World Journal of Engineering Research
Print ISSN: 2959-9865
Online ISSN: 2959-9873
DOI: https://doi.org/10.61784/wjer3040

FAULT DIAGNOSIS OF WIND POWER SYSTEM BASED ON LSTM PREDICTION METHOD

Yuan Ma[1*], JingYu Zhao[2]
[1]State Grid Energy Research Institute Co., Ltd., Beijing 102209, China.
[2]State Grid Talents Exchange And Service Center Company Limited, Beijing 100089, China.
Corresponding Author: Yuan Ma, Email: ma_yuan@126.com

Abstract: As the core unit of the new energy grid, the operational reliability of the wind power system is directly related to the stability and power generation efficiency of the power grid, so it is urgent to improve its ability to accurately predict faults. Aiming at the complex timing fault characteristics of wind power generation systems, a multi-feature fusion operation fault prediction method based on LSTM is proposed. The wind farm SCADA system collects the operation data for preprocessing, screens the highly correlated features by the Pearson correlation coefficient method, and constructs a multi-feature input LSTM fault prediction model to improve the accuracy of wind power generation system operation fault prediction. Experimental results show that compared with the single feature model, the multi-feature fusion strategy can significantly improve the comprehensive performance of the prediction model, and the fault warning accuracy and F1 score are increased by 12.78% and 12.04% respectively.
Keywords: Power system failure; Wind power systems; Fault prediction; LSTM

1 INTRODUCTION

As the core support unit of the new energy grid, the operational reliability of the wind power generation system is directly related to the safety, stability and energy conversion efficiency of the power system. However, wind power equipment has been serving in complex natural environments for a long time, and the core components are easily induced by the coupling effect of wind speed, temperature and other factors, resulting in downtime losses and a surge in operation and maintenance costs. The traditional fault prediction method based on a single parameter is difficult to adapt to the temporal dynamic characteristics of multivariate coupling of wind power generation systems, and there are problems such as prediction lag and insufficient accuracy.

With the rapid development of deep learning algorithms, especially the modeling ability of long short-term memory (LSTM) for time series dependencies, it provides a new path for wind power system failure prediction. Therefore, this study proposes a fault prediction method for wind power generation system based on LSTM, which improves the accuracy and efficiency of fault prediction and ensures the reliable operation of new energy power grids by mining the correlation characteristics of multi-dimensional operation data.

2 FAULT PREDICTION OF WIND POWER SYSTEM

2.1 Wind Power System Fault Data Collection

The geographical dispersion of wind farm clusters has significantly increased the technical complexity of remote management of wind power systems, and the monitoring and data acquisition system (SCADA) has become the core infrastructure for onshore and offshore wind farms to realize digital and intelligent operation and maintenance, supporting remote status monitoring of wind turbines, The operation parameter regulation and full data collection can continuously generate SCADA data with multi-dimensional timing characteristics, covering multi-dimensional monitoring parameters such as power load dynamics, unit power generation efficiency, and fault event logs.

As a wind turbine data acquisition unit, the SCADA system has a wide range of monitoring data, mainly including wind speed, generator speed and other speed data; electrical data such as voltage, current, active power, and reactive power; temperature data such as generator stator and rotor temperature, engine room motor temperature, and ambient temperature; and wind direction, yaw angle, torque and other angle data. Table 1 shows the main parameters of the SCADA system monitoring variables.

Table 1 Main Monitoring Variables and Units of SCADA System

Parameter	Unit	Parameter	Unit
Wind speed	m/s	Wind turbine yaw state	/
Wind direction	°	Hub temperature	℃
Active power	KW	Tower bottom temperature	℃
Reactive power	KVar	10 min average turbulence intensity	m/s
Wind wheel speed	rpm	Turbulence coefficient	m/s

© By the Author(s) 2025, under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0).

Generator speed	rpm	Average torque	N·m
Generator bearing inner Ring temperature	℃	Pitch control cabinet temperature	℃
Yaw speed	m/s	Temperature of the outer ring of the power generation bearing	℃

In summary, the SCADA data generated in the operation of the wind power generation system has obvious spatio-temporal correlation laws, and the current state is restricted by the historical operation state, and the LSTM network model can effectively model long-term dependencies due to its unique gating mechanism and state memory ability.

2.2 Overview of Fault Prediction Methods

Fault prediction and diagnosis in wind power systems is key to ensuring grid stability and falls into three categories: statistical learning, machine learning, and deep learning methods.

Statistical learning methods mainly rely on rule-based expert systems and traditional statistical analysis methods, including the ARIMA model[1], exponential smoothing algorithm[2]. Because this traditional model is built on the basis of time series stability, and the operation data of wind power generation system shows strong time-varying and nonlinearity dynamic characteristics, its stability assumption is fundamentally in conflict with the internal law of the data. Common machine learning methods in power system fault prediction mainly include decision trees[3], support vector machines[4], random forest[5] etc. These methods can extract the characteristics of power system data such as high dimensionality, nonlinearity and complex distribution, but wind power faults are usually induced by the coupling of multiple physical parameters, it is difficult to fully describe the fault evolution law of single-dimensional features. Therefore, it is necessary to manually construct feature engineering, which makes it difficult to predict the fault of wind power generation system through machine learning methods. In contrast, deep learning methods can automatically extract complex features from massive data and capture deep correlations between data without manual intervention[6]. The most commonly used deep learning methods are LSTM[7], gated recurrent unit(GRU)[8], convolutional neural networks(CNN)[9], multi-layer perceptron(MLP)[10] etc.

Given that faults in wind power systems exhibit distinct long-term dynamic evolution characteristics, among several common deep learning algorithms, GRU simplifies its gating mechanism, leading to inadequate depiction of the complex dependencies inherent in wind power faults[11]. MLP as a static feedforward network, lacks the capability to address the temporal evolution characteristics of such faults. CNN excels in extracting spatial features but is not tailored for temporal dynamics. In contrast, the gating mechanism of LSTM can effectively mitigate the gradient vanishing issue in long sequences—a limitation of traditional recurrent neural networks—while accurately capturing cross-time-step dependencies. Moreover, it can directly incorporate temporal information of multi-dimensional features, thereby fulfilling the prediction needs of multi-feature fusion. Accordingly, this study presents an LSTM-based multi-feature fusion fault prediction method for wind power systems.

3 FAILURE PREDICTION MODEL OF WIND POWER SYSTEM BASED ON LSTM

3.1 Data Processing and Feature Analysis

Data in wind power systems primarily originates from SCADA systems, characterized by large volume and redundancy. This necessitates data cleaning to remove missing values and outliers, normalization and standardization to eliminate scale differences among distinct features, and feature selection to screen out features that significantly impact the operational fault prediction of wind power systems. Such processes remove redundant or irrelevant features, thereby reducing computational load and enhancing model training efficiency.

Based on this, the Pearson correlation coefficient method was used to screen the characteristics of SCADA data and analyze the correlation of the relevant factor sets. The Pearson correlation coefficient, a standardized form of covariance, is a statistic used to measure the strength and direction of linear relationships between variables, independent of variable units and scales, and is defined as follows:

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 (Y_i - \bar{Y})^2}} \tag{1}$$

where X_i sum Y_i is the first observation of the two variables i ; $\bar{X}$ and $\bar{Y}$ are the average of X the sums Y , respectively. The value range of the Pearson correlation coefficient is $[-1,1]$ as follows, $r > 0$ indicating the existence of a positive correlation and $r < 0$ indicating the existence of a negative correlation, and its absolute value is used to characterize the degree of correlation. r The degree of correlation is shown in Table 2.

Table 2 Pearson Correlation Determination Table

Correlation degree	Irrelevant or	Weak	Moderately	Strongly	Extremely

	extremely weak				strong
Absolute value of correlation coefficient	0-0.2	0.2-0.4	0.4-0.6	0.6-0.8	0.8-1

3.2 LSTM Model Design

Long short-term memory networks are developed from recurrent neural networks (RNN), which introduce four parts: memory cells, input gates, forgetting gates, and output gates to regulate the flow of information, which can not only overcome the problems of gradient explosion and gradient disappearance in RNN processing long-term dependencies, but also capture longer dependencies, which is more suitable for processing and predicting important events with long intervals in time series[12]. The cell state is the core component of LSTM, which can retain important historical information and be applied to the prediction task of the current moment, the input gate determines whether to write some or all of the current input information to the memory cell, the forgetting gate can selectively forget outdated or irrelevant historical information, ensuring that the state in the memory cell is always relevant to the current task, and the output gate is used to decide which information in the memory cell will be output to the subsequent network layer. This mechanism enables LSTMs to selectively expose information in memory cells according to task needs, thereby effectively regulating the transmission of information flow.

In this study, highly correlated feature data screened via Pearson correlation serve as input to the LSTM network, with wind power system fault results as the model output. Its structure is shown in Figure 1.

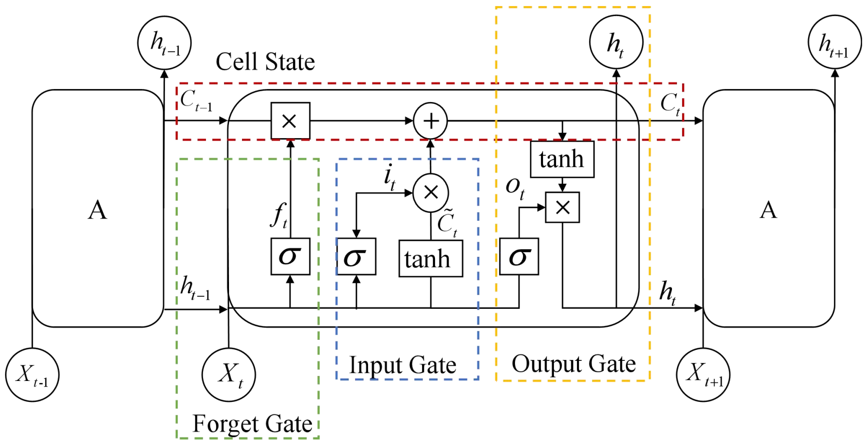

Figure 1 LSTM Network Structure Diagram

The forward propagation process of LSTM network is as follows:

Firstly, the forgetting gate screens σ the information to be retained or discarded in the C_{t-1} cell state through the gating mechanism, and outputs the gating vector that matches the cell state x_t using the current input h_{t-1} and the hidden state of the previous moment f_t. $\sigma(x)$ Used to press inputs into a $[0,1]$ range, defined as S type nonlinear functions, 0 represents complete discarding of the corresponding dimension information, and 1 represents complete retention. The formula is:

$$f_t = \sigma[W_f(h_{t-1}, x_t) + b_f] \tag{2}$$

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{3}$$

Secondly, the input gate generates σ a gating vector through x_t function processing h_{t-1} and sum i_t, identifies the dimension that needs to be updated in the cell state, and $\tanh$ performs a nonlinear transformation of the same input layer to generate a candidate cell state $\tilde{C}_t$. Represents potential updates, $\tanh(x)$ is a hyperbolic tangent function that compresses its input into a $[-1,1]$ range. The mathematical expression is:

$$i_t = \sigma[W_i(h_{t-1}, x_t) + b_i] \tag{4}$$

$$\tilde{C}_t = \tanh[W_c(h_{t-1}, x_t) + b_c] \tag{5}$$

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{6}$$

Then, the old cell state C_{t-1} is retained in proportion f_t by the forgetting gate, and then the part screened $\tilde{C}_t$ i_t by the input gate with the candidate state is updated through gated weighted fusion to obtain the new cell state at the current moment C_t. The mathematical expression is:

$$C_t = f_t\ C_{t-1} + i_t\ \tilde{C}_t \tag{7}$$

Finally, the output gate x_t generates a gating vector h_{t-1} based on and σ through the gating mechanism o_t, identifies C_t the dimension to be output in the updated cell state, and C_t multiplies $\tanh$ it by element after layer mapping o_t. Get the current moment hidden state h_t. The formula is as follows:

$$o_t = \sigma[W_o(h_{t-1}, x_t) + b_o] \tag{8}$$

$$h_t = o_t\ \tanh(C_t) \tag{9}$$

W_f, W_i, W_c, W_o is the corresponding parameter matrix, b_f, b_i, b_c, b_o is the bias vector.

Since multi-layer and hierarchical architectures are more efficient than shallow models in feature representation, and can extract abstract features of time series data by adding additional layers, the sequence output of a single-layer LSTM is widely used to deal with sequence prediction problems as the input of another layer of LSTM.

3.3 Model Evaluation

For the operation fault prediction model of wind power generation system based on LSTM, the performance evaluation needs to be evaluated to reveal the defects of the model after training. The fault prediction of wind power generation system is a binary classification problem, and there will be the following four typical judgment results in the prediction process: actual fault and predicted fault, true positive (TP), actual fault but predicted non-fault, is a false negative class (FN), which is actually a non-fault but is predicted to be a failure, which is called a false positive class (FP), which is actually non-fault and predicted to be non-fault, is a true negative class (TN), in order to quantify the prediction efficiency of the LSTM model on the fault state of the wind power generation system, the following parameters are used as the evaluation indexes.

(1) Accuracy means that the model predicts the correct ratio. The expression looks like this:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

(2) Precision is used to measure the proportion of actual failures predicted by the model as failure samples. The expression looks like this:

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

(3) Recall is used to describe the model's ability to capture fault samples, that is, the proportion of correctly predicted fault samples to actual fault samples. The expression looks like this:

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

(4) F1 score is an indicator that comprehensively considers accuracy and recall. The expression for the score is as follows:

$$F1\ score = \frac{2Precision \times Recall}{Precision + Recall} \tag{13}$$

In addition, the loss function is also a core component in machine learning and deep learning model training, which can be used to measure the difference between the predicted results and the actual results of the model, and the loss function of binary cross-entropy is suitable for the binary classification problem of predicting failure states. Therefore, in this study, the cross-entropy function is used as the loss function to adjust the parameter combination of the LSTM model to make the prediction results as close to the actual results as possible.

The mathematical expression of the cross-entropy function is:

$$Loss = -\frac{1}{N}\sum_{i=1}^{N}[y_i\ \ln\hat{y}_i + (1-y_i)\ \ln(1-\hat{y}_i)] \tag{14}$$

where represents N the number of samples, y_i represents the actual value, and $\hat{y}_i$ represents the predicted value.

Through the above indicators, the performance shortcomings of the LSTM model in the fault prediction of wind power generation system can be systematically evaluated, and a clear direction can be provided for subsequent model optimization.

4 EXPERIMENT ANALYSIS

4.1 Dataset Processing

The dataset used in this study is derived from the operation data of an onshore wind farm from February 2022 to January 2023, with a time interval of 1 minute, to test the effectiveness of the LSTM-based wind power generation system operation failure prediction method proposed in this paper.

Firstly, after data cleaning of the wind power system operation dataset, a total of 40,300 historical SCADA data were extracted, and the training set, test set and verification set were divided according to the ratio of 70%, 15% and 15%, of which the training set had a total of 28,208 data, the test set had a total of 6,046 data, and the verification set had a total of 6,046 data.

Then, 17 monitoring parameters of the SCADA system including fault state are extracted, and the SCADA data of the wind power generation system obtained by Pearson correlation coefficient method is used to screen the characteristics.

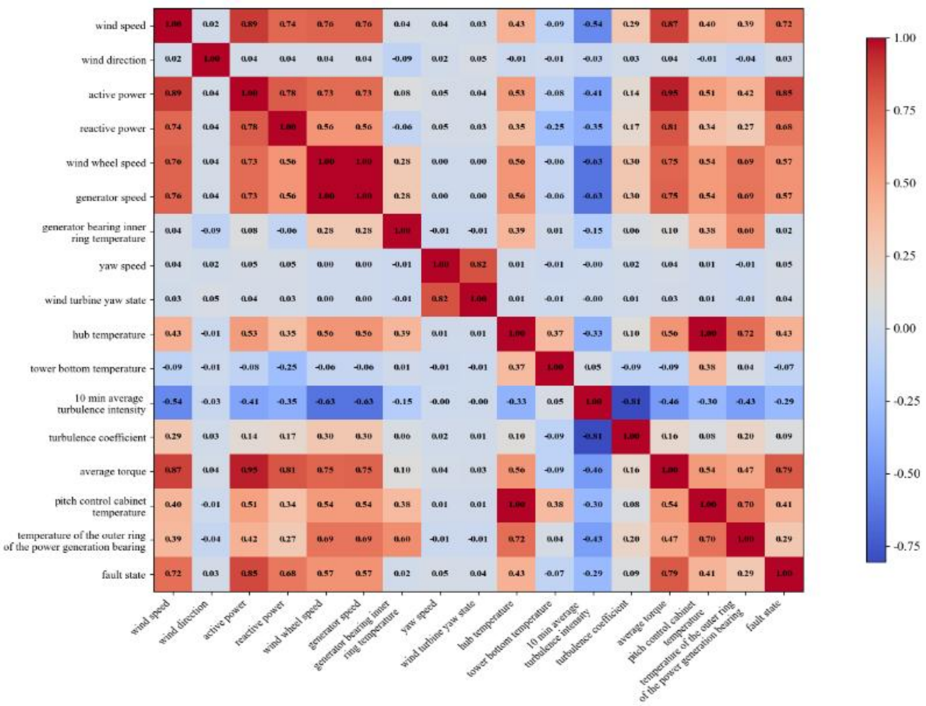

Figure 2 Pearson Correlation Coefficient Heat Map

The correlation coefficient heat map results are shown in Figure 2. It shows that there is a strong correlation between active power and system operation failure, with a correlation coefficient of 0.85, and the correlation between average torque, wind speed and reactive power and system operation failure is between 0.6 and 0.8, and there is a strong correlation.

Therefore, in this study, four characteristics of active power, average torque, wind speed and reactive power are selected as the inputs of the LSTM model for wind power system failure prediction, and the fault prediction is carried out by fusing multiple feature data to improve the accuracy of wind power system failure prediction.

4.2 Model Parameter Optimization and Performance Evaluation

In order to effectively prevent overfitting and improve the generalization ability of the model during model training, this study uses the method of cross-validation to determine the optimal hyperparameters, including learning rate, batch size and number of LSTM layers. The expression looks like this:

$$CV = \frac{1}{K}\sum_{k=1}^{K} L(\hat{y}^{(k)}, y^{(k)}) \tag{15}$$

wherein, $\hat{y}^{(k)}$ is the k prediction result of the folding verification set; $y^{(k)}$ is the actual value, and the final result is the K average loss value of the discount verification.

In the process of model training and optimization, the optimal combination of parameters is selected by cross-validation method, and the advantages and disadvantages of the parameters are evaluated by the loss function curve.

As shown in Figure 3, it is a schematic diagram of the loss function of the LSTM model and the optimization process of the model accuracy.

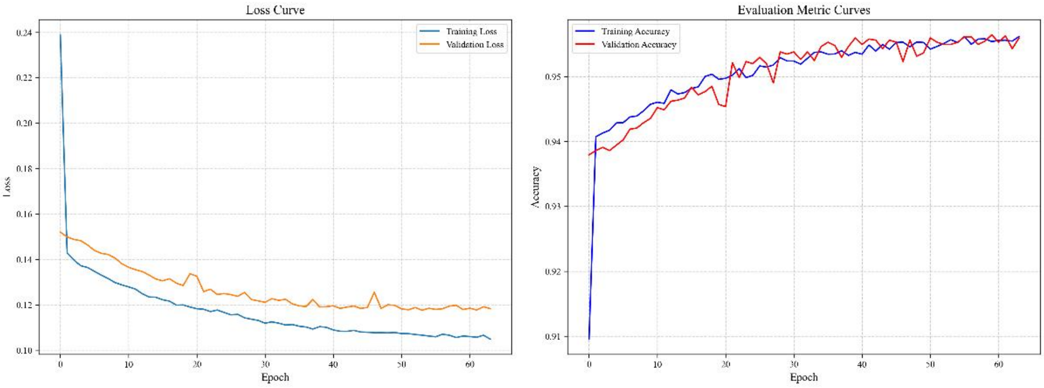

Figure 3 Schematic Diagram of LSTM Model Training Results

As shown in Figure 3, the training process of the LSTM model shows typical convergence characteristics, the training loss curve decreases rapidly in the first 10 rounds, and the verification loss curve decreases synchronously and stabilizes after about 50 rounds, and the two finally converge in a similar low value range, indicating that the model does not have obvious overfitting. In addition, the training accuracy and verification accuracy are simultaneously increased to more than 0.95, and the verification accuracy is always slightly higher than the training accuracy, indicating that the model has good generalization ability. The model training results show that the current set time step, number of LSTM units, dropout rate and learning rate constitute the optimal combination of hyperparameters, so that the model can maximize the failure prediction performance without overfitting.
As shown in Figure 4, the confusion matrix of the test set and the verification set obtained after using the LSTM model for wind power system operation fault prediction.

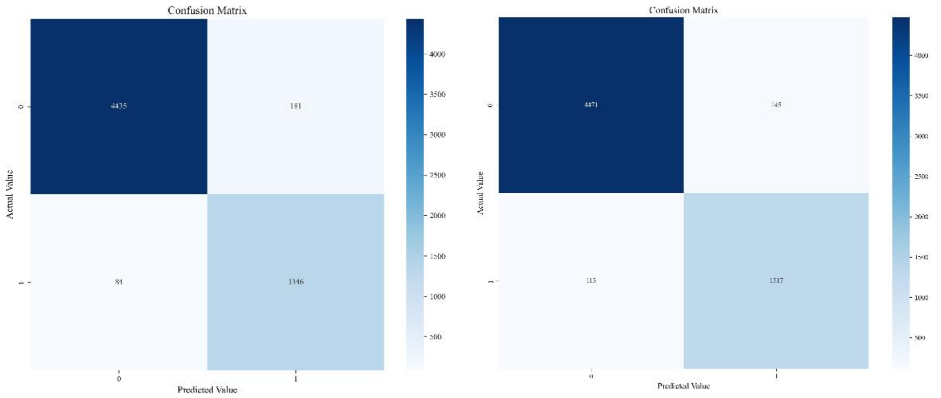

Figure 4 Confusion Matrix

In addition, the accuracy, recall, and F1 values can be further calculated by the confusion matrix, as shown in Table 3, which shows the prediction results of the LSTM model for wind power system operation failures.

Table 3 Failure Prediction Results of LSTM Model

Data subsets	Accuracy	Precision	Recall	F1 Score
Validation set	0.9567	0.8896	0.9308	0.9097
Test set	0.9583	0.9045	0.9210	0.9127

As shown in Table 3, the LSTM-based wind power fault prediction model performed well in both the validation set and the test set, with an overall accuracy of 95.67% and 95.83%, respectively, and in terms of the key indicators to measure the reliability of fault prediction, the accuracy of the model reached 90.45% in the test set, the recall rate reached 92.10%, both exceeded the 90% threshold, and the F1 score reached 91.27%, which was close to the optimal theoretical value. On the whole, the performance of the model in the verification set and the test set is consistent, indicating that the model has good generalization ability, not only has high accuracy, but also balances the risk of fault false alarm and false alarm, and has significant advantages in engineering practice.

4.3 Analysis and Discussion of Results

In order to explore the influence of multi-feature fusion on the fault prediction ability of LSTM model, this study experimentally compares the prediction performance of single operating parameters such as wind speed, active power, reactive power, and average torque with multi-feature fusion strategy, and the evaluation indicators cover accuracy, precision, recall rate and F1 score. The specific comparison results are shown in Table 4.

Table 4 Comparison of fault predictions for different feature inputs

Enter the characteristics		Accuracy	Precision	Recall	F1 Score
Single feature forecast	Wind Speed	0.9421	0.880	0.8741	0.8772
	Active Power	0.9535	0.903	0.9007	0.9016
	Reactive Power	0.8841	0.7334	0.8007	0.7656
	Average Torque	0.9386	0.875	0.8636	0.8694
Multi-feature fusion	This study	0.9562	0.882	0.9413	0.9104
Performance improvement ratio		2.86%	12.78%	11.22%	12.04%

As shown in Table 4, the active power is the best in a single feature with an accuracy rate of 95.35%, but its recall rate is only 90.07%, and the accuracy of wind speed and reactive power is close, but the accuracy and F1 score of reactive power are significantly lower, and the average torque index is lower than the active power. Compared with the prediction results of single feature failure, the accuracy of the multi-feature fusion strategy proposed in this study has been improved to 95.62%, the recall rate has jumped to 94.13%, the risk of missed judgment has been significantly reduced, and the F1 score has reached 91.04%, of which the accuracy and F1 score performance have been significantly improved, which are 12.78% and 12.04%, respectively.

In summary, the multi-feature fusion strategy can fully mine the coupling correlation of wind speed, power, torque and other parameters, and strengthen the recognition ability of LSTM for fault timing patterns through multi-dimensional information complementarity, avoiding the loss of information caused by the limitations of a single feature due to the limitations of physical mechanisms, thus effectively improving the accuracy of fault prediction. Therefore, it is difficult to fully characterize the complex evolution of wind power system faults due to the one-sided nature of physical correlation, while the multi-feature fusion strategy can significantly improve the comprehensive performance of the prediction model through cross-dimensional information collaboration, verify the key role of feature engineering in wind power fault prediction, and provide a direction for subsequent model optimization.

5 CONCLUSION

In this study, based on the operation data of the SCADA system, the highly correlated features were screened by the Pearson correlation coefficient method to construct an LSTM model with multi-feature input. In the future, the dynamic update mechanism of the model can be further constructed to optimize the LSTM architecture, enhance the long-time series dependency characterization ability and computing efficiency, and help the intelligent operation and maintenance system of wind power generation system to be accurately upgraded.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Kozitsin V, Katser I, Lakontsev D. Online forecasting and anomaly detection based on the ARIMA model. Applied Sciences-BASEL, 2021, 11(7): 3194.

[2] Yang Bofan, Zhang Lin, Zhang Bo, et al. Online prediction method for dynamic multi-model exponential smoothing fusion. Systems Engineering and Electronic Technology, 2020, 42(09): 2013-2021.

[3] Zhang Z, Dong S, Li D, et al. Prediction and Diagnosis of Electric Vehicle Battery Fault Based on Abnormal Voltage: Using Decision Tree Algorithm Theories and Isolated Forest. Processes, 2024, 12(1): 136.

[4] Arumbu P V, Karthikeyan D. Reliability Assessment and Fault Prediction in a 13-Level Multilevel Inverter Through Machine Learning with SVM. Journal of Electrical Engineering & Technology, 2024, 20(1): 1-16.

[5] Sachin Kumar R, Sakthiya Ram S, Arun Jayakar S, et al. Failure prediction of turbines using machine learning algorithms. Materials Today: Proceedings, 2022, 66(3): 1175-1182.

[6] Gawali B M, Gawali S S, Patil M. Fault prediction model in wind turbines using deep learning structure with enhanced optimization algorithm. Journal of Control and Decision, 2025, 12(3): 471-488.

[7] Bharatheedasan K, Maity T, Kumaraswamidhas L, et al. Enhanced fault diagnosis and remaining useful life prediction of rolling bearings using a hybrid multilayer perceptron and LSTM network model. Alexandria Engineering Journal, 2025, 115, 355-369.

[8] Jin X, Chunyan L, Bo L, et al. Prediction of wind turbine blade icing fault based on selective deep ensemble model. Knowledge-Based Systems, 2022, 242.

[9] Daohua Z, Xinxin J, Piao S. Research on power system fault prediction based on GA-CNN-BiGRU. Frontiers in Energy Research, 2023, 11: 1245495. DOI: 10.3389/fenrg.2023.1245495.

[10] Zare A, Babalou M, Torkaman H. Open-circuit fault diagnosis of modular DC-DC converter based on multi-layer perception//2025 16th Power Electronics, Drive Systems, and Technologies Conference, PEDSTC. Tabriz, Iran, Islamic Republic of. IEEE, 2025, 1-5. DOI: 10.1109/PEDSTC65486.2025.10912043.

[11] Zhang W G, Li H R, Tang L B, et al. Displacement prediction of Jiuxianping landslide using gated recurrent unit

(GRU) networks. Acta Geotechnical, 2022, 17(4): 1367-1382.
[12] Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Computation, 1997, 9(8): 1735-1780.