Wind Turbine Fault Diagnosis Based on Feature Selection and Stacking Model Fusion with Small-scale Data

Na Liu , Jinxing Che*, Yu Ye

Abstract—**Wind energy, as new energy, becomes important support in the low-carbon transformation of the power industry. However, in wind farms, wind turbine fault diagnosis based on small-scale data is a thorny problem. To this end, this paper advances a wind turbine fault diagnosis method given feature selection and stacking model fusion. For unbalanced data, smote oversampling method is used to the effective instance of small class data and increase its proportion. RFECV is used to rank the importance of features, and then the features with high correlation are deleted according to the thermal map to reduce the dimension and obtain the feature subset. Then, XGBoost and LightGBM models based on 6-fold cross-validation is used to train the filtered data. To further improve the stability and generalization ability, a stacking fusion model based on logistic regression is trained using logistic regression. The results of this experiment are compared to other traditional methods by accuracy, ROC, and other indicators. The experimental results show that the strategy used in this paper has more satisfactory accuracy results than the traditional methods and can be used in wind turbine fault diagnosis engineering.**

Keywords—**unbalanced data, smote, RFECV, SCADA, model fusion**

I. Introduction

CLIMATE change is a major challenge related to human survival and long-term social development. At this stage, China's goal for climate change is to achieve carbon neutrality. The carbon emissions from fossil energy activities in China account for about 85% of the total carbon emissions, so the low-carbon transformation of energy is the key to carbon neutralization [1]. At present, as new energy, wind energy is an important support in the low-carbon transformation of the power industry. By the first half of 2020, China's installed wind power capacity had exceeded 216.75 million KW. In terms of proportion, the proportion of China's cumulative installed capacity is generally on the rise, ranking among the highest in the world [2]. At the same time, the safe operation and maintenance of wind turbines have become more and more important, which requires us to invest a lot of humans, material, and financial resources in this issue. Doing a good job of wind turbine fault diagnosis will greatly reduce the cost of the wind farm.

The research on wind turbine fault diagnosis mainly focuses on the fault prediction method model based on fault physical model, data-driven model, and fusion analysis using the characteristics of these two models [3]. Based on the fault physical model, the remaining life of the unit equipment is estimated based on an in-depth understanding of the operation mechanism of the unit equipment. The data-driven method is to establish a model based on historical data and life data for fault diagnosis and analysis. At present, a large number of scholars begin to analyze the information collected in the SCADA system for fault diagnosis. SCADA system is a special system for monitoring and collecting real-time data of wind turbines. It can contain characteristic information related to the operation of the wind turbine, including wind speed, temperature, current, power, and voltage. For data processing work, if we want to predict whether the wind turbine fails, we need to obtain some effective information from the information collected by the SCADA system to facilitate us to predict the failure. Guo et al. observed and diagnosed the fan fault by analyzing the changing trend of temperature [4]. Wang et al. constructed a data-driven fault diagnosis method for wind turbines by using discrete entropy (Artemide) and the improved time-shift multiscale fluctuation of cosine pairwise constrained supervised manifold mapping (cpcsmm) [5]. Dong et al. calculated the accompanying changes in blade icing performance characteristic parameters by studying the phases between the transmitted energy of wind turbine blades in various stages and the parameters of characteristic information recorded in SCADA [6]. Zhang et al. used a weighted network and meta-network cloned from the original RESNET to solve the wind turbine fault diagnosis problem [7]. Pandit used a Gaussian processing algorithm based on the fan blade state variable evaluation operation curve to detect blade fault [8]. However, for the case of a large amount of data, the traditional algorithm will have some shortcomings, such as a long convergence process, slow operation, and low accuracy of training results.

Manuscript received December 7, 2021; revised October 24, 2022.

The research is supported by the Jiangxi Provincial Education Department (Program No. GJJ190961), the National Natural Science Foundation of China (Grant No. 71971105 and 12161058), the National Statistical Science Research Project of China (Grant No. 2020LZ03), and the Jiangxi Provincial Natural Science Foundation (Program No. 20212BAB201020).

Na Liu is a postgraduate student of Nanchang Institute of Technology, Nanchang 330099, Jiangxi, China. (e-mail: 18300275371@163.com).

Jinxing Che is an associate professor of Nanchang Institute of Technology, Nanchang 330099, Jiangxi, China (corresponding author, e-mail: jinxingche1@163.com).

Yu Ye is a postgraduate student of Nanchang Institute of Technology, Nanchang 330099, Jiangxi, China. (e-mail: 2980618232@qq.com).

The emergence and development of machine learning, and deep learning networks have greatly improved the shortcomings of traditional algorithms. Machine learning and deep learning methods play an extremely effective role in predicting whether faults occur, and are widely used by researchers, and have been proved to be very effective when there are many parameters and a large amount of data [9-15]. In the state detection field of machine and deep learning methods, in Zhao et al.'s literature, the deep automatic coding network (DAE) is regarded as an index to detect the working state of the wind turbine, and then restructure the error between the input value and the output value, to monitor the effective operation of the wind turbine [16]. Chen et al. defined a new comprehensive index that can evaluate the performance of different algorithms, and proposed a new migration learning algorithm to deal with data imbalance and different distribution [17]. Wang et al. contributed an idea of fan blade fault detection using the SCADA data depth automatic encoder (DA) model [18]. Xiang et al. contributed a fault diagnosis idea using an attention mechanism to assign weight to the features of LSTM, and finally, infer the fault component according to the residual [19]. Li et al. used transfer learning based on shared parameters to take the operation data like a wind turbine as the training data for fault detection, in order to solve the problem of insufficient data of wind turbines [20]. To solve the problem of data imbalance, Tong et al. adopted the strategy of combining adaptive weighting with traditional fixed weighting and constructed an adaptive weighted kernel limit learning machine algorithm [21]. Afrasiabi et al. proposed a wind turbine fault identification method using a generation countermeasure network (GAN) to extract features and a time convolution neural network (TCNN) to classify faults [22]. Literature first analyzed the SCADA data of wind turbines, then obtained the reconstructed value of SCADA data through DAE and established an XGBoost multi-classification fault identification model [23]. Reference used out-of-pocket estimation of random forest for feature selection and established a diagnostic model based on the XGBoost algorithm [24]. Zhang et al. proposed a new non-smooth signal processing technique in order to mitigate the background noise, which can interfere with the fault signals of the wind turbine bearings, and introduced a particle swarm algorithm to extract the periodic shock fault information of the wind turbine [25]. Taghinezhad et al. used an artificial neural network model to predict the turbine power profile of wind turbines, which is a very important aspect of the study of wind turbines [26]. Xiao et al. combined convolutional neural networks with recurrent neural networks, convolutional neural networks with long and short-term memory neural networks, convolutional neural networks with selected-pass recurrent units, and finally with convolutional neural networks to form four focal-loss-based cost-sensitive deep neural networks as the basic prediction models [27]. By analyzing the monitoring and data acquisition data, Korkos et al. proposed an adaptive neuro-fuzzy inference system, this can be used to detect failure problems of wind turbine blades in different combinations of parameters [28].

The desire for a large amount of data is generally difficult to achieve. For instance, for turbines that haven't been used for a year, the information stored in SCADA obviously can't meet the demand. In addition, there are some cases of SCADA data loss, especially in wind farms in remote areas. In these cases, the data information stored in SCADA is particularly unbalanced. For example, the proportion of fault type and normal type is very different, which seriously affects the accuracy of the research results. Deep learning and many other machine learning methods will also be deeply affected by the problem of data imbalance. The fault information of wind turbines provided by unbalanced and large characteristic data is very complex, which is needed in the process of model training and parameter adjustment. In this case, these methods may not contribute well. This motivates us to propose a new fault diagnosis method suitable for dealing with the above problems.

In the field of fault diagnosis research based on SCADA data, subjective feature selection and traditional machine learning methods may produce unsatisfactory results. In this paper, a fault diagnosis method for wind turbines based on feature selection and stacked model fusion is proposed. Firstly, oversampling of the small-scale data is performed to balance the proportion of fault categories. Next, correlation analysis is performed on the data features. Then, the 5-fold cross-validation of RFECV is combined with the implementation of feature importance ranking to extract valid features. Finally, the stacking model fusion method is used to fuse two strong learners, and the stability and generalization ability of the model is improved through logical regression. The results show that the operation effect of this method is better than that of the traditional machine learning method.

II. SCADA DATA PREPROCESSING

SCADA (supervisory control and data acquisition) is the abbreviation of data acquisition and monitoring system. It is responsible for real-time monitoring of the operation data of wind turbines in wind farms and can be stably transmitted to the terminal. The data is collected every 10 minutes, including wind speed, wind direction angle, gearbox temperature, blade angle, and power generation.

In this paper, SCADA is classified into normal and fault states based on warning data and status data. The abnormal data caused by communication signal and transmission equipment faults are then eliminated. This is done to reduce the interference of abnormal data in the model prediction results. When there is a large proportion difference between data categories, it will seriously affect the prediction of the model and cause meaningless classification results, because the prediction model built by unbalanced data will be more inclined to the label of multi-category samples.

A. Sampling method via SMOTE

In this paper, SMOTE (synthetic minority oversampling Technology) is used to synthesize a few class oversampling techniques [29]. Because the traditional random oversampling technology directly copies the samples of a few category labels, it is easy to cause the overfitting of classification results. SMOTE algorithm is improved on this basis. It randomly selects a sample $\hat{X}_i$ from the near-neighboring

minority samples of each minority sample X_i and establishes a link between the two samples. Then a sample is randomly selected from the connected samples and classified as a newly formed minority sample.

Smote algorithm, its main thought is to synthesize new samples by analyzing samples on a small scale and then integrate those new samples into the original data set. The smote process is just as follows:

First, regarding per sample X, the Euclidean distance from each minority sample to the other minority samples is calculated.

$$d = \sqrt{\sum_{i=1}^{N} \left(X_{1i} - X_{2i} \right)^2} \tag{1}$$

Second, the sampling ratio N is determined according to the set unbalanced sampling proportion of samples, and samples $\hat{X}_i$ are randomly selected from the nearest neighbor minority samples of each minority sample X_i.

Third, replace the original samples $\hat{X}_i$ in the previous steps, which are obtained by random selection, into the following formula to obtain new samples.

$$X_{\text{new}} = x + rand(0,1) \times (\hat{X} - X) \tag{2}$$

B. Feature Selection

For the data collected by wind farm SCADA, there are many kinds of features, including features related to normal and abnormal categories, irrelevant features, and redundant features related to category labels. We can carry out feature selection according to the unit characteristics and the correlation analysis between features. We can also carry out feature selection according to the machine learning method. Compared with the former, the latter is more efficient and can fully analyze the relationship between features and category labels to retain the features with the best performance. In order to get the number of features that can make the experiment produce the best results, we will use rfecv, which is recursive feature elimination combined with cross-validation.

RFECV is divided into two stages [30]. The flow chart is shown in Figure 1.

The first stage is RFE: recursive feature elimination and feature importance ranking.

The second stage is CV: cross-validation. On the premise that the importance of features has been sorted, the number of features with the best performance is selected by cross-validation.

RFE, the full name is recursive feature elimination. Its core is to build the model repeatedly so as to find the optimum functionality. At the beginning of each model construction, first, delete the last optimal feature, then build the model in the remaining features, and continue to select the optimal feature. The first deleted feature is the most important feature.

The specific steps of recursive feature elimination are:

① All features are initial features.

② Modeling on the current data set through the prediction model (random forest classifier is used in this paper), and each feature is given a weight.

③ Then delete the least important features in the feature set.

④ Execute step ② again, to perform recursively.

Finally, the importance of all features will be evaluated.

Cross-validation is used to prevent overfitting caused by two complex models [31]. Sometimes called cyclic estimation, it is a practical method that can statistically cut data samples into smaller subsets. In this way, a subset can be analyzed first, and the remaining subsets can be used as the subsequent confirmation and verification of this analysis. This method can be explained by the following formula.

$$CV(f,\alpha) = \frac{1}{N} \sum_{i=1}^{N} L\left(y_i, \hat{f}^{-k(i)}\left(x_i, \alpha \right) \right) \tag{3}$$

The function $CV(f,\alpha)$ provides an estimate of the test error curve, and we find the tuning parameter α that minimizes it.

Specific steps of cross-validation: ① according to the importance of the features obtained by recursive feature elimination, take out the data sets with a different number of features each time. ② cross-validate these samples with the different number of features respectively. ③ finally select the number of features with the highest average score, that is, the feature with the best performance.

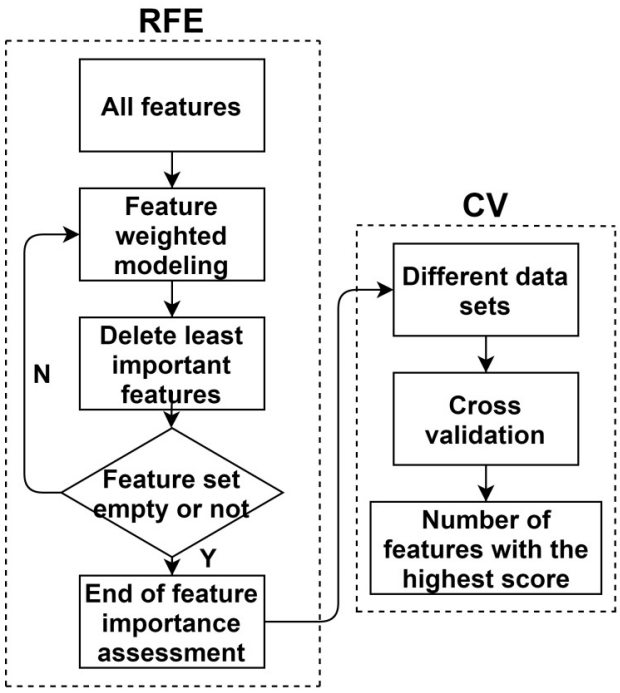

Fig. 1. Flow chart of RFECV

III. Fault Diagnosis Model

A. Stacking Model Fusion

Stacking is a hierarchical structure as shown in Figure 2. The first layer is the base layer, which contains the prediction results of several base strength learners. The second layer is a

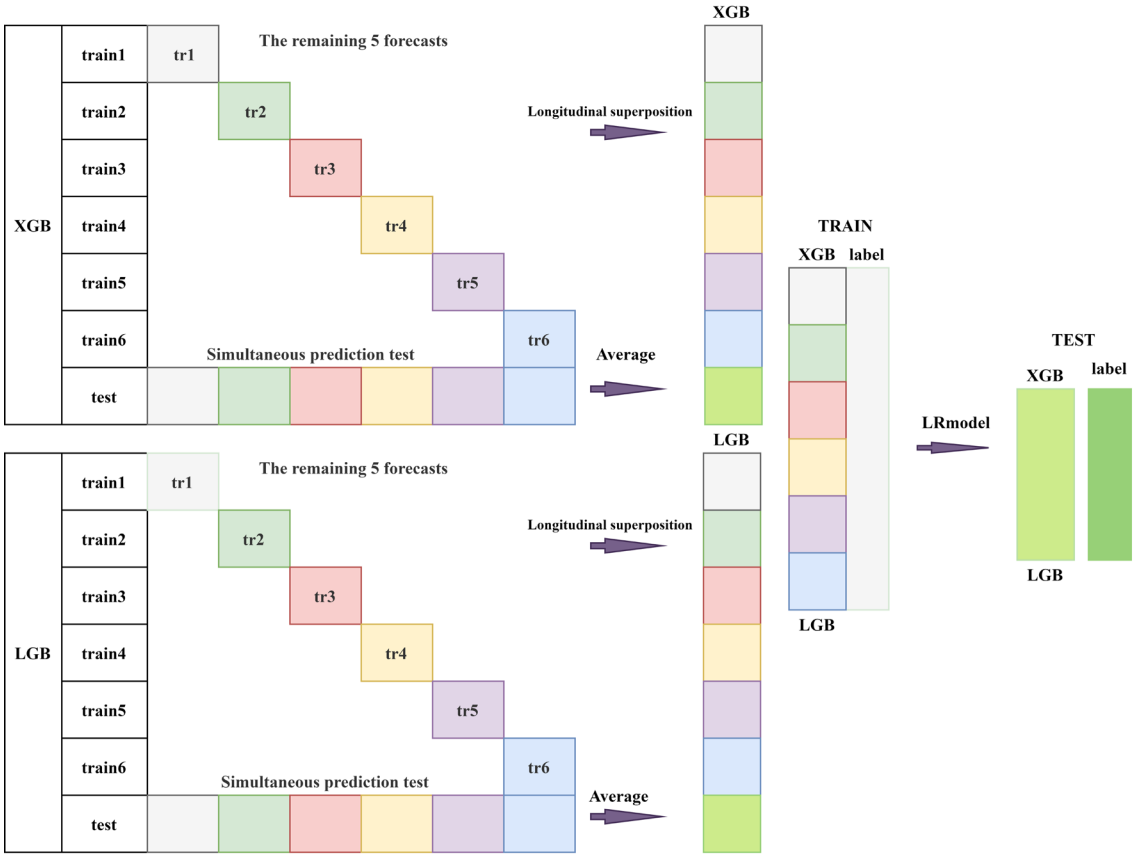

Fig. 2. Model stacking process

model that can integrate the operation results of the grassroots level, usually a weak learner. Stacking is to integrate the operation results of the grass-roots level into sample characteristics, and the original sample is regarded as a new data set to divide a new training set, to train a new model and predict the sample.

Because each learner in the basic level will predict the training set and test set, in the second layer, if they directly integrate the predicted training set and test set as the training set and test set of weak learners, it is very easy to produce overfitting. In this paper, the k-fold cross-validation method is used to deal with this problem. At the grassroots level, cross-validation is used for prediction. The basic learners cross verifies the feature list of the exercise set and test set of the training relative to the exercise set and test set of the second layer model respectively.

B. XGBOOST

XGBoost algorithm belongs to the boosting framework. The essence of the boosting framework algorithm is that the definition of gain is different when fitting the residual tree in each round [32]. The gain used by XGBoost is the difference between the structural score before splitting and the structural score after splitting. One of the highlights of XGBoost is to define a splitting criterion so that the loss of each optimal split point can be reduced the most than that without splitting. This is one of the reasons why XGBoost is efficient. The meaning of structural score: the minimum value of the loss function when the structure of the tree is known. XGBoost's gain definition is to subtract the structural score before splitting from the structural score after splitting and select the segmentation point with the largest gain as the optimal segmentation point. Its significance is the segmentation point where the loss of the model after splitting is largest than that before splitting. The result of this round of residual trees fitted by this gain definition method is very good.

Model expression in iteration k :

$$f^{(k)}(x_i) = f^{(k-1)}(x_i) + h^{(k)}(x_i) \tag{4}$$

Where $h^{(k)}(X_i)$ represents the residual tree of the kth round of fitting.

The loss function of the model in the round k is:

$$L^{(k)} = \sum_{i=1}^{n} L\left(y_i - f^{(k)}(X_i)\right) + \Omega\left(h^k(X_i)\right) \tag{5}$$

Where $\Omega\left(h^k(X_i)\right)$ represents the regularization term of the round k .

$$\Omega\left(h^k(X_i)\right) = \gamma T + \frac{1}{2}\lambda * \sum_{j=1}^{n} w_j^2 \tag{6}$$

The Taylor second-order expansion expression of the loss function is:

$$L^k = \sum_{i=1}^{n} \begin{pmatrix} L\left(y_i - f^{(k)}\left(x_i\right)\right) \\ +g_i * h^{(k)}\left(x_i\right) \\ +\frac{1}{2} h_i * \left(h^{(k)}\left(x_i\right)\right)^2 \end{pmatrix} + \Omega\left(h^k\left(x_i\right)\right) \tag{7}$$

g_i represents the first derivative of the i-th sample, also known as the first-order residual. h_i represents the second derivative of the i-th sample, also known as the second-order residual.

Because the meaning of the structure score is the minimum value of the loss function when the structure of the tree is known. Our purpose is to require the minimum value of the loss function. In the process of finding the extreme value, the constant term 1 can be removed and simplified as:

$$L^k = \sum_{i=1}^{n} \begin{pmatrix} g_i * h^{(k)}\left(x_i\right) \\ +\frac{1}{2} h_i * \left(h^{(k)}\left(x_i\right)\right)^2 \end{pmatrix} + \Omega\left(h^k\left(x_i\right)\right) \tag{8}$$

C. LightGBM

LightGBM algorithm belongs to the boosting framework. The essential difference between the algorithms of boosting framework is how to fit a residual tree in each round of iteration [33].

The innovation of the LightGBM algorithm:

①Introduce the Goss algorithm to eliminate samples with small weight and reduce the amount of data before fitting the residual tree.

② Then EFB algorithm is introduced to bind mutually exclusive features in the case of high-dimensional data, so as to reduce the number of features.

③ The discrete values are processed so that the input discrete values are directly supported by the model.

④In the process of fitting the residual tree, the sample leaf-wise method is used to reduce the number of splitting nodes, so as to reduce the amount of calculation.

⑤The histogram method is used to improve the speed of finding the optimal segmentation point.

The objective function of LightGBM is

$$L_{\mathrm{n}} = \sum_{i=1}^{n} l\left(y^i, y_{n-1}^i + f_n\left(x^i\right)\right) + \Upsilon T + \frac{1}{2}\lambda \sum_{j=1}^{T} W_j^2 \tag{9}$$

After dividing node O into two parts with a certain segmentation point, the gain, in this case, is: the variance of node O before segmentation minus the sum of the variance of two child nodes after segmentation

$$\begin{aligned} & p(O)\,\mathrm{Var}(y|O) \\ & -p(Left)Var(y|Left) - p(Right)\,\mathrm{Var}(y|Right) \\ & = \frac{1}{N}\left[\left(\sum_{i:\bar{X}_i \in L} g_i\right)^2 / n_{l|O}(d)\right. \\ & \left. + \left(\sum_{i:\bar{X}_i \in R} g_i\right)^2 / n_{r|O}(d) + \left(\sum_{i:\bar{X}_i \in O} g_i\right)^2 / n_O\right] \end{aligned} \tag{10}$$

Because our goal is to compare the gain, we can remove $\frac{1}{N}$ constant terms $\frac{\left(\sum_{i:\bar{X}_i \in O} g_i\right)^2}{n_O}$ that have no effect on the comparison.

So again, it is obtained:

$$\mathrm{Gain} = \left[\frac{\left(\sum_{i:\bar{X}_i \in L} g_i\right)^2}{n_{l|O}(d)} + \frac{\left(\sum_{i:\bar{X}_i \in R} g_i\right)^2}{n_{r|O}(d)}\right] \tag{11}$$

In Goss, in the process of dividing node O, some samples may be discarded randomly, so the total number of samples of node O decreases. Therefore, redefining the variance gain:

$$\mathrm{Gain} = \frac{1}{n_O}\left[\frac{\left(\sum_{i:\bar{X}_i \in L} g_i\right)^2}{n_{l|O}(d)} + \frac{\left(\sum_{i:\bar{X}_i \in R} g_i\right)^2}{n_{r|O}(d)}\right] \tag{12}$$

n_O represents the total number of samples of the O node, $n_{l|O}$ represents the total number of samples of the left child node of the O node, and $n_{r|O}$ represents the total number of samples of the right child node of the O node.

The fitting process of LightGBM in the residual tree is to draw their respective histograms for the features of each node before node splitting, and then calculate the respective gain of each feature according to the histogram. This gain is divided according to different value ranges. This method of using the histogram to calculate the gain is to discretize the value, which can greatly improve the efficiency of selecting the optimal segmentation point in the later work. After that, the optimal splitting point of the leaf nodes is divided. Considering that LightGBM itself adopts a leaf-by-leaf growth mode, the splitting point with the greatest gain among the leaf nodes is considered as the optimal splitting point. This can be more efficient and achieve higher accuracy.

D. Modeling Process

This paper realizes wind turbine fault diagnosis based on smote, rfecv, and stacking model fusion as shown in Figure 3. The specific steps are as follows:

① In the data pre-processing stage, SCADA data are screened to eliminate invalid data that interfere with the prediction results. Then a sampling oversampling is performed on a small scale of data to prevent meaningless classification results caused by data imbalance.

②After preprocessing, the recursive feature elimination and cross-validation RFECV algorithm based on a random

forest classifier is used. In this paper, 6-fold cross-validation is used.

③After feature selection, divide the data into a training set and test set according to the ratio of 0.67:0.33.

④The stacking model fusion method is used to model the training set. The basic learner uses XGBoost and LightGBM. The second layer uses logical regression to fuse the results obtained by the basic learner to improve the stability. Meanwhile, 6-fold cross-validation is used to prevent overfitting during model training.

⑤Test the prediction results according to the test set. Then, view the classification results through the confusion matrix. Test according to the model performance measurement indicators such as accuracy, AUC, recall, and balance score F1. The experimental results are compared with gradient lifting tree, XGBoost, and LightGBM.

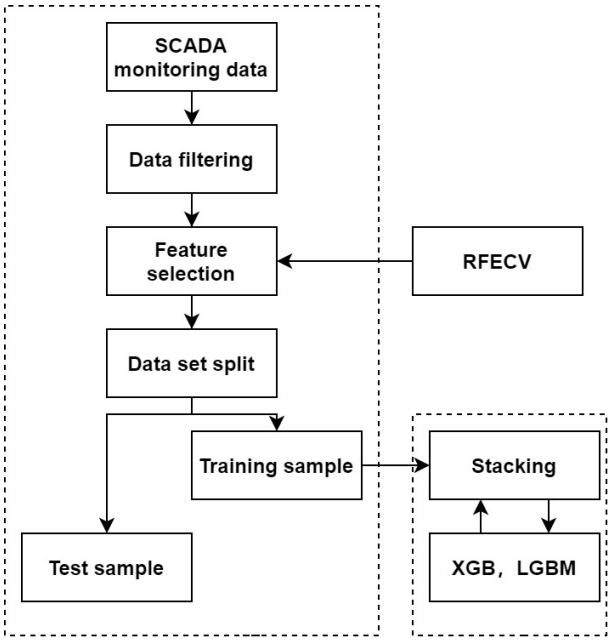

Fig. 3. Flowchart of fault detection for a wind turbine

IV. Case Analysis

A. Over-sampling

This paper uses the status data, warning data, and SCADA data of a foreign wind farm from May 1, 2014, to April 9, 2015. The SCADA data contains 49026 records and 61 features. As shown in Table I.

TABLE I
CHANGE OF SAMPLE SHAPE COUNTER THROUGH SMOTE

	Normal	Failure
Proportion	95.84%	4.16%
Original dataset shape Counter	6869	298
Resampled dataset shape Counter	6869	6869

After feature preprocessing, 6869 normal samples and 298 fault samples are retained. After smote, the number of normal samples remains unchanged, and the fault samples become 6869.

B. Feature Selection

Through RFECV feature selection, it can be seen that when the number of selected features is 48, the performance reaches the best.

After correlation analysis combined with thermal diagram, this paper will select 46 features for later model training.

As shown in Figure 4, the 48 features are sorted according to their importance, and 46 of them are selected as representatives in combination with the correlation thermodynamic diagram between the features. These characteristics are shown in Table II.

In machine learning, a confusion matrix is actually used to calculate the classification error [34]. The confusion matrix allows researchers to intuitively understand the effectiveness of the classification algorithm they use. A confusion matrix is an n-dimensional square matrix, and the dimension of the square matrix represents the category. The row coordinates and column coordinates of the matrix can represent reality and prediction respectively.

Through the confusion matrix, it is easy to see whether the system will confuse the two classes, which is also the origin of the name of the confusion matrix. The stacking model fusion is used to train the data after feature selection, and the normal data and fault data are classified. The confusion matrix is shown in Figure 5.

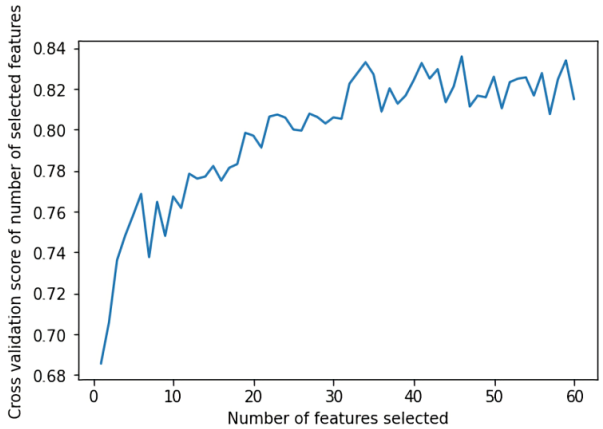

Fig. 4. RFECV via automatic tuning of the number of features selected with cross-validation

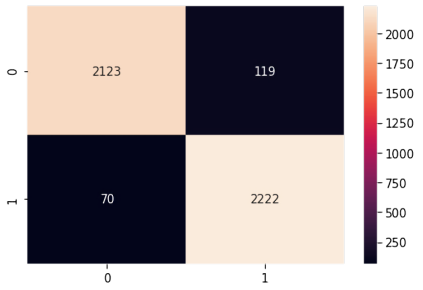

Fig. 5. Confusion matrix of normal and fault samples

TABLE II
THE FEATURE DESCRIPTION OF THE WIND TURBINE

No.	Feature	Units	No.	Feature	Units
1	ava. windspeed	m/s	25	Sys 2 inverter 2 cabinet temp	℃
2	max. windspeed	m/s	26	Sys 2 inverter 3 cabinet temp	℃
3	min. windspeed	m/s	27	Sys 2 inverter 4 cabinet temp	℃
4	ava. Rotation	r/s	28	Front bearing temp	℃
5	max. Rotation	r/s	29	Rear bearing temp	℃
6	min. Rotation	r/s	30	Pitch cabinet blade A temp	℃
7	ava. Power	Kw	31	Pitch cabinet blade B temp	℃
8	max. Power	Kw	32	Pitch cabinet blade C temp	℃
9	min. Power	Kw	33	Rotor temp. 1	℃
10	ava. Nacelle position including cable twisting	°	34	Rotor temp. 2	℃
11	Operating	Hours	35	Stator temp. 1	℃
12	Production	kWh	36	Stator temp. 2	℃
13	Production	minutes	37	Nacelle ambient temp. 1	℃
14	ava. reactive Power	Kw	38	Nacelle ambient temp. 2	℃
15	max. reactive Power	Kw	39	Nacelle temp	℃
16	min. reactive Power	Kw	40	Nacelle cabinet temp	℃
17	ava. available P from wind	Kw	41	Main carrier temp	℃
18	ava. available P technical reasons	Kw	42	Rectifier cabinet temp	℃
19	ava. Available P force majeure reasons	Kw	43	Yaw inverter cabinet temp	℃
20	ava. Available P force external reasons	Kw	44	Fan inverter cabinet temp	℃
21	ava. blade angle A	°	45	Ambient temp	℃
22	Sys 1 inverter 3 cabinet temp	℃	46	Tower temp	℃
23	Sys 1 inverter 6 cabinet temp	℃	47	Control cabinet temp	℃
24	Sys 2 inverter 1 cabinet temp	℃	48	Transformer temp	℃

C. Model evaluation

As can be seen from the classification report of the confusion matrix in Table III, where 0 represents normal data and 1 represents fault data. The recall rate can explain the correctness of this category. The greater the recall rate, the better the recognition effect of this category. F1 score, also known as balanced f score, is defined as the harmonic average of accuracy and recall. The F1 score indicator combines the results of the outputs of precision and recall. F1 score is a measure of classification problems. The maximum is 1 and the minimum is 0.

$$F1 = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \tag{13}$$

Support indicates the practical number of samples belonging to a category. Micro avg does not distinguish between sample categories and calculates the overall accuracy, recall, and F1. Weighted average, as the name suggests, is to add weight to the macro average. Its weight refers to the ratio of the number of samples belonging to a certain category to the amount of all samples. It can be seen from the accuracy, recall, or F1 score that the classification results have reached a good level.

TABLE III
CLASSIFICATION REPORT OF THE CONFUSION MATRIX

	Precision	Recall	F1 score	Support
0	0.97	0.95	0.96	2242
1	0.95	0.97	0.96	2292
Accuracy			0.96	4534
Macro avg	0.96	0.96	0.96	4534
Weighted avg	0.96	0.96	0.96	4534

TABLE IV
COMPARISON OF EVALUATION INDEXES OF EACH MODEL

COMPARISON OF RESULTS				
index	roc_acu	precision	recall	f1_score
stacking	0.99044	0.94916	0.96945	0.95920
XGBoost	0.98821	0.93797	0.96989	0.95366
LightGBM	0.98908	0.93813	0.97251	0.95501
GBDT	0.97988	0.91779	0.94502	0.93121

Finally, the accuracy of stacking model fusion experiment results is compared with XGBoost, LightGBM, and GBDT, as shown in the following Table IV.

We can see whether, from AUC, precision, recall, or F1, the classification scores after the fusion of score and stacking model are better than the other three classifiers. Although in recall the score of LightGBM on the score is better than that of stacking model fusion. However, after comprehensive consideration, for the AUC index that can best represent the classifier performance, the AUC value of stacking is significantly greater than that of XGBoost, LightGBM, and GBDT. Therefore, in this paper, the classifier performance after the integration of the stacking model will be better.

V. SUMMARY

This thesis advances a stacking wind turbine fault diagnosis algorithm on account of feature selection, which uses the actual detection data of wind farm SCADA to realize fault diagnosis. Aiming at the problem of unbalanced data of sample labels, smote oversampling algorithm is adopted to achieve data balance. Recursive feature elimination and 5-fold cross-validation RFECV algorithm are used to select features with high performance according to the importance of features. Finally, using the stacking model fusion method, the training sets are trained by two basic learners respectively, and the 6-fold cross-validation is used. Then these training sets and test sets are fused by the logistic regression LR algorithm in the second layer. Experiments show that the accuracy of stacking is 0.99044, which is higher than that of other single learners. Wind turbine fault diagnosis on account of stacking model fusion provides new thinking for wind turbine fault diagnosis through data mining and big data analysis.

REFERENCES

[1] Li Quansheng.Discussion on China';s energy transformation path under the goal of carbon neutralization [J]. China coal, 2021,47 (08): 1-7.

[2] Prospective industry research institute.2020Analysis on market status and development prospect of China's wind power industry in 2025 [EB / OL] (2020 10 29) [2021 10 5]https://bg.qianzhan.com/trends/detail/506/201029-59546d95.html

[3] PECHT M G . Prognostics and health management of electronics [M] . New Jersey: John Wiley & Sons, 2008 .

[4] P. Guo, D. Infield, and X. Yang, "Wind Turbine Generator Condition-Monitoring Using Temperature Trend Analysis," in IEEE Transactions on Sustainable Energy, vol. 3, no. 1, pp. 124-133, Jan. 2012.

[5] Zhenya Wang, Gaosong Li, Ligang Yao, et al, Data-driven fault diagnosis for wind turbines using modified multiscale fluctuation dispersion entropy and cosine pairwise-constrained supervised manifold mapping, Knowledge-Based Systems, vol 228,2021.

[6] Xinghui Dong, Di Gao, Jia Li, et al, Blades icing identification model of wind turbines based on SCADA data, Renewable Energy, vol 162, 2020,575-586.

[7] Kai Zhang, Baoping Tang, Lei Deng, et al, A fault diagnosis method for wind turbines gearbox based on adaptive loss weighted meta-ResNet under noisy labels, Mechanical Systems and Signal Processing, vol 161,2021.

[8] Pandit R, Infield D. Gaussian Process Operational Curves for Wind Turbine Condition Monitoring. *Energies*. 2018; 11(7):1631.

[9] X. Ding and Q. He, "Energy-Fluctuated Multiscale Feature Learning Deep ConvNet for Intelligent Spindle Bearing Fault Diagnosis," in IEEE Transactions on Instrumentation and Measurement, vol. 66, no. 8, pp. 1926-1935, Aug. 2017.

[10] Y. Gao, X. Liu and J. Xiang, "FEM Simulation-Based Generative Adversarial Networks to Detect Bearing Faults," in IEEE Transactions on Industrial Informatics, vol. 16, no. 7, pp. 4961-4971, July 2020.

[11] Xiaoyang Liu, Haizhou Huang, Jiawei Xiang, A personalized diagnosis method to detect faults in gears using numerical simulation and extreme learning machine, Knowledge-Based Systems, 195, 2020, 105653, https://doi.org/10.1016/j.knosys.2020.105653.

[12] Q. Gao, H. Tang, J. Xiang, et al., A Walsh transform-based Teager energy operator demodulation method to detect faults in axial piston pumps, Measurement 134 2019 293–306.

[13] Z. Pan, Z. Meng, Z. Chen, et al., A two-stage method based on extreme learning machine for predicting the remaining useful life of rolling-element bearings, Mech. Syst. Signal Process. 144 2020.

[14] Wei Zhang, Xiang Li, Qian Ding, Deep residual learning-based fault diagnosis method for rotating machinery, ISA Transactions, 95, 2019, 295-305, https://doi.org/10.1016/j.isatra.2018.12.025.

[15] W. Zhang, X. Li, Q. Ding, et al., Machinery fault diagnosis with imbalanced data using deep generative adversarial networks, Measurement 152 2020.

[16] H. Zhao, H. Liu, W. Hu, et al., Anomaly detection and fault analysis of wind turbine components based on deep learning network, Renew. Energy 127 (2018) 825–834.

[17] Wanqiu Chen, Yingning Qiu, Yanhui Feng, et al, Diagnosis of wind turbine faults with transfer learning algorithms, Renewable Energy, vol 163,2021.

[18] L. Wang, Z. Zhang, J. Xu, et al., Wind Turbine Blade Breakage Monitoring With Deep Autoencoders, IEEE Trans. Smart Grid 9 (4) (2018) 2824–2833.

[19] Ling Xiang, Penghe Wang, Xin Yang, et al, Fault detection of wind turbine based on SCADA data analysis using CNN and LSTM with attention mechanism, Measurement, vol 175, 2021

[20] Yanting Li, Wenbo Jiang, Guangyao Zhang, et al, Wind turbine fault diagnosis based on transfer learning and convolutional autoencoder with small-scale data, Renewable Energy, vol 171,2021

[21] Ruining Tong, Peng Li, Xun Lang, et al, A novel adaptive weighted kernel extreme learning machine algorithm and its application in wind turbine blade icing fault detection, Measurement, vol 185,2021

[22] S. Afrasiabi, M. Afrasiabi, B. Parang, M. Mohammadi, M. M. Arefi and M. Rastegar, "Wind Turbine Fault Diagnosis with Generative-Temporal Convolutional Neural Network," 2019 IEEE International Conference on Environment and Electrical Engineering and 2019 IEEE Industrial and Commercial Power Systems Europe (EEEIC / I&CPS Europe), 2019, pp. 1-5, doi: 10.1109/EEEIC.2019.8783233.

[23] Zhao Hongshan, Yan Xihui, Wang Guilan, Yin Xianglong. Fault diagnosis of wind turbine generator using deep self coding network and xgboost [J]. Power system automation, 2019,43 (01): 81-86.

[24] Jin Zhijie,Huo Zhihong, Xu Chang, Guo Hongyu, Zhou Huajian. Wind turbine fault diagnosis based on feature selection and xgboost [J]. Renewable energy, 2021,39 (03): 353-358

[25] Yi Zhang, Yong Lv, Mao Ge, Time–frequency analysis via complementary ensemble adaptive local iterative filtering and enhanced maximum correlation kurtosis deconvolution for wind turbine fault diagnosis, Energy Reports, 7, 2021, 2418-2435, https://doi.org/10.1016/j.egyr.2021.04.045.

[26] Javad Taghinezhad, Samira Sheidaei, Prediction of operating parameters and output power of ducted wind turbine using artificial neural networks, Energy Reports, 8, 2022, 3085-3095, https://doi.org/10.1016/j.egyr.2022.02.065.

[27] Jin Xiao, Chunyan Li, Bo Liu, Jing Huang, Ling Xie, Prediction of wind turbine blade icing fault based on selective deep ensemble model, Knowledge-Based Systems, 242, 2022, 108290, https://doi.org/10.1016/j.knosys.2022.108290.

[28] Panagiotis Korkos, Matti Linjama, Jaakko Kleemola, Arto Lehtovaara, Data annotation and feature extraction in fault detection in a wind turbine hydraulic pitch system, Renewable Energy, 185, 2022, 692-703, https://doi.org/10.1016/j.renene.2021.12.047.

[29] SMOTE: Synthetic Minority Over-sampling Technique.https://export.arxiv.org/pdf/1106.1813.

[30] sklearn.feature_selection.RFECV — scikit-learn 1.0.1 documentation

[31] Li Hang. Statistical learning methods [M]. Beijing: Tsinghua University Press, 2019. 23-24

[32] XGBoost: A Scalable Tree Boosting System. https://arxiv.org/pdf/1603.02754.pdf.

[33] LightGBM: A Highly Efficient Gradient Boosting Decision Tree (neurips. cc)

[34] Confusion matrix. In Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Confusion_matrix

Na Liu is studying for a master's degree in Nanchang Institute of Technology.

Jinxing Che received his B. S. degree from Jiujiang University in 2007 and his M. S. degree in applied mathematics from Lanzhou University in 2010, as well as his Ph. D. degree in mathematical statistics from Xidian University, China in 2019. He is currently an associate professor and Master's Supervisor in School of Science, Nanchang Institute of Technology, China. His main research interest is data analysis theory and application, hydrological information processing as well as prediction theory and method.

Yu Ye is studying for a master's degree in Nanchang Institute of Technology.