Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.36474

# *Research on Machine Learning Regression Prediction of Insulator Contamination Degree for the Digital Transformation of Transmission Line Operation and Maintenance Services*

**Feng Wang*, Gangqiang Cao, Dingguo Liu**

*State Grid Hunan Electric Power Co., Ltd., Changsha, China*
**Corresponding Author. Email: wangf22@hn.sgcc.com.cn*

***Abstract.*** The insulators of transmission lines are prone to accumulate dirt when exposed to the outdoor environment for a long time, which can lead to pollution flashover. Traditional manual detection is lagging and costly. Therefore, there is an urgent need for high-precision online prediction technology. Although machine learning can fit the pollution pattern under the coupling of multiple factors, a single model has deficiencies such as insufficient extraction of temporal features and poor stability. Based on the on-site monitoring platform, this paper obtains 1638 sets of environmental and equipment operation data, and conducts correlation analysis after preprocessing to clarify the feature associations. Then, the BiLSTM-Bayes-SVM regression prediction algorithm is proposed. This algorithm uses the bidirectional long short-term memory network to mine the global temporal correlation features, and uses the Bayesian algorithm to allocate probability weights to eliminate redundant interference. Finally, it achieves high-precision fitting through support vector machine regression. Experimental results show that compared with traditional models, the root mean square error of this fusion algorithm is as low as 0.020, the average absolute percentage error is 10.571, and the coefficient of determination is 0.944. All error indicators are comprehensively superior to algorithms such as XGBoost, effectively improving the accuracy and generalization performance of pollution degree prediction, and meeting the requirements of safe operation and maintenance of smart grids.

***Keywords:*** Transmission line insulators, pollution flashover, BiLSTM, SVM

## 1. Introduction

Transmission lines are the core carriers for the transmission of electrical energy in the power system, and their safe and stable operation directly affects the overall power supply reliability of the power grid [1]. Insulators, as the key insulating components of transmission lines, are exposed to the complex natural environment outdoors for a long time, and are prone to adsorbing dust, salt, and industrial pollutants in the air, forming a surface contamination layer. As the contamination accumulates, under the influence of humid weather conditions, the insulation performance of

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.36474

insulators will significantly decline, easily triggering power faults such as pollution flashover tripping, and in severe cases, causing large-scale power outages, bringing significant losses to the operation of the power system and social production and life [2]. Currently, the detection of insulator contamination degree in the power industry mainly relies on manual regular inspections and offline sampling detection. These traditional detection methods have many drawbacks, such as strong lag, high labor costs, and limited detection accuracy, and cannot achieve real-time dynamic perception of the contamination state of insulators, making it difficult to meet the development needs of regular safety operation of the intelligent power grid. Therefore, conducting research on high-precision online prediction of insulator contamination degree has important engineering practical value [3].

With the rapid development of intelligent sensing technology and artificial intelligence technology, machine learning algorithms have been widely applied in the field of power equipment condition prediction and fault diagnosis, providing a new technical idea for online prediction of insulator contamination degree [4]. Traditional mathematical modeling methods highly rely on equipment physical parameters and manual experience, and are difficult to accurately fit the changing patterns of contamination deposition under the coupling of multiple factors such as meteorological environment and operating conditions. The generalization ability of prediction is poor [5]. However, machine learning algorithms have powerful nonlinear data fitting and feature mining capabilities, and can autonomously learn the complex correlations hidden in massive monitoring data, effectively avoiding the limitations of traditional modeling methods. By deeply mining multi-dimensional environmental data such as temperature, humidity, wind speed, rainfall, and equipment operation data, the machine learning model can accurately capture the dynamic changes of contamination accumulation and achieve continuous prediction of contamination degree, significantly improving the intelligence and accuracy of power equipment condition assessment.

Single machine learning models and traditional time series prediction algorithms often have problems such as insufficient extraction of time series features, poor fitting effect on small sample data, and insufficient stability of prediction results when dealing with insulator contamination monitoring data. They are difficult to meet the engineering requirements for high-precision prediction. To further improve the prediction accuracy and robustness of insulator contamination degree for transmission line insulators, this paper combines the advantages of bidirectional long short-term memory network, Bayesian optimization algorithm, and support vector machine regression to propose a new BiLSTM-Bayes-SVM regression prediction algorithm. This algorithm relies on the bidirectional long short-term memory network to fully mine the sequential correlation features of monitoring data, uses the Bayesian optimization algorithm to intelligently optimize model parameters, solves the problem of low efficiency in manual parameter debugging of traditional models, and completes high-precision fitting prediction through support vector machine regression, effectively compensating for the prediction defects of a single model and being better adapted to the dynamic prediction scenarios of insulator contamination degree in complex environments.

## 2. Data sources

The insulator pollution prediction data set adopted in this research relies on the on-site online monitoring platform of transmission lines to complete the data collection work, and collects the measured data related to various environmental meteorological parameters and equipment operation status in the power operation and maintenance site for a long time. At the same time, the corresponding measured values of pollution are obtained by combining on-site physical and

DOI: 10.54254/2755-2721/2026.36474

chemical tests. In the process of research, the original collected data is systematically preprocessed, and the basic operations such as eliminating abnormal data and filling up missing values and screening redundant information are completed in turn, and the data statistical standards of various monitoring indicators are further unified to ensure that the overall data sample quality meets the basic requirements of model training. There are 1638 groups of valid sample data in the whole data set, which are mainly divided into three categories: related characteristics of environmental monitoring, operating status and parameters of parameters and equipment, and measured target parameters of pollution degree. The correlation analysis was performed on the individual variables and the correlation heat map was plotted, as shown in Figure 1.

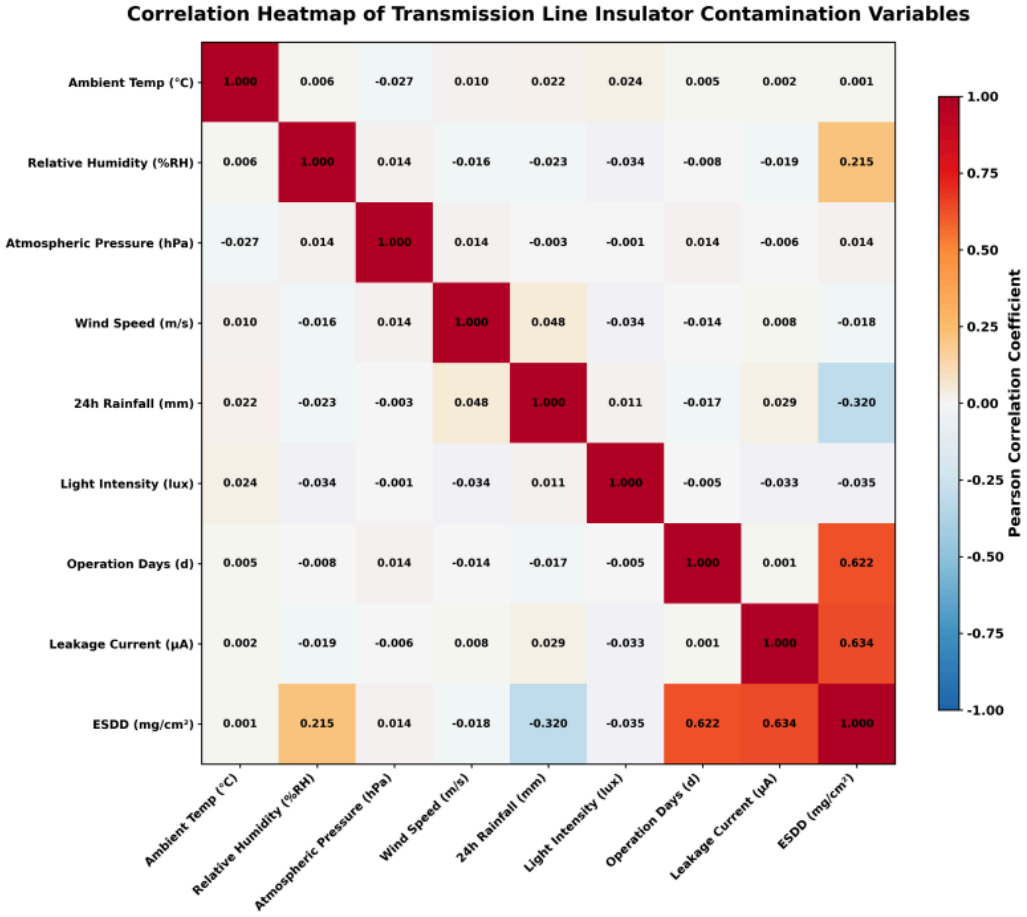


Figure 1. The correlation heat map

## 3. Method

### 3.1. BiLSTM

BiLSTM is a two-way long short-term memory neural network, which adds a reverse information propagation channel to the traditional one-way LSTM. The algorithm relies on a special gating structure to regulate the information flow state, which can effectively capture the long-term internal dependencies of time series data. The forward network reads the sequence data in chronological order to extract forward features, and the reverse network traverses forward from the end of the sequence to mine post-hidden features [6]. After the feature information output by the two types of

DOI: 10.54254/2755-2721/2026.36474

networks is fused and spliced, the global semantics and changing laws before and after time series data can be completely covered. The network structure of BiLSTM is shown in Figure 2.

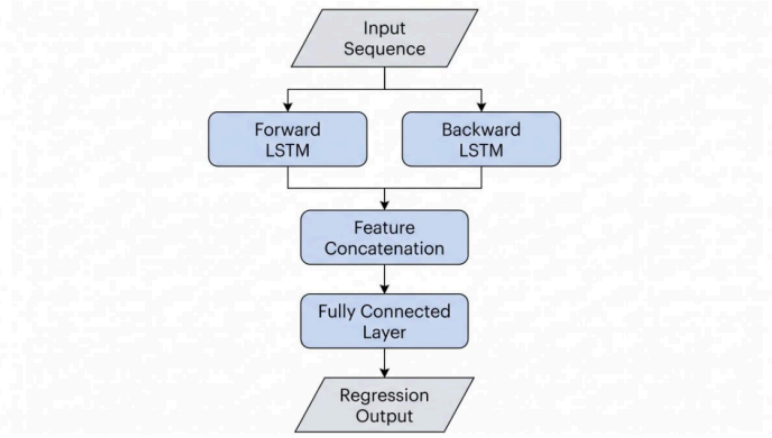


Figure 2. The network structure of BiLSTM

## 3.2. Bayes

Bayes algorithm relies on the basis of probability and statistics to carry out data modeling operation, and the core relies on prior probability combined with sample characteristics to calculate posterior probability to complete data deduction. The algorithm follows the basic probability operation logic, relies on the known data distribution law to determine the initial probability value, and continuously modifies the probability judgment result based on the newly added sample attribute characteristics [7]. In the operation process, various characteristic information are defaulted to be independent of each other, which simplifies the correlation operation process between complex data and greatly reduces the overall operation complexity. The network structure of Bayes is shown in Figure 3.

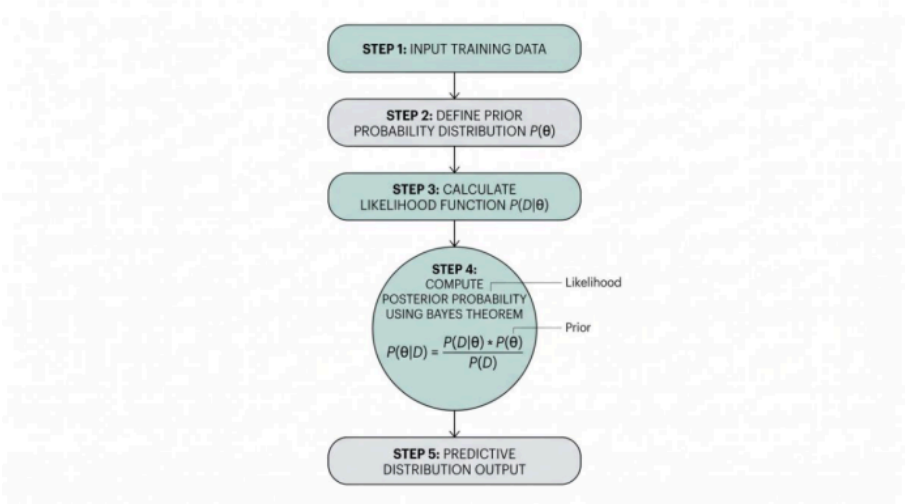


Figure 3. The network structure of Bayes

## 3.3. SVM

SVM algorithm takes the optimal classification hyperplane as the core operation basis, and relies on the interval maximization criterion to complete data space division and feature fitting. The algorithm

DOI: 10.54254/2755-2721/2026.36474

maps the sample data that is difficult to distinguish in low-dimensional space to high-dimensional feature space through kernel function, and finds the optimal plane structure that can accurately separate different data samples in the new space [8]. In the process of calculation, the distance between the two types of samples and the separation plane is controlled, and the plane position is continuously optimized to maximize the global interval. The network structure of SVM is shown in Figure 4.

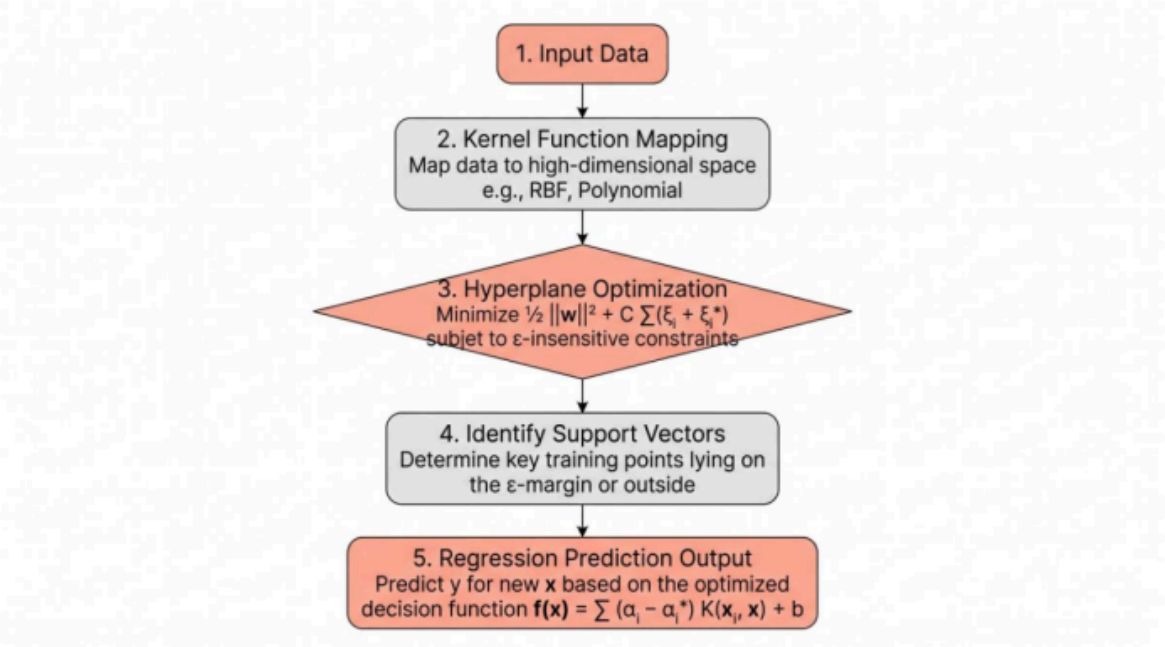


Figure 4. The network structure of SVM

### 3.4. BiLSTM-Bayes-SVM

The fusion prediction algorithm completes the three-layer operation process of feature extraction probability correction and accurate fitting in turn. Firstly, BiLSTM is used to comprehensively mine the global before and after correlation features of time series prediction data, and effective feature information with complete time series law is extracted. Then Bayes algorithm is introduced to allocate probability weights to the extracted feature data, and the calculation proportion of various features is adjusted according to the data distribution law, so as to eliminate the calculation interference caused by invalid redundant features [9]. Finally, the feature data of weight optimization is input into SVM model, and the nonlinear regression operation model is constructed with the help of SVM high-dimensional space fitting ability, and the numerical trend fitting is completed by the optimal interval criterion. The three types of algorithms give full play to their own computational advantages layer by layer, realize the mutual cooperation between the probability optimization and adjustment of deep mining of time series features and high-precision numerical fitting, rely on the multi-layer computational process to refine the inherent change law of data, and gradually improve the accuracy of the overall regression prediction results [10]. The network structure of BiLSTM-Bayes-SVM is shown in Figure 5.

DOI: 10.54254/2755-2721/2026.36474

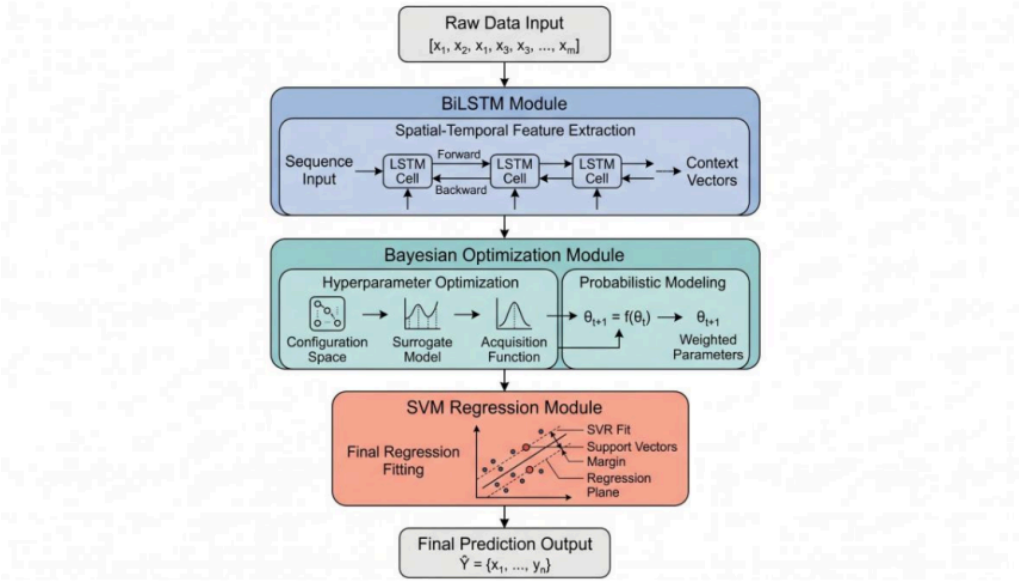


Figure 5. The network structure of BiLSTM-Bayes-SVM

## 4. Result

In this project, the regression prediction algorithm based on BiLSTM-Bayes-SVM uses Bayesian optimization to automatically tune the core hyperparameters of the model. The number of hidden layer neurons of BiLSTM is set to 64 to fully extract time series features, the penalty coefficient of SVM is optimized to 1.5 to balance the model complexity and fitting error, and the learning rate is set to 0.01 to ensure the stable convergence of the training process. All sample data are divided into training set and test set according to the ratio of 7:3.

Comparative experiments were performed using a variety of machine learning algorithms, and the results are shown in Table 1.

Table 1. The results of the comparative experiment

| Model | MSE | RMSE | MAE | MAPE | $R^2$ |
|---|---|---|---|---|---|
| ExtraTrees | 0.001 | 0.027 | 0.022 | 13.254 | 0.898 |
| Decision Tree | 0.001 | 0.036 | 0.028 | 17.407 | 0.83 |
| Random Forest | 0.001 | 0.025 | 0.02 | 11.947 | 0.914 |
| GBDT | 0.001 | 0.023 | 0.018 | 11.039 | 0.927 |
| AdaBoost | 0.001 | 0.024 | 0.019 | 11.372 | 0.918 |
| KNN | 0.006 | 0.08 | 0.064 | 33.133 | 0.128 |
| CatBoost | 0.001 | 0.023 | 0.018 | 10.831 | 0.928 |
| XGBoost | 0 | 0.022 | 0.018 | 10.702 | 0.934 |
| BiLSTM-Bayes-SVM | 0 | 0.02 | 0.015 | 10.571 | 0.944 |

Output the index comparison histogram of each algorithm, as shown in Figure 6.

DOI: 10.54254/2755-2721/2026.36474

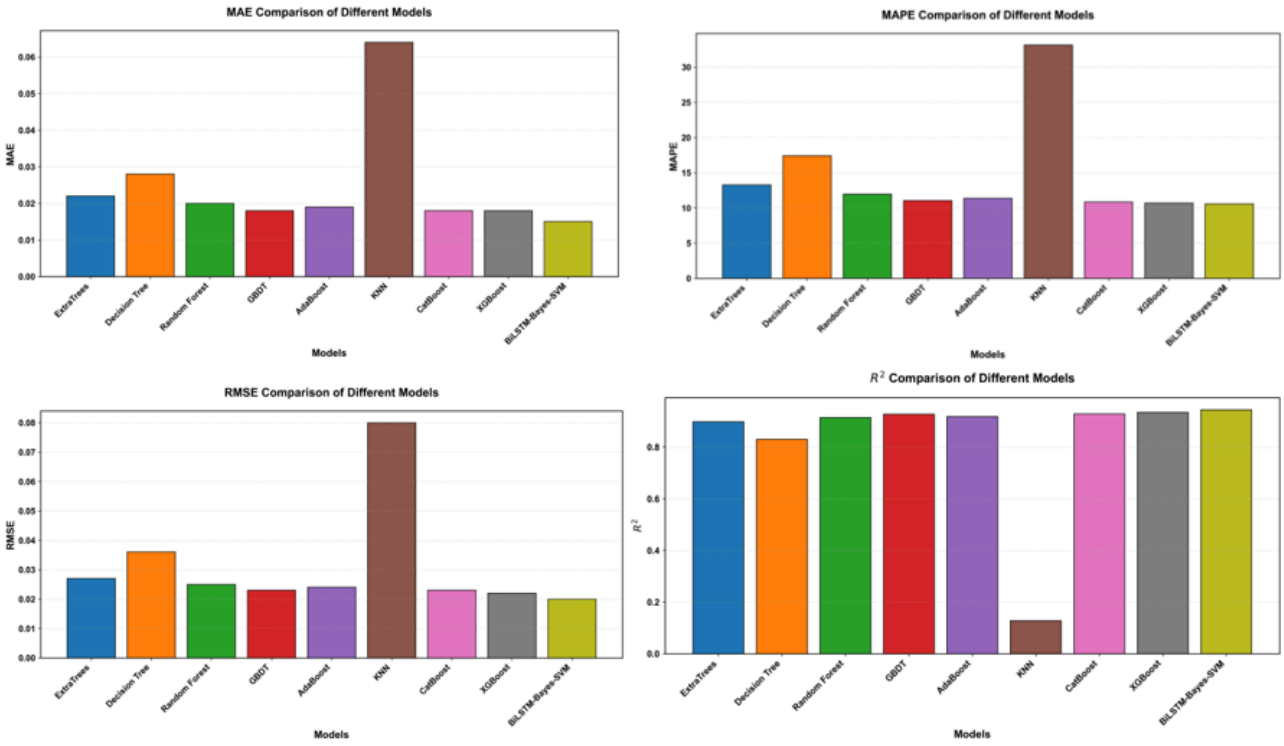


Figure 6. The index comparison histogram of each algorithm

Comparing the prediction performance of various machine learning models, it can be seen that the overall performance of the traditional ensemble learning model is obviously better than that of the basic decision tree and KNN algorithm. The error value of KNN model is high and the fitting degree is poor, and only the determination coefficient of 0.128 is obtained, so the prediction effect is difficult to meet the actual demand. In the conventional tree-class ensemble model, XGBoost has excellent comprehensive performance, with an average absolute percentage error of 10.702 and a coefficient of determination of 0.934. CatBoost and GBDT are closely followed, and all indexes of the basic decision tree are in the middle and lower levels. The evaluation indexes of the BiLSTM-Bayes-SVM algorithm proposed in this paper are better than all comparative models. The root mean square error of this algorithm is as low as 0.020, the average absolute percentage error is only 10.571, and the determination coefficient is improved to 0.944. All error indexes are lower than those of mainstream algorithms such as XGBoost random forest, which can deeply mine the deep feature information of data, effectively improve the prediction accuracy and model fitting ability, and have better prediction ability and generalization performance than traditional machine learning algorithms.

## 5. Conclusion

Ensuring the insulation reliability of transmission lines is very important to the stability of power grid. The risk of power outage caused by pollution accumulation cannot be ignored. Manual inspection is difficult to cope with the dynamic perception demand in complex environment. Through the correlation analysis of multi-dimensional meteorological and operational parameters, the potential coupling correlation between multi-variables is revealed, which lays the foundation for feature optimization. The constructed fusion algorithm gives full play to the advantages of each model layer by layer. The bidirectional network deeply captures the forward and backward semantic

DOI: 10.54254/2755-2721/2026.36474

dependence of the sequence, the probability and statistical correction effectively suppresses the invalid feature interference, and the high-dimensional mapping fitting ensures the accuracy of numerical deduction. This multi-level collaborative computing mechanism breaks through the bottleneck of insufficient generalization of conventional machine learning, and the comparative experiment confirms its excellent performance. Not only is the error index fully refreshed, but the determination coefficient also reaches the highest level. This research realizes the high-reliability dynamic evaluation of insulator pollution situation, provides an innovative paradigm for intelligent operation and maintenance of power equipment, and has great engineering practical significance for preventing pollution flashover faults and improving power supply safety.

## Acknowledgment

State Grid Hunan Electric Power Co., Ltd. Professional Consulting Project: Evaluation of the Digital Transformation Effects of Transmission Operation Inspection Model and Optimization of Implementation Strategies (B616AJ25Z000).

## References

[1] Khan, U. A., et al. (2025). Experimental validation of machine learning for contamination classification of polluted high voltage insulators using leakage current. *Scientific Reports, 15*(1), 13246.
[2] de Santos, H., & Sanz-Bobi, M. Á. (2023). A machine learning approach for condition monitoring of high voltage insulators in polluted environments. *Electric Power Systems Research, 220*, 109340.
[3] Ahmed, M. F., Mohanta, J. C., & Sanyal, A. (2022). Inspection and identification of transmission line insulator breakdown based on deep learning using aerial images. *Electric Power Systems Research, 211*, 108199.
[4] GöRGöZ, İ., & Cebeci, M. (2025). Analysis of electric field and temperature distributions of non-uniformly contaminated silicone composite insulators using deep learning. *IEEE Access*.
[5] Wang, H., et al. (2024). Deep learning based insulator fault detection algorithm for power transmission lines. *Journal of Real-Time Image Processing, 21*(4), 115.
[6] da Silva, M. G. L., et al. (2026). Monitoring leakage current in insulator strings for flashover risk prediction: A systematic literature review. *IEEE Transactions on Power Delivery*.
[7] Zhang, Z., et al. (2026). Non-contact detection technology of operation status for transmission line insulators: Characteristics, perspectives, and challenges. *Energies, 19*(3), 636.
[8] Liu, Y., et al. (2024). Optical imaging technology application in transmission line insulator monitoring: A review. *IEEE Transactions on Dielectrics and Electrical Insulation, 31*(6), 3120-3132.
[9] Li, X., et al. (2026). Insulator defect detection in transmission line in sandy and dusty scenarios based on deep learning. *Electric Power Systems Research, 256*, 112852.
[10] Sun, J., et al. (2021). Contamination degree prediction of insulator surface based on exploratory factor analysis-least square support vector machine combined model. *High Voltage, 6*(2), 264-277.