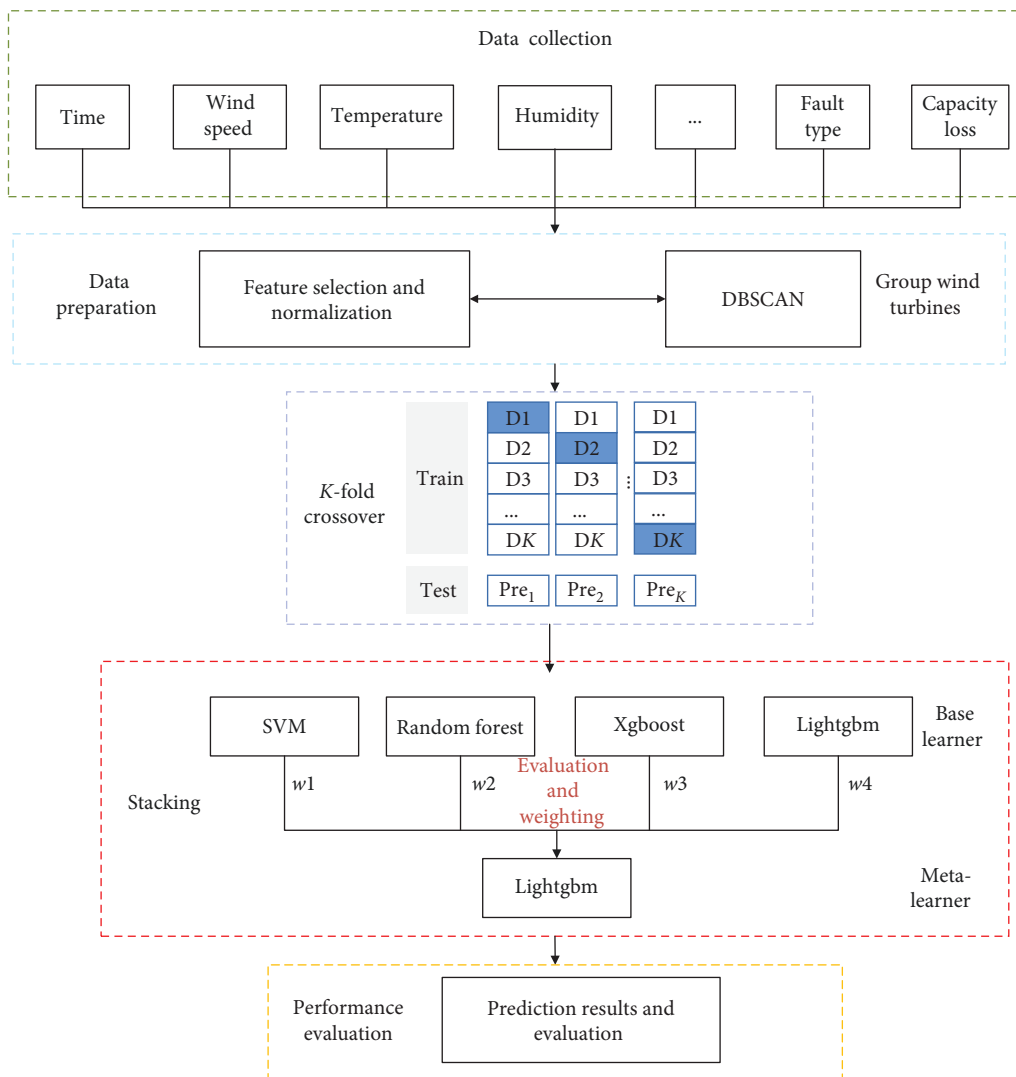FIGURE 4: Improved model working principle flowchart.

FIGURE 5: The flowchart of the integrated stacking model with improved feature weights.

TABLE 1: Group clustering results based on DBSCAN.

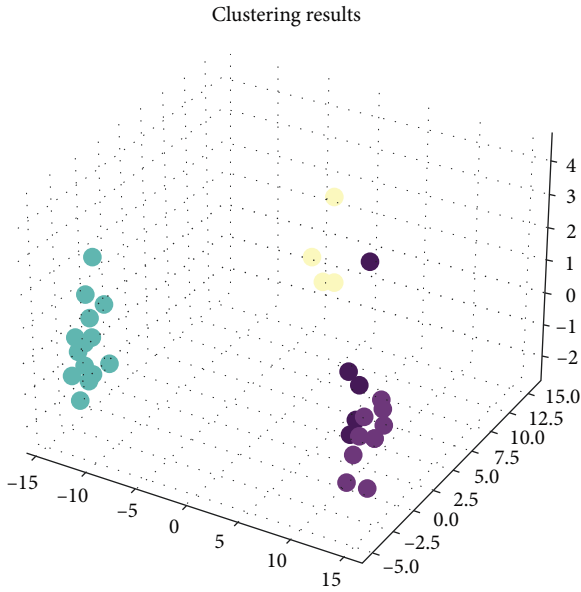| Clustering criterion	Clustering results	Fan unit number	Unit group capacity (MW)	Representative unit
	Cluster 1	101, 102, 103, 104, 105, 106, 107, 108, 109, 201, 202, 203, 208, 209, 306, 307	38	102
Characteristics of wind turbine operating state	Cluster 2	204, 205, 206, 207	10	204
	Cluster 3	301, 302, 303, 304, 305, 401, 402, 403, 404, 405, 406, 407, 408, 409, 410	32	409

Clustering results

FIGURE 6: Cluster distribution map. Green color represent Cluster 1 all units, yellow represent Cluster 2, and purple represent Cluster 3.

were obtained; the clustering results are shown in Table 1, and the Cluster distribution map is shown in Figure 6.

At the same time, in order to characterize the operation status of the cluster, the average intracluster correlation coefficient (AICC) of each unit in the cluster is used in Equation (11) [32], and the unit with the highest AICC in the cluster is the representative unit, and its operation status is defined as the cluster operation status.

$$\text{AICC}_p = \frac{1}{m} \sum_{q \in C} \frac{\text{Cov}(X_p, X_p)}{\sqrt{\text{Var}(X_p)\text{Var}(X_q)}}, \tag{11}$$

where p and q are any two units in the cluster, m is the number of units in the cluster, X_p, X_p are the real-time operating characteristics of the two units, $\text{Cov}(X_p, X_p)$ is the covariance of X_p, X_p, $\text{Var}(X_p)$, $\text{Var}(X_q)$ are the variances of X_p, X_p, respectively.

4.2. Feature Selection. The proposed Spearman correlation analysis method is used to analyze the correlation between the multidimensional NWP information in the wind farm and the operation status of the wind farm. Taking the correlation degree of multidimensional weather characteristics,

including wind speed, wind direction, temperature, and humidity measured at the 10-m height of the wind tower in the wind farm and the fault type of each unit group in the wind farm as an example, Equation (3) is used to calculate the correlation of each index, and the correlation result is obtained. The thermal diagram is drawn as shown in Figure 6.

As shown in Figure 7, the correlation between unit group 2 and the weather characteristics measured at this height is low, and the correlation between different unit groups and different weather characteristics is also different. Therefore, when forecasting different unit groups, not only the same weather characteristics measured at different heights but also the correlation analysis between different characteristics measured at the same point should be carried out, and the characteristics with the highest correlation should be selected as the model input to predict the operation status of the unit group.

4.3. Fault Identification and Capacity Loss Prediction of Wind Power Station under Extreme Weather. Two experiments were used to predict the fault and capacity loss of wind farm stations under extreme weather. Experiment 1 takes the whole wind farm station as the experimental object for preliminary analysis, and Experiment 2 takes each wind turbine group in the wind turbine group scheme proposed in Section 4.1 as the experimental object for in-depth analysis. The above feature weighted improved stacking ensemble learning is used as the prediction model to build the overall prediction framework of the two target features.

Experiment 1: The experimental data used are more than 2,800 sets of data collected from the wind farm during a period of time affected by cold currents and typhoons during November–December 2022, and the wind turbine in the wind farm has abnormal operating conditions, including icing fault and strong wind cutting. When a certain fault occurs to a wind turbine in the wind farm, the operation status of the wind farm will be recorded as the fault condition. Therefore, the operation status of the wind farm station in extreme weather is divided into three types: normal, icing, fault, and cutout by strong wind. The data are randomly reconstructed according to the proportion of running states. The training set accounts for 90%, and the rest is the test set.

The comparison of the prediction results of the same sample data by single classification models such as RF, XGBoost, and stacking ensemble learning model based on feature weighting is shown in Figure 8, where the parameters of each model are obtained by Bayesian optimization, and the model parameters and optimization settings are shown in
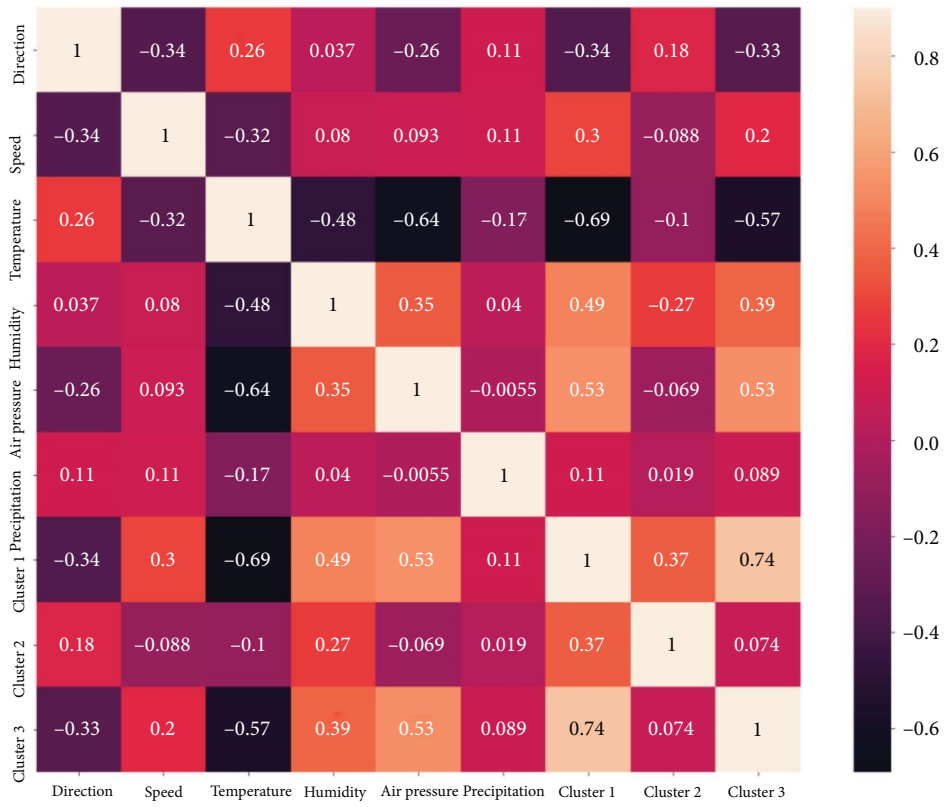
Figure 7: Spearman correlation coefficient of wind farm multidimensional data.
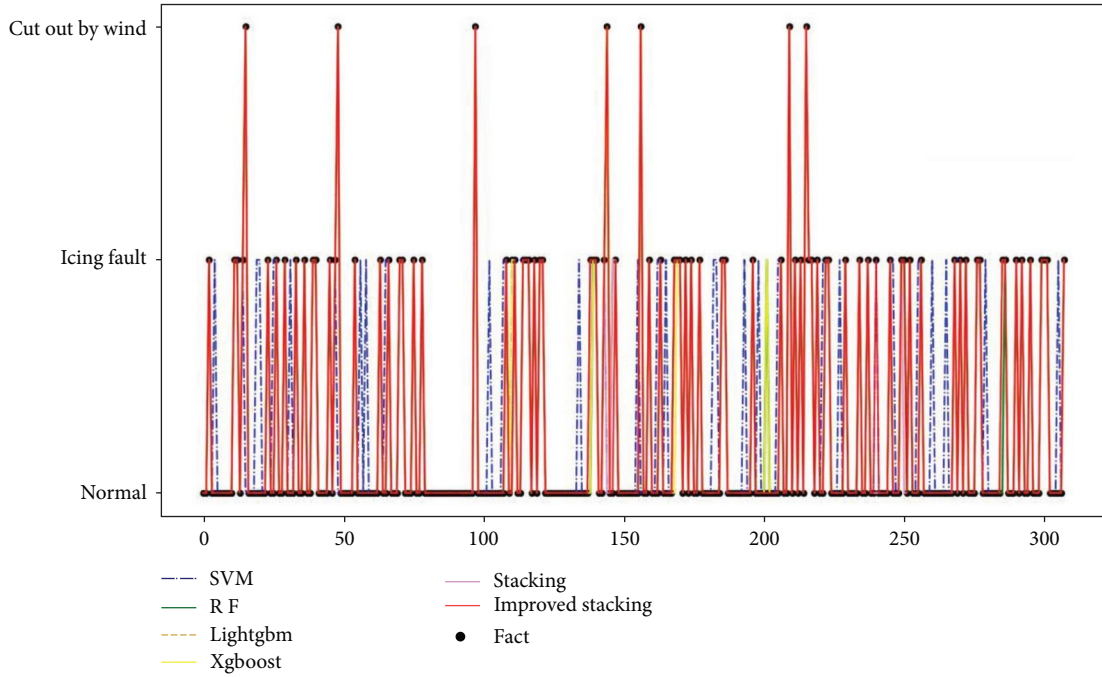

Figure 8: Fault prediction results of wind farm stations in extreme weather based on feature-weighted improved stacking.

TABLE 2: Table of parameters and optimization settings for different models.

Model	Parameter setting	Hyperparameter optimization range	Optimal parameter
SVM	"kernel," "C"	["linear," "RBF," "sigmod"]; [0.01, 0.03, 0.1, 0.3, 1, 3]	"linear"; 1
RF	"n_estimators"; "max_depth"; "max_features"	[10, 50, 100, 500]; [3, 5–7, 5–7, 9, 12, 15]; [2, 4, 6, 8, 10, 12]	100; 9; 8
XGBoost	"n_estimators"; "learning_rate"; "max_depth"	[10, 50, 100, 500]; [0.01, 0.025, 0.05, 0.1]; [3, 5–7, 5–7, 9, 12, 15]	100; 0.05; 7
LightGBM	"n_estimators"; "num_leaves"; "learning_rate"; "max_depth"	[10, 50, 100, 500]; [10, 20, 30; 50]; [0.01, 0.025, 0.05, 0.1]; [3, 5–7, 5–7, 9, 12, 15]	100; 20; 0.1; 6

TABLE 3: Comparison table of fault discrimination accuracy score of different models.

Model	SVM	RF	Xgboost	LightGBM	Stacking	Improved stacking
accuracy_score	95.2%	96.0%	96.4%	98.9%	98.2%	99.6%

Table 2. The parameters of each base learner in the stacking ensemble model are the same as those of the single model.

The accuracy score described in Equation (8) was used to compare the model prediction results, and the results are shown in Table 3.

Combining the results in Figure 7 and Table 2, we can find that the proposed feature-weighted improved stacking ensemble model has higher accuracy in the task of wind farm station fault discrimination prediction under extreme weather, and it is applied to the prediction of wind turbine group faults due to its own strong generalization.

Experiment 2: Due to the insufficient amount of data for other faults, such as high wind cutouts and lightning strikes for some groups of units, it is not easy to guarantee experimental accuracy. Therefore, this experimental data selects more than 1,300 sets of data from wind farms with a high proportion of icing faults to reclean the data. For each unit group, the feature-weighted stacking integration model is used for training prediction. The model input data select the wind speed, wind direction, temperature, humidity, and other weather characteristics of the point with the highest correlation with each unit group as the model input, and the ratio of the training set to the test set is set at 8 : 2. The output is the operation status of each unit group in the wind farm station under extreme weather, and the affected capacity of the station is calculated according to the status of each unit group at the same time. The predicted results of the model are shown in Figure 8. Figure 9(a) shows the comparison between the faults of the three unit clusters and the actual situation, and Figure 9(b) shows the predicted results of the capacity loss of the comprehensive unit group operation state.

As shown in Figure 9(a), the prediction accuracy_score of the operation status of each unit group has reached 0.996,

0.996, and 1, respectively, compared with other models, it has higher accuracy and good prediction effect for unit groups with different reliability characteristics, with strong generalization. As shown in Figure 8, the prediction error of impact capacity is small, indicating that the proposed method can effectively predict the impact capacity of extreme weather faults. The causes of the errors are analyzed and summarized as follows: first, the characteristics of the units in the same unit group affected by extreme weather cannot be completely consistent, and the individual units and the unit group are not consistent with wind turbine clusters; secondly, the selected representative units cannot fully reflect the operating state characteristics of the test. The proposed prediction model cannot guarantee complete accuracy for the fault prediction of the wind turbine cluster.

Combining the above two experimental results, the results of Experiment 1 and Experiment 2 prove that the feature-weighted improved stacking model has the advantages of adaptability of the traditional stacking model compared with a single learning tool. Meanwhile, the feature-weighted improved stacking model highlights the training model with higher accuracy by introducing a feature-weighted mechanism. The critical characteristics of extreme weather are better extracted. Compared with the traditional Stacking model, the feature weighted improved stacking model improves the discrimination accuracy of extreme weather fault types. Experiment 2 proves that the proposed fault-affected capacity prediction method combining wind farm, wind turbine cluster, and feature-weighted improved stacking model can effectively predict the impact of faults on the overall stacking capacity of wind farm wind turbines and yards. Therefore, the method described in this paper can accurately and effectively predict the failure scenario and loss capacity of wind farm stations in extreme weather.
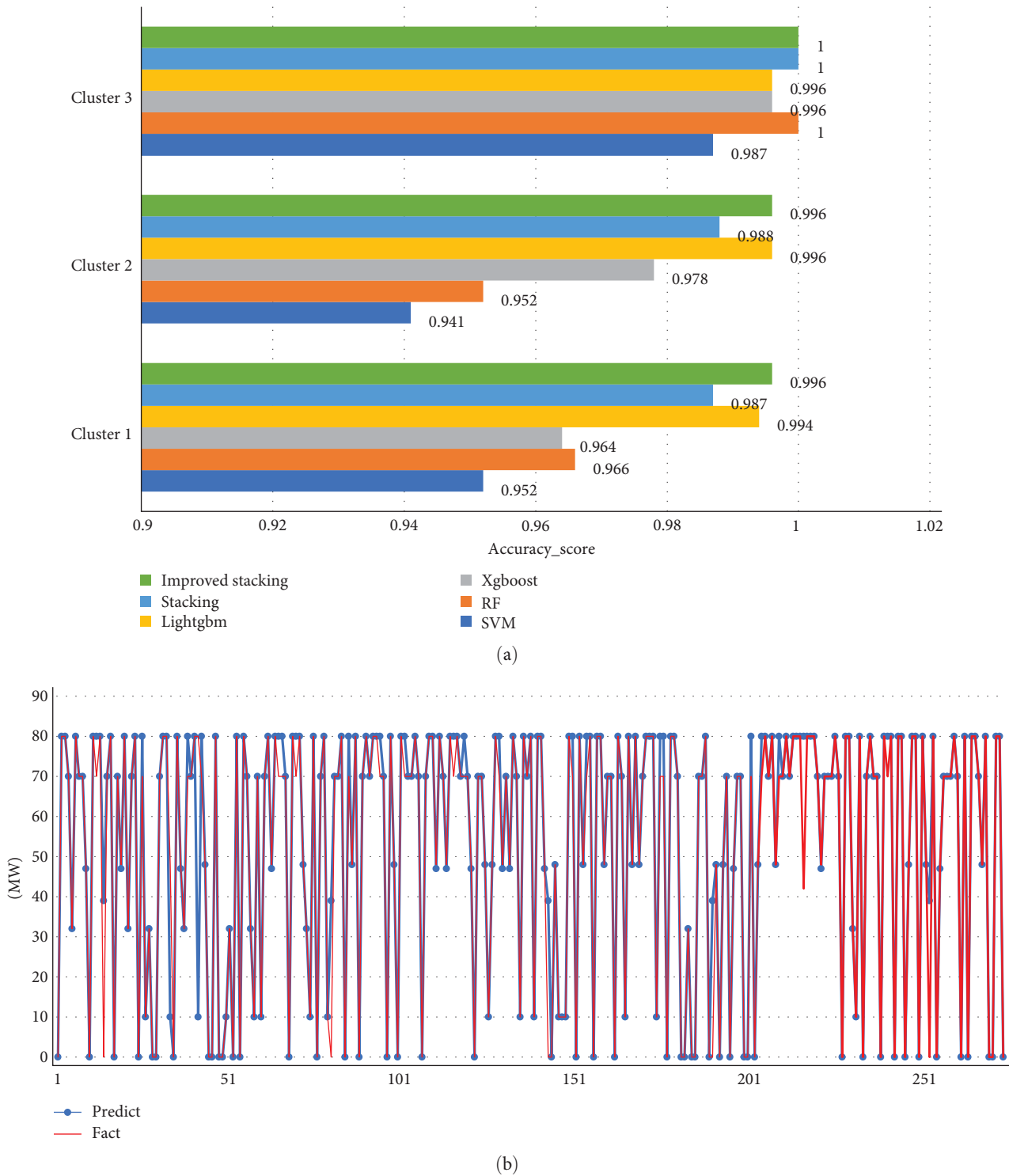
(a)

(b)

FIGURE 9: Graph of model prediction results. (a) Comparison of prediction accuracy of different algorithms for each unit group. (b) Capacity loss forecast due to failure.

5. Conclusion

We developed a feature-weighted improved stacking ensemble learning model to solve the problem of wind turbine fault identification and prediction of capacity loss caused by faults under extreme weather scenarios. This method completes the prerequisite work of modeling by grouping the wind farm units and extracting the extreme weather features, and it uses a multitype ensemble learning framework to enhance the prediction accuracy and adaptability for different faults. We verified our prediction scheme with experimental cases. It showed that our stacking model improved its ability to integrate multiple

learners and adapt to complex multidimensional extreme weather scenarios by introducing a feature weighting mechanism. Our model outperformed single models in prediction accuracy and reduced the calculation error of wind power fault prediction caused by strong fluctuation data under extreme weather conditions. Implementing effective and high-precision prediction of wind farm failures under extreme weather conditions provides support for improving the level of safe operation of the power grid under extreme scenarios.

For future research, we aim to verify the wind turbine fault prediction results under extreme weather and to apply our method to other renewable energy equipment operation predictions. We also plan to study the medium and long-term operation evaluation under the mixed action of multiple fault factors and to promote the implementation of more general algorithms and applications.

Nomenclature

SVM: Support vector machine
RF: Random forest
XGBoost: Extreme gradient boosting
LightGBM: Light gradient boosting machine
NWP: Numerical weather prediction
DBSCAN: Density-based spatial clustering of applications with noise
ML: Machine learning
GBDT: Gradient augmented decision tree
Eps: Epsilon
MinPT: Minimum Points
AICC: The average intracluster correlation coefficient
Cov: Covariance
Var: Variance.

Data Availability

The data used to support the findings of this study were supplied by the Dispatching Control Center of Guangxi Power Grid under license and so cannot be made freely available. Requests for access to these data should be made to Ze Chen, 2729174391@qq.com.

Conflicts of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Authors' Contributions

LL and YZ were responsible for the specific work of this manuscript. ZC and KW carried out some of the calculation work. HW and ZL guided the work of this manuscript.

Acknowledgments

References

[1] L.-L. Li, X. Zhao, M.-L. Tseng, and R. R. Tan, "Short-term wind power forecasting based on support vector machine with improved dragonfly algorithm," *Journal of Cleaner Production*, vol. 242, Article ID 118447, 2020.

[2] J. Wang, H. Zhu, Y. Zhang, F. Cheng, and C. Zhou, "A novel prediction model for wind power based on improved long short-term memory neural network," *Energy*, vol. 265, Article ID 126283, 2023.

[3] H. Zhang, L. Yongqian, J. Yan, S. Han, L. Li, and Q. Long, "Improved deep mixture density network for regional wind power probabilistic forecasting," in *IEEE Transactions on Power Systems*, vol. 35, no. 4, pp. 2549–2560, 2020.

[4] O. Krogsaeter and J. Reuder, "Validation of boundary layer parameterization schemes in the weather research and forecasting model under the aspect of offshore wind energy applications—Part I: Average wind speed and wind shear," *Wind Energy*, vol. 18, no. 5, pp. 769–782, 2015.

[5] Zhuochun Wu, Xiangjie Xia, Liye Xiao, and Yilin Liu, "Combined model with secondary decomposition-model selection and sample selection for multi-step wind power forecasting," *Applied Energy*, vol. 261, Article ID 114345, 2020.

[6] M. Aslam and M. Albassam, "Forecasting of wind speed using an interval-based least square method," *Frontiers in Energy Research*, vol. 10, Article ID 469, 2022.

[7] C. Cai, J. Guo, X. Song et al., "Review of data-driven approaches for wind turbine blade icing detection," *Sustainability*, vol. 15, no. 2, Article ID 1617, 2023.

[8] J. Xu, W. Tan, and T. Li, "Predicting fan blade icing by using particle swarm optimization and support vector machine algorithm," *Computers & Electrical Engineering*, vol. 87, Article ID 106751, 2020.

[9] L. Zhang, K. Liu, Y. Wang, and Z. B. Omariba, "Ice detection model of wind turbine blades based on random forest classifier," *Energies*, vol. 11, no. 10, Article ID 2548, 2018.

[10] T. Tao, Y. Liu, Y. Qiao et al., "Wind turbine blade icing diagnosis using hybrid features and stacked-XGBoost algorithm," *Renewable Energy*, vol. 180, pp. 1004–1013, 2021.

[11] M. Tang, Z. Peng, and H. Wu, "Fault detection for pitch system of wind turbine-driven doubly fed based on IHHO-LightGBM," *Applied Sciences*, vol. 11, no. 17, Article ID 8030, 2021.

[12] Y. Liu, S. Li, P. W. Chan, and D. Chen, "On the failure probability of offshore wind turbines in the China coastal waters due to typhoons: a case study using the OC4-deepcwind semisubmersible," *IEEE Transactions on Sustainable Energy*, vol. 10, pp. 522–532, 2019.

[13] F. Li, H. Cui, H. S. Iderchuluun, Z. Ma, Y. Zhu, and Y. Zhang, "Icing condition prediction of wind turbine blade by using artificial neural network based on modal frequency," *Cold Regions Science and Technology*, vol. 194, Article ID 103467, 2022.

[14] X. Wang, Z. Zheng, G. Jiang, Q. He, and P. Xie, "Detecting wind turbine blade icing with a multiscale long short-term memory network," *Energies*, vol. 15, no. 8, Article ID 2864, 2022.

[15] Z. Zhiqiang, F. Bin, L. Yong, Z. Peng, W. Jianguo, and W. Du, "Rapid warning of wind turbine blade icing based on MIV-tSNE-RNN," *Journal of Mechanical Science and Technology*, vol. 35, no. 12, pp. 5453–5459, 2021.

[16] L. Gao, T. Dasari, and J. Hong, "Wind farm icing loss forecast pertinent to winter extremes," *Sustainable Energy Technologies and Assessments*, vol. 50, Article ID 101872, 2022.

[17] L. Swenson, L. Gao, J. Hong, and L. Shen, "An efficacious model for predicting icing-induced energy loss for wind turbines," *Applied Energy*, vol. 305, Article ID 117809, 2022.

[18] S. T. Hallowell, A. T. Myers, S. R. Arwade et al., "Hurricane risk assessment of offshore wind turbines," *Renewable Energy*, vol. 125, pp. 234–249, 2018.

[19] A. Ghaemi, A. Safari, H. Afsharirad, and H. Shayeghi, "Situational awareness and deficiency warning system in a smart distribution network based on stacking ensemble learning," *Applied Soft Computing*, vol. 128, Article ID 109427, 2022.

[20] P. Waqas Khan and Y.-C. Byun, "Multi-fault detection and classification of wind turbines using stacking classifier," *Sensors*, vol. 22, no. 18, Article ID 6955, 2022.

[21] L. Breiman, "Stacked regressions," *Machine Learning*, vol. 24, no. 1, pp. 49–64, 1996.

[22] M. N. Akram and S. Lotfifard, "Modeling and health monitoring of DC side of photovoltaic array," *IEEE Transactions on Sustainable Energy*, vol. 6, no. 4, pp. 1245–1253, 2015.

[23] A. Abdellatif, H. Mubarak, S. Ahmad et al., "Forecasting photovoltaic power generation with a stacking ensemble model," *Sustainability*, vol. 14, no. 17, Article ID 11083, 2022.

[24] H. Hafdaoui, E. A. K. Boudjelthia, A. Chahtou, S. Bouchakour, and N. Belhaouas, "Analyzing the performance of photovoltaic systems using support vector machine classifier," *Sustainable Energy, Grids and Networks*, vol. 29, Article ID 100592, 2022.

[25] G.-F. Fan, L.-Z. Zhang, M. Yu, W.-C. Hong, and S.-Q. Dong, "Applications of random forest in multivariable response surface for short-term load forecasting," *International Journal of Electrical Power & Energy Systems*, vol. 139, Article ID 108073, 2022.

[26] J. Fang, H. Wang, F. Yang, K. Yin, X. Lin, and M. Zhang, "A failure prediction method of power distribution network based on PSO and XGBoost," *Australian Journal of Electrical and Electronics Engineering*, vol. 19, no. 4, pp. 371–378, 2022.

[27] P. Tao, H. Shen, Y. Zhang, P. Ren, J. Zhao, and Y. Jia, "Status forecast and fault classification of smart meters using LightGBM algorithm improved by random forest," *Wireless Communications and Mobile Computing*, vol. 2022, Article ID 3846637, 11 pages, 2022.

[28] H. Pingping, X. Yu, Z. Yan, and L. Kui, "Equivalent model of wind farm based on DBSCAN," in *2017 IEEE Innovative Smart Grid Technologies—Asia (ISGT-Asia)*, pp. 1–6, IEEE, Auckland, New Zealand.

[29] M. Pei, L. Ye, Y. Li et al., "Short-term regional wind power forecasting based on spatial-temporal correlation and dynamic clustering model," *Energy Reports*, vol. 8, pp. 10786–10802, 2022.

[30] Z. Lin, F. Wen, Y. Ding, and Y. Xue, "Data-driven coherency identification for generators based on spectral clustering," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 3, pp. 1275–1285, 2018.

[31] N. Demir and G. Dalkiliç, "Modified stacking ensemble approach to detect network intrusion," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 26, no. 1, pp. 418–433, 2018.

[32] Y. Wang, Y. Liu, L. Li, D. Infield, and S. Han, "Short-term wind power forecasting based on clustering pre-calculated CFD method," *Energies*, vol. 11, no. 4, Article ID 854, 2018.