EAI Endorsed Transactions
on Energy Web

Research Article EAI.EU

Fault Diagnosis Algorithm Based on Power Outage Data in Power Grid

Haiyan Wang[1,*], Xinping Yuan[1], Shanfei Gao[1], Shoushan Gao[1]

1.Yunnan Power Grid Co., Ltd. Kunming Enersun Technology Co., Ltd. Kunming 65000, Yunnan, China

Abstract

INTRODUCTION: With the rapid development of the power industry, the power system has become more and more complex and prone to failures, which seriously impacts power supply and safety.
OBJECTIVES: Development of efficient and accurate fault diagnosis algorithms for power systems.
METHODS:Proposes a fault diagnosis algorithm based on outage data to construct an outage fault prediction model using accurate data. First, the outage data are collected, pre-processed, feature extracted and reduced to obtain a more efficient data set. Then, an optimized fault diagnosis algorithm is designed based on logit, support vector machine (SVM) and decision tree (DT) to improve the accuracy and efficiency of fault diagnosis.
RESULTS: The method is applied to the natural power system, and the results show that the optimization algorithm outperforms the traditional methods. Specifically, the accuracy of the optimization algorithm can reach 100%, while the accuracy of the traditional logit algorithm and SVM algorithm is only 84% and 93%, which is a significant improvement in the model prediction performance.
CONCLUSION: The author can significantly optimize the performance of its model and construct an outage data mining algorithm with a good predictive ability to achieve grid fault research and judgment, which has a specific application value in the practical field.

Keywords: Power grid, Fault diagnosis, Support vector machines, Decision tree

Received on 27 April 2023, accepted on 15 December 2023, published on 20 December 2023

doi: 10.4108/ew.4657

1. Introduction

The power grid is the infrastructure of modern society and plays an irreplaceable role in all aspects of People's Daily lives. With the increase in electricity demand and the grid system's expansion, the power system's complexity and scale have also increased dramatically. Therefore, the power system is increasingly prone to failure, resulting in power outages, equipment damage, economic losses, security risks, and other serious consequences. Therefore, developing an efficient and accurate fault diagnosis algorithm for a power system is very important.

Fault diagnosis identifies the cause and location of faults in the power grid system. The traditional fault diagnosis methods mainly rely on manual inspection and expert experience, which are time-consuming, expensive, and may have errors. With the development of information technology, data-driven fault diagnosis methods, which include artificial intelligence techniques such as machine learning, deep learning, and data mining, are becoming more popular. Among these techniques, support vector machines (SVM) and decision trees (DT) are widely used in fault diagnosis due to their excellent performance and easy implementation.

Based on this idea, this paper proposes a prediction model using SVM and DT algorithms. Accurate power outage data is used to construct the prediction model, which is

[*] Corresponding Author. Email: 18208891009@163.com

trained to identify faults effectively. Chapter 2 of this paper outlines the main characteristics of past research, Chapter 3 explains the data processing process, Chapter 4 details the model prediction, and Chapter 5 provides a summary.

2. Related Work

Many researchers have proposed various power grid fault diagnosis algorithms with rich data sources and various accumulated research types. Relevant literature can be classified into two categories for power outage fault detection: data-driven methods and model-based methods.

2.1 Data-Driven Methods

Data-driven methods use data collected from the power grid system to construct fault diagnosis models, which can be artificial neural networks (ANNs), decision trees (DTs), support vector machines (SVMs), or deep learning networks (DLNs). The advantage of data-driven methods is that they do not require precise models and parameters and can learn from massive data. Support vector machines and DT techniques have been widely used in fault diagnosis due to their excellent performance and interpretability.

Liu et al. propose an extensive data-driven data collection method to build a scientific customer outage time key indicator around the power grid comprehensive customer service system, establish a customer outage time indicator responsibility system, and establish a customer outage process monitoring platform[1]. Using big data analytics, the authors created a data-driven customer power outage analysis system to handle customer outages effectively. Chinthavali proposed a standardized format for outage data in Seattle City Light. The standardized format includes specific information such as outage start time, end time, number of affected customers, circuit ID, and cause of the outage[2]. This format is expected to improve outage data collection and analysis accuracy and consistency, eventually leading to better outage management and customer service. Chunyan et al. presented a method to identify power blackout-sensitive users in the energy system using big data analytics. The authors proposed an approach to analyze social media data to identify customers susceptible to power disruptions[3]. They used the results to develop a customer prioritization scheme for power restoration work. Yue et al. proposed a Bayesian approach-based outage prediction in electric utility systems using radar measurement data[4]. Based on radar data, the authors used Bayesian inference to estimate the probability of an upcoming outage. The approach improved the prediction accuracy and the data processing speed, thus providing a powerful tool for grid operators.

2.2 Model-Based Methods

Model-based methods rely on mathematical models to analyze the physical and electrical characteristics of the power grid system and then conduct fault diagnosis based on the obtained models. The diverse analysis methods provide model references for the discussions in this paper. However, these methods require precise models and parameters, often challenging to obtain in practice, affecting such models' application value.

The paper by Liu et al. proposed a data inference-based maintenance method to mitigate the risk of cascading blackouts[5]. The authors developed a mathematical model to analyze the correlation between various maintenance indicators and the risk of blackouts. The approach resulted in a better understanding of the risk of blackouts and reduced the likelihood of cascading blackouts. Jun et al. address the issue of frequent outage complaints based on data mining[6]. Using data mining techniques, the authors proposed a model to construct early warning for frequent outage complaints. The approach involved identifying the critical predictors of customer complaints and leveraging them to develop a predictive model. Gurara and Tessema analyzed the impact of blackouts on firms using firm-level data[7]. The authors developed a model to estimate the impact of blackouts on firm-level productivity and profitability in different sectors. The results showed that blackouts have a significant negative impact on firm performance.

2.3 Literature Review

In summary, the research mentioned above analyzes power outage fault identification problems from different perspectives, and their research methods and results provide references and inspiration for the discussions in this paper. Both data-driven and model-driven methods are valuable for collecting and analyzing power outage data, but their application scope differs. On the one hand, data-driven methods use extensive data analysis to provide insights into customer behavior and power grid operation, thus requiring large-scale data to support the validity of their analysis results, mainly relying on objective data rather than simulated models[8-10]. On the other hand, model-driven methods use mathematical models to analyze the correlation between various factors and power outage risks, significantly increasing the models' universality. However, their fit to specific sample data is often weaker than the training results of large-scale data samples[11-12]. Both methods can improve power grid management and maintenance, thereby minimizing power outages and enhancing customer satisfaction[13-14]. Montanari and Dimitriou's paper discusses developing and utilizing the IAEA stopping power database, which is crucial in understanding how ions interact with matter[15]. In Michael M. Li and Brijesh Verma's paper, they propose using Radial Basis Function (RBF) neural networks for nonlinear curve fitting[16]. Stanko Novakovic et al.'s paper investigates the challenges and implications of load imbalance in distributed systems, a critical issue in ensuring efficient and reliable data serving[17]. Bahiru Egziabiher, Scott Thomsen, and John Simmins' paper discuss the importance of collaboration, data standards, and APIs in the utility industry, focusing on Seattle City Light's outage data initiative[18]. This paper refers to and draws inspiration from model-driven analysis methods based on existing research, processes a wide range of electrical parameter values, and uses a large amount of data-driven models to improve their fitting effectiveness.

3. Proposed Method

In this chapter, the author proposes a fault diagnosis algorithm based on power outage data in a power grid system. The proposed algorithm aims to accurately diagnose faults in power grid systems and improve the overall reliability and efficiency of power grid operations.

3.1 Step 1: Data Collection and Pre-processing

The first step in the proposed algorithm is to collect and pre-process power outage data from sensors in the power grid system.

The indicator system constructed in this article contains 16 leading indicators. These 16 indicators can be divided into several aspects.

Voltage-related indicators: maximum voltage, minimum voltage, voltage mean, voltage standard deviation, etc., reflecting the magnitude and stability of voltage fluctuations in the power grid. If the maximum voltage exceeds the rated value, electrical equipment faults or other issues may require maintenance or replacement. If the minimum voltage exceeds the rated value, electrical equipment faults or other issues may require maintenance or replacement. The voltage mean is very useful in checking the overall operating status of the electrical equipment in the system. For example, if the voltage is low, electrical equipment may have faults or inadequate power supply. Voltage standard deviation: The standard deviation quantifies the degree of voltage fluctuation and is an essential indicator for checking the stability of the electrical equipment in the system. If the voltage standard deviation is large, it indicates significant fluctuations in the power network, which may require adjustment or upgrading.

Current-related indicators: maximum current, minimum current, current mean, current standard deviation, etc., reflecting the magnitude and stability of current in the power grid. If the maximum current exceeds the rated value of the equipment, there may be electrical equipment faults or power supply system instability. If the minimum current is lower than the rated value, electrical equipment faults or power supply system instability may occur. The current mean can determine whether the electrical equipment is functioning normally. If the current is low, there may be electrical equipment faults or inadequate power supply. The standard deviation quantifies the degree of current fluctuation and is an essential indicator for checking the system's electrical equipment's stability.

Power-related indicators: active power, reactive power, apparent power, power factor, etc..

If the active power is low, electrical equipment faults or inadequate power supply may occur. If the reactive power is low, adjusting or upgrading the electrical equipment may be necessary to improve the operating efficiency. If the actual power is low, there may be electrical equipment faults or other issues. If the power factor is low, upgrading or replacing older electrical equipment may be necessary.

Frequency-related indicators: power grid frequency, frequency deviation, etc., reflecting the balance of power supply and use in the power grid. If the frequency is low, there may be electrical equipment faults or power supply issues. Frequency deviation affects the stability and regular operation of electrical equipment. Therefore, it is a crucial monitoring indicator.

Transformer-related indicators: transformer temperature, transformer load rate, etc., reflecting the operating status and health of transformers in the power grid. Electrical equipment faults may require maintenance or replacement if the transformer temperature exceeds the rated value. If the transformer load rate is high, upgrading or replacing older electrical equipment may be necessary.

Load-related indicators: load size, load rate, etc., reflecting the size and stability of the load in the power grid.

These indicators can comprehensively reflect the operating status and health of the power grid, which helps diagnose and predict power grid faults.

Then, the data is pre-processed to remove noise, outliers, and missing values.

3.2 Step 2: Feature Extraction and Reduction

To eliminate the dimensional differences between feature attributes, the monitoring data of all state parameters are normalized respectively. The normalization method is shown in Equation (1) :

$$x^* = \frac{x_i}{\sum_{i=1}^{N} x_i} \quad i = 1,2,\cdots,N \tag{1}$$

The equation includes N as the monitoring period and x_i any state parameter's monitoring data on the i-th day.

Using normalized transformer state parameter data, a high-dimensional spatiotemporal state monitoring matrix is constructed in both time and spatial sequence, as shown in matrix W.

$$W = \left\{ \begin{matrix} w_{11} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{m1} & \cdots & w_{mn} \end{matrix} \right. \quad (2)$$

In the equation, m = 1, 2, M, where M is the monitoring period; n = 1, 2, N, where N is the number of state parameters; the matrix element i j w is the feature value of the jth state parameter on the ith day.

The above pre-processing and analysis methods laid a necessary foundation for the subsequent analysis of this paper. They avoided the impact of dimensional differences on the fluctuation analysis of sample data in the case of the overall limited regression of sample data.

After data pre-processing is completed, the next step is to extract the relevant features from the signal and reduce the dimension of the feature space. The raw data has high dimensional characteristics. On this basis, the dissimilarity matrix D between any two nodes i and j in the high-dimensional space-time state monitoring matrix W is calculated by using Euclidean distance:

$$d_{ij} = \sqrt{\sum_{k=1}^{n} \left(x_{ik} - x_{jk} \right)^2} \quad k = 1,2,\cdots,n \quad (3)$$

The coordinates of node i are:

$$X_i = \left[x_{i1}, x_{i2}, \cdots, x_{in} \right] i = 1,2,\cdots,m \quad (4)$$

The coordinates of node j are:

$$X_j = \left[x_{j1}, x_{j2}, \cdots, x_{jn} \right] \ j = 1,2,\cdots,m \quad (5)$$

Thus, the doubly centered matrix B can be calculated based on the dissimilarity matrix B:

$$B_{ij} = \frac{1}{2}\left(d_{ij}^2 - \frac{1}{l}\sum_{i=1}^{l} d_{ij}^2 - \frac{1}{l}\sum_{j=1}^{l} d_{ij}^2 + \frac{1}{l^2}\sum_{i=1}^{l}\sum_{j=1}^{l} d_{ij}^2 \right)$$

$$i = 1,2,\cdots,m \,,\ j = 1,2,\cdots,m \quad (6)$$

By calculating the doubly centered matrix B, the high-dimensional spatiotemporal monitoring matrix W can be represented in two-dimensional space, thereby reducing data dimensionality.

3.3Step 3: Fault Diagnosis Algorithm Optimization

Subsequently, further analysis is conducted on the raw data. The various terminal nodes and their inter-regions in the distribution network are numbered for each sample, and the network correlation matrix is presented in Table 1.

Table 1 The Correlation Relationships between Nodes and Regions and Their Corresponding Correlation Values

Correlation Value	Correlation Relationship
0	Node is outside the region
1	Node is inside the region with voltage direction pointing towards the region.
-1	Node is inside the region with voltage direction pointing away from the region.

As the difference between the initial monitoring data of the fault node and that of the standard node is insignificant, it is challenging to conduct fault detection directly. Therefore, the method of regional differential processing is introduced to process the initial monitoring data, which is used to increase the difference between the fault and ordinary nodes. The regional differential matrix for each characteristic quantity is calculated as follows:

$$T_i = AT_i \quad (7)$$

In the above equation, A is the network association matrix and T_i is the column vector composed of the characteristic data monitored by each node. The state monitoring matrix for a single characteristic quantity in a single period is calculated as follows:

$$C_i = \left| A^T \right| R_i \quad (8)$$

The state monitoring matrix iC for a single characteristic quantity in a single time is extended spatially to form a state monitoring matrix iW for a single time with multiple characteristic quantities:

$$W_i = \left[C_1 C_2 \cdots C_n \right] \quad (9)$$

iW is then extended temporally to form a high-dimensional temporal-spatial state monitoring matrix W with multiple periods and characteristic quantities:

$$W = \left[W_1 W_2 \cdots W_n \right] \quad (10)$$

In Equations 4-6, the time length of matrices C_i W_i is one power frequency cycle, and the matrix W is formed by extending 5 matrices W_i .

The final step of the algorithm is to optimize the fault diagnosis algorithm using classification methods such as SVM and DT. Specifically, using extracted and simplified features, the algorithm processes accurate data and performs data segmentation. The data samples are then used for training to classify different fault states in the power grid system. Based on the test samples, the algorithm's performance is evaluated using the accuracy, precision and recall rate to choose the best classification method to diagnose the power system fault. Chapter 4 gives the optimization results of SVM and DT models.

4. Experimental Results

4.1Sample Characteristics

To verify the effectiveness and accuracy of the proposed method, MATLAB software is used to conduct experiments on the actual power system data set. The dataset contains 1500 samples of power outage data, and each sample has 16 features that represent the electrical characteristics of the power grid system.

The results of descriptive statistics for the entire sample data after pre-processing are presented in Table 2.

Table 2 Descriptive Statistics

Variable	Mean	Standard Deviation	Median	Kurtosis	Skewness
Maximum voltage	0.496	0.405	0.198	-1.924	0.016
Minimum voltage	0.171	0.132	0.14	-0.323	0.847
Voltage mean	0.218	0.191	0.149	-0.161	1.014
Voltage standard deviation	0.025	0.033	0	-0.683	0.917
Maximum current	0.52	0.428	0.199	-1.957	-0.005
Minimum current	0.099	0.058	0.101	-1.183	0.006
Current mean	0.17	0.132	0.136	-0.237	0.886
Current standard deviation	0.062	0.081	0	-0.538	0.969
Active power	0.428	0.424	0.035	-1.959	0.046
Reactive power	0.017	0.018	0.011	2.249	1.585
Apparent power	0.338	0.172	0.342	-0.826	-0.035
Power factor	0.001	0.01	0	1194.378	33.307
Power grid frequency	0.5	0.287	0.504	-1.206	-0.009
Frequency deviation	0.101	0.057	0.102	-1.193	-0.006
Transformer temperature	0.494	0.287	0.499	-1.164	0.003
Transformer load rate	0.098	0.058	0.096	-1.211	0.028
Fault discrimination	0.204	0.5	0	-2.002	1.016

These statistics show that the range of maximum and minimum voltages is extensive, with an average of 0.218 and a standard deviation of 0.191. The voltage standard deviation is relatively small, with a mean of 0.025 and a standard deviation of 0.033.

Similarly, there is a wide range of maximum and minimum currents, with a mean of 0.17 and a standard deviation of 0.132.

The active power range is wide, the mean value is 0.428, and the standard deviation is 0.424. Reactive power, however, has a minimal range, with a mean of 0.017 and a standard deviation of 0.018.

The mean value of the power factor is minimal, 0.001, and the standard deviation is 0.01, indicating that the use of power in the system is relatively consistent.

The average frequency of the grid is 0.5, with a standard deviation of 0.287, and the average temperature of the transformer is 0.494, with a standard deviation of 0.287.

It is worth noting that the kurtosis of the power factor is very high, at 1194.378, indicating that the kurtosis is very high.

This suggests that there is likely a particular and consistent pattern of power usage in the system. The power grid frequency has a negative skewness, meaning there are more outliers on the lower end of the scale.

In contrast, the transformer temperature has a positive skewness, indicating more outliers on the higher end of the scale.

Some variables, such as the reactive power, have minimal ranges and standard deviations. This means these variables are consistent and do not vary much in the data set.

Others, such as the active power and the maximum voltage, have more comprehensive ranges and extensive standard deviations, indicating that these variables can have more variability in their values.

It is also important to note that some variables, such as the power factor, have highly skewed or not normally distributed values. This can have implications for statistical analyses and modeling, as these techniques often assume normal distributions.

Careful consideration should be given to the appropriate statistical methods for each variable based on its distribution.

These descriptive statistics provide a helpful summary of the data and can inform further analysis and modeling of the electrical grid system. The wide range and variability of some variables suggest that important factors may be at play in the system's functioning that should be examined in greater detail. Additionally, the non-normal distributions of some variables highlight the need for caution in statistical analyses using these variables and the need for alternative approaches or transformations.

4.2Traditional Logit Model Is Applied to Diagnose the Faults in Power Grid System

The Logit model is a probability-based statistical model commonly used to describe and predict the probability distribution of binary or multiclass classification problems. The model transforms the classification problem into a logical function that inputs input variables and outputs the corresponding classification probabilities.

The most commonly used model in Logit is the binary Logit regression model. This model assumes that the response variable y follows a binomial distribution, with only two possible outcomes: success or failure, presence or absence, etc. This distribution transforms the probability density function into a logarithmic odds function, which can be described using a linear combination of input variables, i.e.,

$$l\left(p\right)=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_nx_n+\chi+\varepsilon \quad (11)$$

X$_1$ to *Xn* are the model's variables. The inverse function of this function is the sigmoid function, which transforms the logit into the probability of the event occurring, i.e.,

$$p = \frac{1}{1 + \exp(-l)} \tag{12}$$

The Logit model can be applied in many fields, such as medicine, finance, social sciences, etc. Its main advantage is that it can control the impact of multiple input variables, classify and predict multiple variables, and provide interpretable results. This article also uses this algorithm for fault analysis based on power outage data.

The ROC curve and confusion matrix for the predicted results of the logit model are shown in Figure 1.

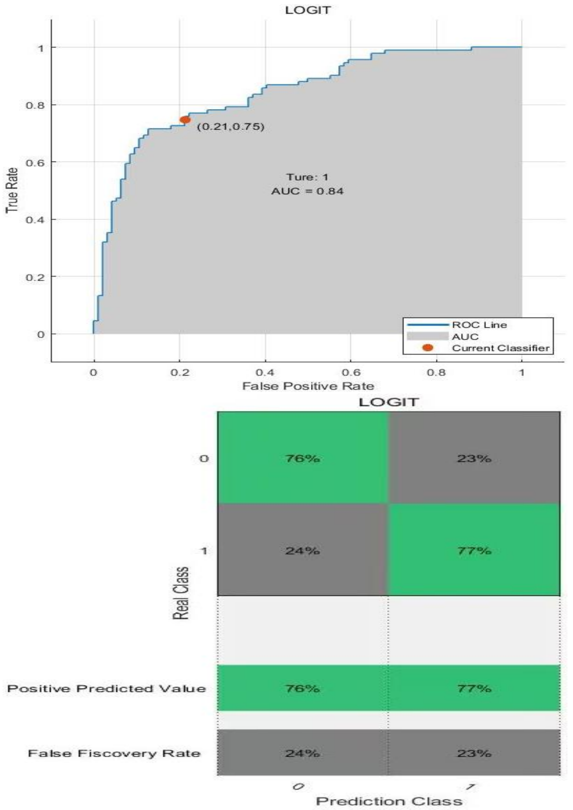

Figure 1 ROC Curve and Confusion Matrix of Model Prediction Results

As shown in the above ROC curve, the model performance is poor, with a prediction accuracy of only 84%. Further analysis of the confusion matrix reveals that the model has a high false positive rate and false negative rate, indicating that traditional Logit analysis methods may not be able to effectively adapt to algorithm optimization and prediction on a large-scale data basis. Therefore, this article selects heuristic algorithms to optimize the prediction model further.

4.3Traditional SVM Algorithm Is Applied to Diagnose the Faults in Power Grid System

SVM is a well-known algorithm for classification and regression analysis. SVM can process high-dimensional nonlinear data and be used in fault diagnosis of power systems. The traditional support vector machine model consists of support vectors defining the feature space's hyperplane. The support vector machine algorithm seeks an optimal hyperplane that maximizes the margin between the support vector and the decision boundary between different classes. The author needs to first extract the relevant features from the power grid data to apply the SVM algorithm to diagnose faults in power grid systems. Once the features are extracted, the author can train a traditional SVM model to classify the faults in the power grid system. The support vector machine model will learn the characteristics of different types of faults and use them to make predictions about new data. The performance of support vector machine models can be evaluated using metrics such as classification accuracy, specificity, and sensitivity.SVM model parameters are set as shown in Table 3.

Table 3 SVM Model Parameter Setting

Project	Value
Default setting	Fine Gaussian SVM
Kernel function	Gaussian
Kernel scale	1
Box constraint level	1
Box constraint level:	Box constraint level:
Normalized data	true

The ROC curve and confusion matrix for the predicted results of the SVM model are shown in Figure 2.

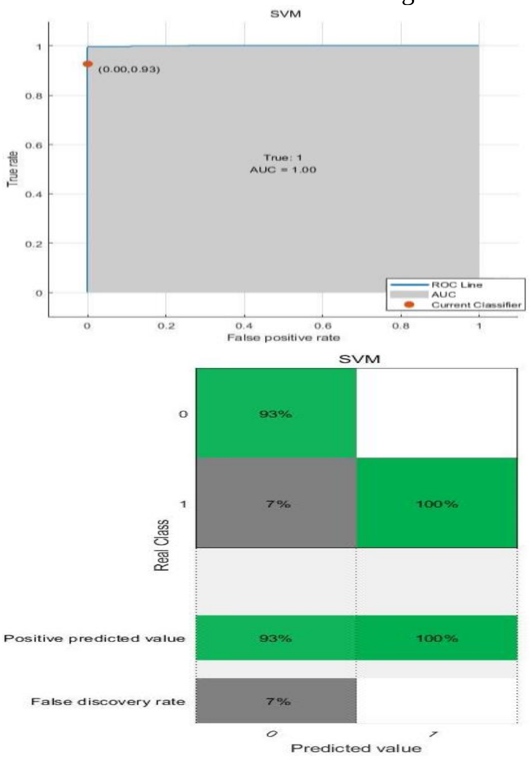

Figure 2 ROC Curve and Confusion Matrix of Model Prediction Results

As shown in the figure above, the results of the SVM model constructed in this article were well-performing. The ROC curve reflects the excellent prediction ability of the model, which can achieve a high actual rate with a small sacrifice of false positive rate. According to the selected current classifier of the model, the prediction accuracy of the model is 93%.

At the same time, the confusion matrix results show that all the predictions labeled as faults were correctly predicted, while 93% of the predictions labeled as non-faults were accurately predicted, and 7% were false negatives.

A certain proportion of false negatives will affect the overall effectiveness of the sample prediction, leading to the failure of some fault information being alerted in advance. However, a 0% false positive rate would prevent unnecessary warning information from interfering with the system's operation. Combined with the performance of the ROC curve, further control of false negatives will inevitably lead to a significant increase in false positives. Therefore, maintaining the existing results of the model is conducive to its application in power system operation.

Based on the results, the SVM model constructed in this article has high predictive accuracy and can effectively detect faults in the power grid system. The model's excellent performance and low false positive rate make it a valuable tool for fault diagnosis and optimization in the power grid industry. Data analysis techniques and machine learning continue to advance, and the model has the potential to improve power grid systems' efficiency further.

Finally, the predictive ability of the sample data was organized, and the performance of the model was composed, as shown in table 4:

Table 4 Model Performance

	Accuracy	Recall	Precision	F1
Training Set	0.9800	0.9595	0.9516	0.9748
Cross-validation Set	0.9771	0.9626	0.9639	0.9997
Test Set	0.9616	0.9805	0.9984	0.9515

The high accuracy of this model indicates its reliability in predicting the operating status of electrical equipment. Furthermore, by examining the recall and precision rates, it can be observed that the model has a high recall rate but a lower precision rate in the test set. This may suggest that the model has some false negatives when predicting faults. In addition, the F1 score is a measure that considers both recall and accuracy, and the model's F1 score of 0.9515 on the test set indicates that the model has some accuracy and reliability in predicting electrical equipment failures.

D. Improved DT algorithm is applied to diagnose the faults in a power grid system

Decision tree (DT) is another famous classification and regression analysis algorithm. DT recursively splits the data into different groups using the most informative features. The final model is a tree structure that humans can easily interpret.

The DT algorithm can also be applied to diagnose faults in power grid systems. The main advantage of the DT algorithm is its ability to handle numerical and categorical data and detect complex nonlinear relationships between the features. However, the traditional DT algorithm has some limitations, such as sensitivity to small perturbations in the data and overfitting to the training data. The improved DT algorithm can be applied to diagnose faults in power grid systems by selecting informative features and training a decision tree model.

To optimize the DT model, this paper further processes the data used by the DT model to improve the model analysis performance. To meet the requirement of the DT model, this paper introduces the correlation coefficient calculation and uses the correlation coefficient method as the data basis for feature selection. Then ISOMAP is used to reduce the data dimension, the neighborhood parameter is 5, and ARPACK decomposition is used. Floyd-Warshall algorithm and brute force search method calculate the shortest path.

The parameter settings for the DT model are shown in Table 5.

Table 5 DT Model Parameter Setting

Project	Value
Setting:	Fine-tuning Tree
Fine-tuning Tree	100
Splitting criterion	Gini diversity index

The ROC curve and confusion matrix for the predicted results of the DT model are shown in Figure 3.

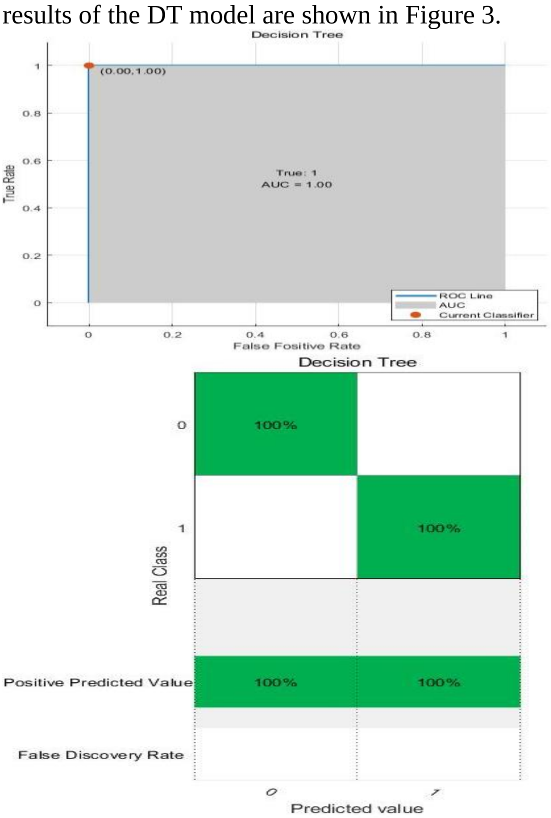

Figure 3 ROC Curve and Confusion Matrix of Model Prediction Results

Finally, the predictive ability of the sample data was organized, and the performance of the model was composed, as shown in table 6:

Table 6 Model Performance

	Accuracy	Recall	Precision	F1

Training Set	1	1	1	1
Cross-validation Set	1	1	1	1
Test Set	1	1	1	1

The performance of this model is excellent.It achieved high accuracy, recall, precision, and F1 scores in all the datasets (training, cross-validation, and testing sets). This indicates the model performs well on all datasets and can accurately classify and predict the data. The DT model can better complete the prediction task than the SVM model.

5. Discussion

This paper proposes an improved fault diagnosis algorithm based on the DT algorithm and large-scale blackout data (N=1500) to enhance the accuracy and efficiency of fault diagnosis, resulting in 100% accuracy in blackout fault prediction based on electrical parameters. Experimental results show that the proposed method has better performance (93%) than the traditional SVM method, which is attributed to the effectiveness of the DT algorithm in predicting such models with the same raw data and pre-processing methods. Moreover, the exceptionally high accuracy of the model suggests that standardizing the raw data of multidimensional electrical parameters can significantly reduce errors caused by non-matching among data of different dimensions and thus optimize the model's ability to utilize raw data.

However, the proposed method has some limitations and needs further improvement. First, this method only considers the diagnosis of single-phase fault and does not consider the diagnosis of multi-phase fault. Secondly, the proposed method only uses the SVM and DT algorithms, while other algorithms, such as artificial neural networks and deep learning networks, may also be helpful for fault diagnosis. Finally, the proposed method focuses on optimizing parameters or features while integrating with other techniques, such as expert reasoning or knowledge-based systems, which may further improve the performance of the fault diagnosis algorithm.

In future work, the author will improve the proposed method by considering the multi-phase faults and integrating them with other techniques. Moreover, the proposed method will be applied to more real power grid systems to validate its effectiveness and generalization ability.

References

[1] Liu Hengyong, Guo Lu, Liu Yongli & Huang Ziqi. (2020).Research on Efficient Collection Method of Blackout Data in Distribution Network. Journal of Physics: Conference Series(5). doi:10.1088/1742-6596/1549/5/052030.

[2] Supriya Chinthavali. (2019). Seattle City Light Standardizes Outage Data. Transmission & Distribution World.

[3] Chunyan Shuai,Hengcheng Yang,Xin Ouyang,Mingwei He,Zeweiyi Gong & Wanneng Shu.(2019).Analysis and Identification of Power Blackout-Sensitive Users by Using Big Data in the Energy System.. IEEE Access.

[4] Yue Meng, Toto Tami, Jensen Michael P., Giangrande Scott E. & Lofaro Robert. (2018).A Bayesian Approach-Based Outage Prediction in Electric Utility Systems Using Radar Measurement Data. IEEE Transactions on Smart Grid(6). doi:10.1109/tsg.2017.2704288.

[5] Liu Feng,Guo Jinpeng,Zhang Xuemin,Hou Yunhe & Mei Shengwei.(2018).Mitigating the Risk of Cascading Blackouts: A Data Inference Based Maintenance Method. IEEE Access. doi:10.1109/access.2018.2855153.

[6] Jun Fu, Xin Xu, Zhijie Sun, Li Wang, Dongmei Gong & Lingyu Zhang. (2018). Model Construction of Early Warning for Frequently Outage Complaint Based on Data Mining. MATEC Web of Conferences. doi:10.1051/matecconf/201817301002.

[7] Gurara Daniel & Tessema Dawit.(2018).Losing to Blackouts: Evidence from Firm-Level Data. IMF Working Papers(159). doi:10.5089/9781484363973.001.

[8] Erwin Normanyo, and Godwin Diamenu."Predicting Reliability of Electric Power Distribution Grid Using Historical Outage Data." American Journal of Electrical Power and Energy Systems 11.4(2022). doi:10.11648/J.EPES.20221104.11.

[9] Han Yi et al." Improved Fault Location Algorithm for Radial Distribution Network Based on Power Failure Information." Journal of Physics: Conference Series 1848.1(2021). doi:10.1088/1742-6596/1848/1/012049.

[10] Liu Hengyong et al." Research on Efficient Collection Method of Blackout Data in Distribution Network." Journal of Physics: Conference Series 1549.5(2020). doi:10.1088/1742-6596/1549/5/052030.

[11] Qingqing HAO, and Qun YU."Research on blackout simulation model considering hidden failures and reclosing." IOP Conference Series: Earth and Environmental Science 431. (2020). doi:10.1088/1755-1315/431/1/012005.

[12] A. Sathish Kumar, et al."Contingency Analysis of Fault and Minimization of Power System Outage using Fuzzy Controller." International Journal of Innovative Technology and Exploring Engineering (IJITEE) 9.1(2019).

[13] Supriya Chinthavali."Seattle City Light Standardizes Outage Data." Transmission & Distribution World .(2019).

[14] Sroka Krzysztof,and Złotecka Daria."The risk of significant blackout failures in power systems." Archives of Electrical Engineering 68.2(2019). doi:10.24425/aee.2019.128277.

[15] C.C. Montanari, and P. Dimitriou."The IAEA stopping power database, following the trends in stopping power of ions in matter." Nuclear Inst. and Methods in Physics Research, B 408. (2017). doi:10.1016/j.nimb.2017.03.138.

[16] Michael M. Li, and Brijesh Verma."Nonlinear curve fitting to stopping power data using RBF neural networks." Expert Systems With Applications 45. (2016). doi:10.1016/j.eswa.2015.09.033.

[17] Stanko Novakovic et al." An Analysis of Load Imbalance in Scale-out Data Serving." ACM SIGMETRICS Performance Evaluation Review 44.1(2016). doi:10.1145/2964791.2901501.

[18] Bahiru Egziabiher, Scott Thomsen, and John Simmons."Seattle City Light Shares Outage Data Initiative: Collaboration, standards and APIs will improve restoration and drive the next generation of utilities." Transmission & Distribution World: The Information Leader Serving the Worldwide Power-Delivery Industry 68.2(2016).