

Power fault analysis and classification based on correlation analysis and random forest algorithm

Jiahui Yan^{1,3,*}, Xinyu Zhu²

¹International Education College, Zhengzhou university of light industry, Henan, Zhengzhou, 450003, China

²International Education College, Zhengzhou university of light industry, Henan, Zhengzhou, 450003, China

³1620835699@qq.com

*Corresponding author

Abstract. The electric power system is one of the indispensable infrastructures in modern society, and its stable operation is crucial for guaranteeing economic development and social security. By calculating the correlation coefficient and drawing the correlation heat map, we find that there is a correlation between the voltage at each location and the occurrence of electrical faults, and at the same time, there is also a certain correlation between various types of electrical faults and the voltage at each location. These results indicate that there are complex and close interrelationships between different locations and between different types in the power system. In order to better identify and classify different types of electrical faults, we used the random forest model for prediction. The model is used to classify whether an electrical fault occurs or not and to determine its category. The prediction results show that the Random Forest achieved 99.63% in classifying whether an electrical fault occurred or not, and 91.26% in classifying the category of electrical faults. This shows that the Random Forest algorithm model is able to detect electrical faults well, as well as judge the category of electrical faults. In summary, this paper provides insight into the operation of the power system by studying the correlation between the voltage at various locations in the power system and the different types of electrical faults and using the random forest model to make predictions. These findings are of great significance in ensuring the stable operation of power systems.

Keywords: Power Fault Analysis, Correlation Analysis, Random Forest.

1. Introduction

Power system is one of the indispensable infrastructures in modern society, and its stable operation is crucial to ensure economic development and social security. However, there are various faults in the power system, such as short circuit, overload, undervoltage, etc. These faults not only affect the normal operation of the power system, but also may lead to equipment damage, accidents and even affect people's daily life [1,2]. Therefore, the detection and diagnosis of power faults are of great significance.

The traditional power fault detection method mainly relies on professional technicians to carry out inspections or use sensors to collect data for analysis and judgement, which is inefficient, costly, and susceptible to human interference and other problems [3,4]. In recent years, machine learning

technology has been widely used in the field of electric power fault detection and has achieved remarkable results [5].

Machine learning technology can build a model by learning and analysing a large amount of historical data, and use the model to predict and judge new data [6]. In terms of power fault detection, machine learning technology can achieve rapid detection and diagnosis of potential faults through in-depth analysis and mining of power grid data. For example, machine learning algorithms can be used to classify and identify various faults in the power system, thus achieving rapid fault location and processing [7,8]. At the same time, machine learning technology can also help the power system to carry out intelligent operation and maintenance, and improve the reliability and security of the power grid.

Machine learning technology has a broad application prospect in the field of power fault detection, and will make an important contribution to the stable operation and sustainable development of the power system. This paper analyses and classifies power faults based on correlation analysis and random forest algorithm, which provides a certain research basis for subsequent research.

2. Structural design

In this paper, we have selected the open source electrical fault detection dataset, which contains two sets of data, Ia, Ib, Ic, and Va, Vb, Vc denote the voltages at each location, in addition to the voltage conditions at each location, the first set of data records whether or not the electrical relay is faulty, and the second set of data records the location where the electrical fault occurs. We can use the first set of data to predict whether an electrical fault will occur as a binary classification problem. At the same time we can use the second set of data to predict the class of electrical faults, as a multiclassification problem. [G C B A] is the output, and different [G C B A] indicates faults at different locations, and the specific categories are shown in Table 1.

Table 1. Model evaluation indicators.

[G C B A]	Classification
[0 0 0 0]	trouble-free
[1 0 0 1]	LG fault (between phase A and earth)
[0 0 1 1]	LL fault (between phase A and phase B)
[1 0 1 1]	LLG fault (between phase A, B and earth)
[0 1 1 1]	LLL fault (between all three phases)
[1 1 1 1]	LLLG fault (three phase symmetrical fault)

3. Relevance analysis

Pearson correlation analysis is a statistical method used to measure the strength of a linear relationship between two variables. It is based on the Pearson correlation coefficient, which ranges from -1 to +1 and indicates the strength and direction of the relationship between two variables. If the correlation coefficient is positive, the two variables are positively correlated, i.e., one variable increases when the other increases; if the correlation coefficient is negative, the two variables are negatively correlated, i.e., one variable increases when the other decreases; and if the correlation coefficient is close to zero, there is no linear relationship between the two variables.

A correlation heat map is a visualisation tool used to show the correlation between multiple variables. It indicates different levels of correlation through colour coding, and correlation analysis is carried out on the first and second sets of data to explore the correlation between the voltage at each location and the presence of electrical faults and the correlation between the various types of electrical faults and the voltage at each location, respectively. The correlation coefficients are calculated and correlation heat maps are plotted, and the results are shown in Figures 1 and 2.

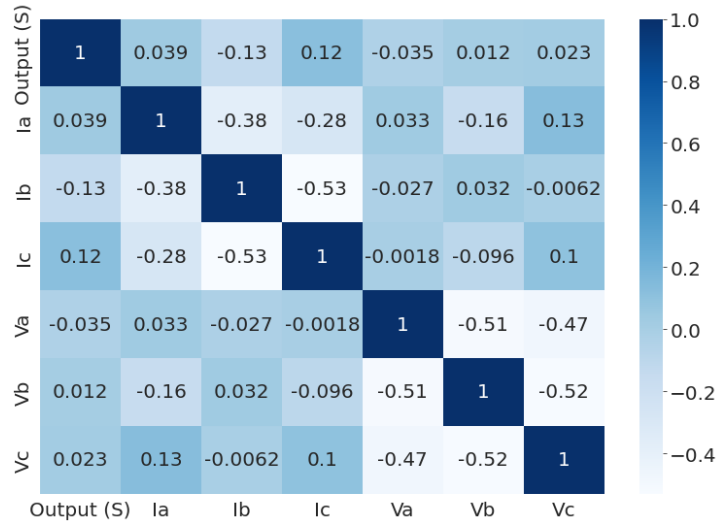


Figure 1. Heat map of correlation between voltage and electrical faults at each position.
 (Photo credit : Original)

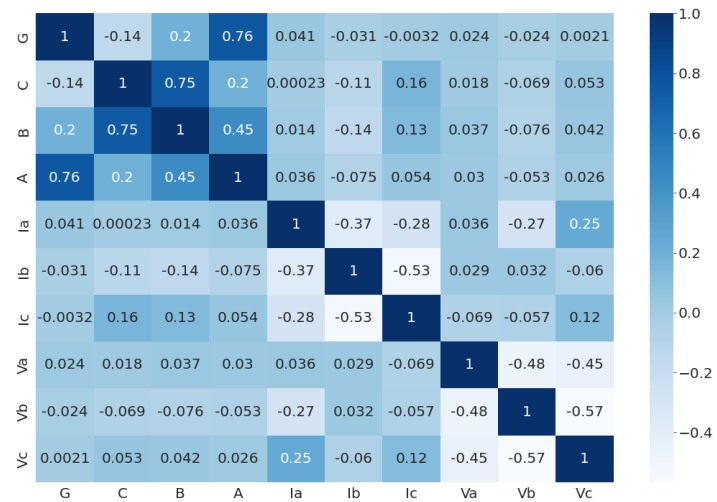


Figure 2. Heat map of correlation between voltage at each position and various types of electrical faults.
 (Photo credit: Original)

From the correlation heat map, it can be seen that there is a correlation between the voltage at each location and the occurrence of electrical faults, as well as a correlation between various types of electrical faults and the voltage at each location. Based on the conclusion of the correlation analysis, this paper uses the random forest machine learning algorithm to explore the quantitative conditions that exist between the variables.

4. Random forest model

Random forest is an integrated learning method that achieves tasks such as classification, regression and feature selection by constructing multiple decision trees. The core idea of the random forest algorithm is to combine multiple decision trees together and obtain the final result by voting or averaging [9].

The training process of the Random Forest algorithm, each decision tree is trained based on a different sample set and feature set. In terms of sample set, Random Forest uses self-service sampling (bootstrap sampling) technique, i.e., samples are taken from the original dataset with put back to construct a new training dataset; in terms of feature set, Random Forest randomly selects a feature subset

for each node to be partitioned [10]. The random forest algorithm has better generalisation ability and robustness compared to a single decision tree, it can effectively avoid the overfitting problem, and it has a better ability to handle high-dimensional data and large-scale data. In addition, since each decision tree is independent of each other, the processing can be parallelised. In conclusion, Random Forest algorithm is a powerful and flexible machine learning method, which is widely used in various data mining and machine learning tasks.

5. Result

In this paper, two sets of classification tasks are performed, the first one is to classify whether an electrical fault has occurred or not based on the voltages at each location (Ia, Ib, Ic, Va, Vb, Vc) and the second one is to classify the category of electrical faults based on the voltages at each location (Ia, Ib, Ic, Va, Vb, Vc).

After getting the dataset, the training and test sets are divided in the ratio of 7:3, 70% of the data is used for model training and 30% of the data is used for model testing, the predictive effectiveness of the model is evaluated using accuracy, recall and precision and the confusion matrix of the prediction results is output. The results of the two sets of classification results are shown in Table 2 and Figure 3. The confusion matrix of the prediction results is shown in Fig. 4 and Fig. 5.

Table 2. Modelling evaluation.

	Accuracy	Precision	Recall
Malfunction	99.63%	99.73%	99.46%
Fault classification	91.26%	85.48%	92.91%

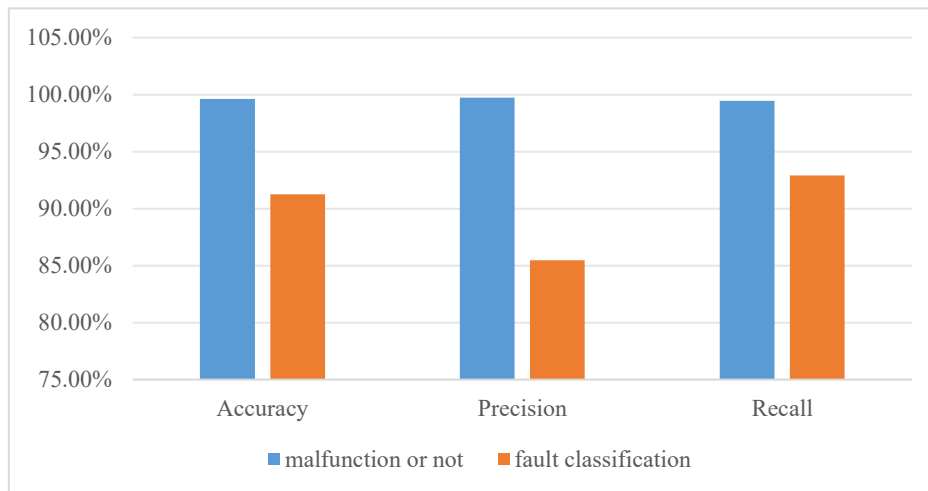


Figure 3. Modelling evaluation.
 (Photo credit: Original)

From the prediction results, it can be seen that in the classification of whether an electrical fault occurs or not, the prediction results of Random Forest reached 99.63%, and in the classification of the category of electrical faults, the prediction results of Random Forest reached 91.26%, the Random Forest algorithm model is able to detect the electrical faults as well as to judge the category of electrical faults very well.

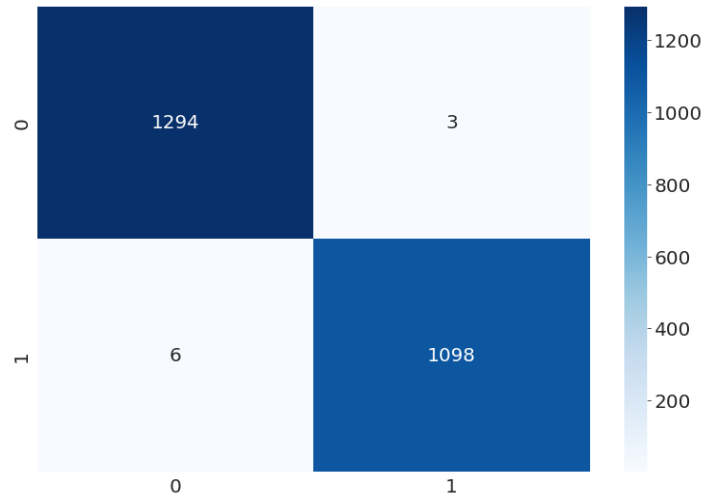


Figure 4. Confusion matrix.
 (Photo credit: Original)

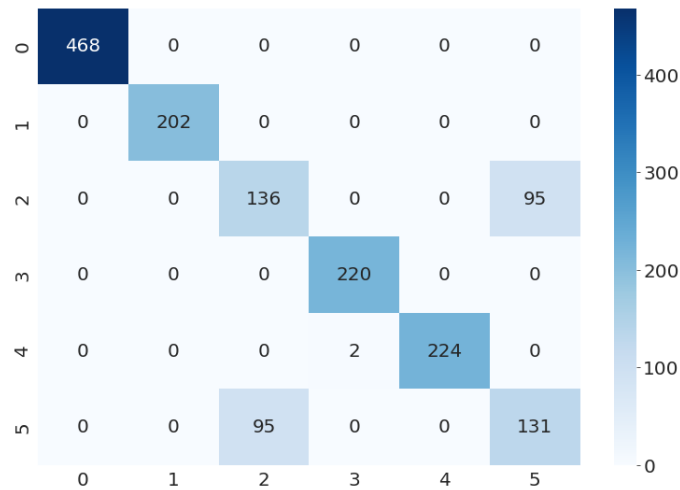


Figure 5. Confusion matrix.
 (Photo credit: Original)

From the results of the confusion matrix, it can be seen that the vast majority of electrical fault detections and electrical fault category detections are correct, indicating that the Random Forest model plays a good predictive role in electrical fault detection.

6. Conclusion

The electric power system is one of the indispensable infrastructures of modern society, and its stable operation is crucial to guaranteeing economic development and social security. In this paper, we draw a series of conclusions by exploring the correlation between the voltage at each location and whether an electrical fault occurs, as well as the correlation between various types of electrical faults and the voltage at each location.

First, we calculated the correlation coefficients and plotted a correlation heat map. From the correlation heat map, it can be seen that there is a certain degree of correlation between the voltage at each location and the occurrence of electrical faults. Also, there is some correlation between various types of electrical faults and the voltage at each location.

Next, we used a random forest model to classify electrical faults and determine their categories. The prediction results show that the Random Forest algorithm achieved 99.63% in classifying whether an

electrical fault occurred or not, and 91.26% in classifying the category of electrical faults. This shows that by using the random forest algorithm model can be very good for the detection of electrical faults as well as judge its category.

In summary, this paper draws a series of useful conclusions by monitoring and analysing the power system. These conclusions are of great significance for the stable operation of the power system as well as for safeguarding economic development and social security. We believe that these conclusions will provide more useful guidance for the optimisation and upgrading of power systems in future research.

References

- [1] Biomechanics A B A .Retracted: A Novel Framework for Automation Technology Based on Machine Vision and Robotics in Electrical Power Inspection Processing.[J].Applied bionics and biomechanics,2024,20249831452-9831452.
- [2] Jiefei Z .Management of Electronic and Electrical Quality Inspection at the Angle of Just and Public Trust[J].Journal of Electronic Research and Application,2023,1(2):
- [3] Accurate Readings in Challenging Electrical Inspection Environments: Meet FLIR's DM93-2 and IM75-2 Meters with METERLINK® App[J].M2 Presswire,2023,
- [4] Wang Y .Electrical Control Equipment Patrol Inspection Method Based on High Quality Image Recognition Technology[J].Traitement du Signal,2023,40(2):
- [5] Lijun D ,Naichao C ,Jiawen L , et al.A review of indoor-orbital electrical inspection robots in substations[J].Industrial Robot: the international journal of robotics research and application,2023,50(2):337-352..
- [6] Jieming P ,Zaifeng Y ,Kit H S Y , et al.Non-destructive online seal integrity inspection utilizing autoencoder-based electrical capacitance tomography for product packaging assurance[J].Food Packaging and Shelf Life,2022,33
- [7] Huang X .Design of motor control system for over-current protection inspection and voltage regulating device of explosion-proof electrical appliances[J].Smart Systems and Green Energy,2021,3(1):
- [8] Sungyoon R ,Younghoon S ,Yusin Y .The novel feature based inspection technique that can detect defects that can affect the deterioration of the electrical properties of semiconductor devices[J].Microscopy and Microanalysis,2021,27(S1):792-793.
- [9] André H L ,Håkon K ,L C S , et al.An autonomous drone-based system for inspection of electrical substations[J].International Journal of Advanced Robotic Systems, 2021,18(2): 17298814211002973-17298814211002973.
- [10] Yurong F ,Kwaiwa T ,Shengyang C , et al.Learning-Based Autonomous UAV System for Electrical and Mechanical (EM) Device Inspection[J].Sensors,2021,21(4):1385-1385.