Machine Learning Based Fault Type Identification In the Active Distribution Network

Baicong Sun, Hengxu Zhang*, Fang Shi
School of Electric Engineering
Shandong University
Jinan, China
sunbaicong@mail.sdu.edu.cn, zhanghx@sdu.edu.cn, shifang@sdu.edu.cn

***Abstract*—To realize the intelligent of the distribution network, it is necessary to identify the fault type accurately. This paper presents the fault type identification method based on machine learning in active distribution networks. The process of machine learning is divided into four steps: data preparation, data preprocessing, feature extraction and model training. When preparing data, a method of generating fault scenarios in the batch of simulation experiments is presented. The IEEE34 Bus System is built in PSCAD to complete the data preparation for machine learning. Variation multiples of voltage and current are extracted as the features to describe the fault type. Various machine learning models are trained by cross-validation method to get the accuracy of identification. The application of decision tree in fault type identification is presented in the form of a tree diagram. The result of fault type identification is shown by the confusion matrix of the decision tree. All the test results show that the proposed fault identifiers can identify all kinds of fault types in the distribution network.**

Keywords—machine learning; fault type identification; active distribution network; batch simulation; feature extraction

I. INTRODUCTION

In modern power grids, fast and accurate fault type identification is an essential operational requirement. Both relay protection and fault location require correct fault type information. There have been a large number of researches on fault classification in the power grid, mainly for the transmission network. Traditional fault type identification is achieved by setting threshold values and relying on logical relationships. The fault location and fault type are inferred based on the logic of the protection and the experience of the operator. This process is difficult to describe by traditional mathematical methods. Artificial intelligence technology has the characteristics of simulating human and has been widely used in this field. The main implementation methods include neural network approach[1], fuzzy neural network[2], expert systems[3], genetic algorithms[4], and Petri net[5].

More than 80% of the faults come from the distribution network in the power system. Fast and accurate fault classification in the distribution network is significant for fault analysis and power restoration, which can effectively improve power supply reliability. In [6], a method for fault type identification using decision trees is proposed. However, the identification of fault types does not involve a specific phase. In [7], a fuzzy logic-based fault-type identification scheme for an unbalanced radial power distribution system has been proposed, which can identify ten types of short circuit faults. The fault diagnosis of distribution network still has the following problems: (1) information source data is huge; (2) uncertainty of information[8, 9]; (3) the selection of intelligent identification methods[10]. This process is complicated to describe by using mathematical techniques, and the artificial intelligence approach plays a vital role in fault type identification.

In this paper, machine learning models are used to realize the identification of fault types in the active distribution network. Firstly, the process of machine learning is introduced. In the fault data preparation, a method for generating fault data in batches in PSCAD is proposed[11], which avoids manual modification of parameters. In the PSCAD, the IEEE 34 Bus System of the distribution network is built to complete the data preparation for machine learning. In the data preprocessing process, the variation multiples of current and voltage are used to describe the fault feature. The extracted features are trained using various machine learning models. The results of fault type identification using decision trees are highlighted[12]. The digital electromagnetic transient simulations presented in this paper have been carried in PSCAD environment, and the machine learning models were done in the MATLAB environment.

II. THE PROCESS OF MACHINE LEARNING

Fig.1 shows the process of machine learning. Machine learning requires a large amount of data to be input. PSCAD is used to build the model, and Python scripts are used to control PSCAD to generate fault simulation data. Because raw data are often not formatted and have too much information, the data need to be preprocessed to extract the required fault feature data. Multiple machine learning algorithms are used to train the data. A trade-off is made between model speed, accuracy, and complexity to select the appropriate model and optimize the model based on the identified results.

978-1-5386-6243-4/19/$31.00 ©2019 IEEE

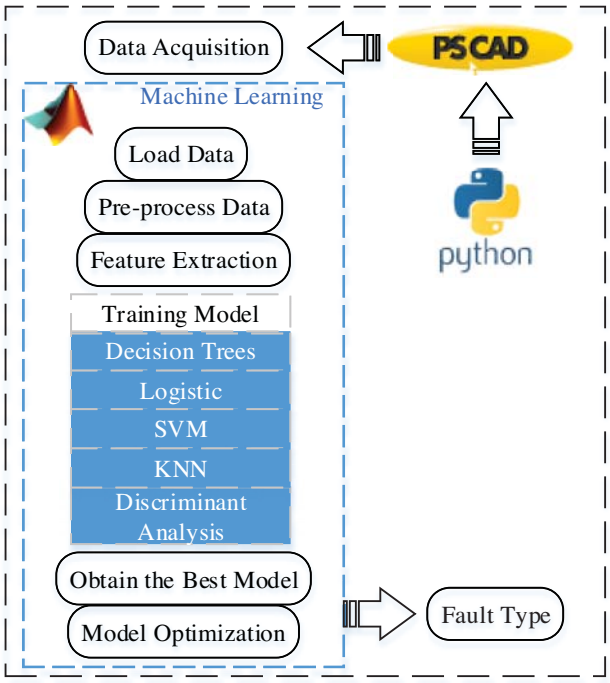


Fig. 1 the process of machine learning

III. DATA PREPARATION- PSCAD BATCH SIMULATION

When using the machine learning method to identify fault types, a large amount of fault data is required. To generate multiple-fault simulation scenarios, it is necessary to repeatedly modify some component parameters such as fault types and transition resistance. The process of manually changing the simulation model is cumbersome, time-consuming and error-prone. A batch generation method for fault simulation scenarios for PSCAD is proposed, and an IEEE34 Bus System is built to generate simulation data.

A. Batch Generation Method for PSCAD

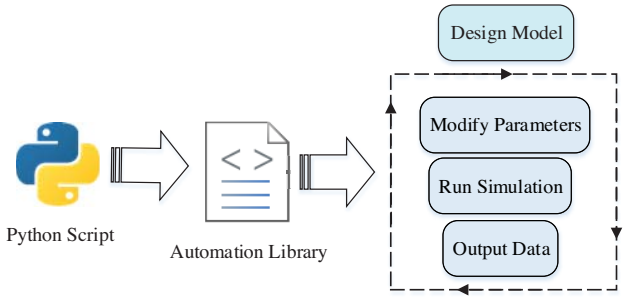


Fig. 2 the process of batch generation for PSCAD

The PSCAD automation library is called by the Python script, and the components are abstracted to realize the control of the simulation model. Fig.2 shows the process of batch generation for PSCAD. Complete the design of the model, run the simulation after modifying the parameters through Python script, process the data and save it to the disk, cycle the above process, and generate the fault simulation scene in batches.

Fig.3 shows the flow of control models by Python scripts. Import the configuration information of the PSCAD software, load the path of the model, get the component ID, and modify the component parameters. By controlling the layer enable, the fault location and the access of the distributed power supply are changed. After adjusting the model parameters, run the simulation. Copy the recorded files (CFG, DAT, HDR) from the default folder to the specified folder. The above process is executed cyclically until the fault simulation scenario is completed.

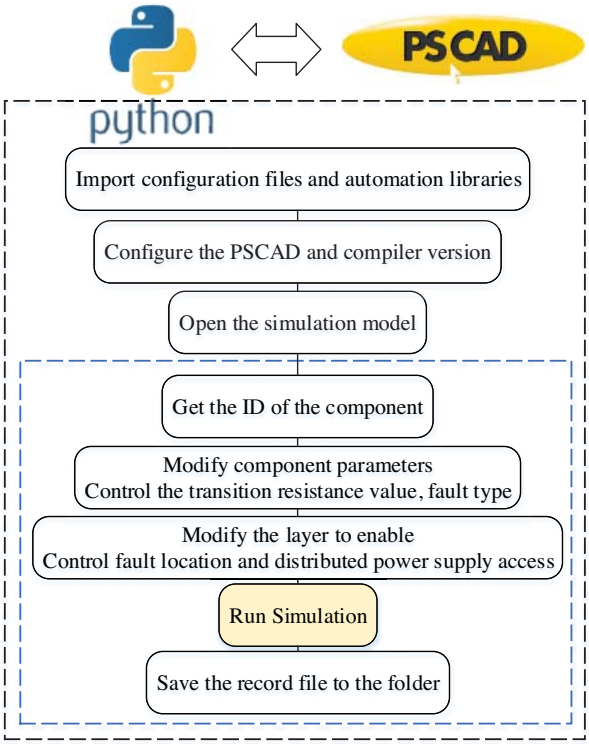


Fig. 3 control models by Python scripts

B. Simulation Model Information

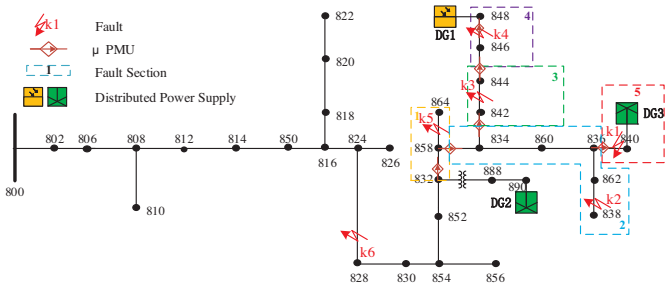


Fig. 4 IEEE 34 Bus System with distributed power supply

Fig.4 shows the IEEE 34 Bus System with distributed power supply. Phasor measurement units (PMU) are installed at nodes 832, 834, 836, 844, 848, and 858, respectively. The voltage and current signals are saved in the standard COMTRADE format. According to the location of the PMU, the IEEE 34 Bus System is divided into five sections. There are ten types of faults at six positions K1-K6, where K1-K5 corresponds to 5 sections and K6 is outside the section. The transition resistance is set to 0.0001, 1, 10, 30, 50, 100, 300, 500 ohm. Four hundred eighty sets of fault simulation scenarios are obtained in PSCAD by using the method of batch generation. Table 1 shows the details of the IEEE 34 Bus System.

TABLE I. DETAILS OF THE IEEE 34 BUS SYSTEM

No.	Configuration	Details
1	Transition resistance	0.0001, 1, 10, 30, 50, 100, 300, 500 ohm
2	Fault type	AG,BG, CG,AB, BC, CA, ABG, BCG, CAG, ABCG
3	Distributed power supply	Total penetration rate 55% DG1: Photovoltaic power generation 0.4MW DG2: Wind power generation 0.33MW DG3: Wind power 0.25MW
4	Neutral grounding method	Not grounded
5	Fault location	K1-K6
6	PMU installation location	Nodes 832, 834, 836, 844, 848, and 858

IV. DATA PREPROCESSING-THE FAULT FEATURE EXTRACTION

It is necessary to select appropriate features to achieve data dimensionality reduction and reduce the difficulty of learning. The COMTRADE file needs to be parsed before the fault feature extraction. The magnitude/phase angle of the voltage/current are calculated by using the analog quantity to extract the fault characteristics.

A. Processing of Recorded Data and Phasor Calculation

The data is processed according to the COMTRADE data format to obtain three-phase voltage analog quantity and three-phase current analog quantity. As shown in Fig.5, the three-phase voltage and three-phase current analog quantities are calculated to get the amplitude/phase angle of the three-phase voltage/current and the amplitude of the zero-sequence voltage in the MATLAB environment.

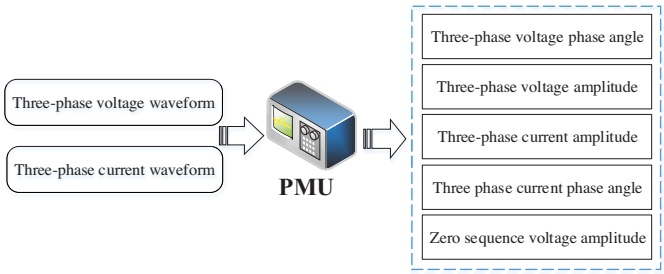
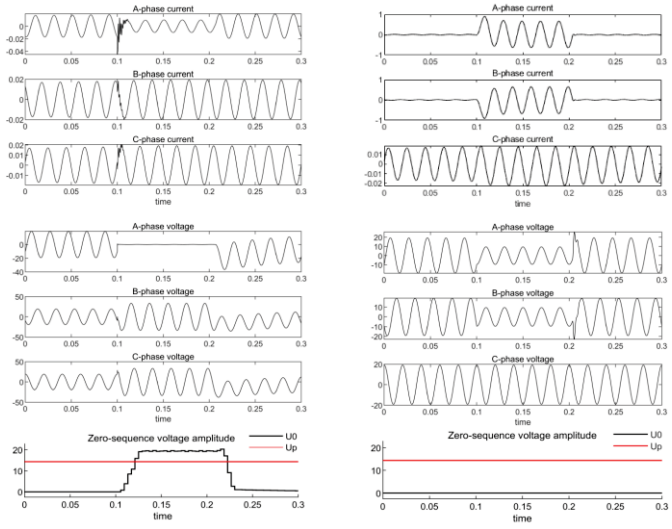


Fig. 5 the process of phasor calculation

B. Feature Extraction

Before feature extraction, it is necessary to observe the change in the electrical quantity at the time of the fault and to derive the factors associated with fault type. Fig.6 shows the three-phase current waveform, three-phase voltage waveform, and zero-sequence voltage amplitude for single-phase ground fault and two-phase short-circuit fault. It can be concluded from the observation that the fault characteristics of the distribution network can be expressed by the phase voltage amplitude, the phase current amplitude and the zero sequence voltage amplitude. Only the magnitude of the voltage and current values can't reflect the fault condition, and the multiple of the voltage and current changes are often more intuitive and reliable. Therefore, in the data pre-processing, the multiplication factor of the phase current amplitude change, the multiple of the phase voltage amplitude change, the ratio of the zero-sequence voltage amplitude to the system voltage are extracted as features.



(a) Single-phase earthing fault(AG) (b) Two-phase short circuit fault(AB)

Fig. 6 Three-phase voltage/current waveform, zero-sequence voltage amplitude

$$\begin{cases} RatioI_a = \dfrac{\left|\dot{I}_a\right|}{\left|\dot{I}_a^{[0]}\right|} \\ RatioI_b = \dfrac{\left|\dot{I}_b\right|}{\left|\dot{I}_b^{[0]}\right|} \\ RatioI_c = \dfrac{\left|\dot{I}_c\right|}{\left|\dot{I}_c^{[0]}\right|} \end{cases} \quad (1)$$

Where, I_a , I_b , I_c are phase current after failure; $I_a^{[0]}, I_b^{[0]}, I_c^{[0]}$ are phase current before failure; $RatioI_a$, $RatioI_b$, $RatioI_c$ are ratio of phase current after and before failure.

$$\begin{cases} RatioU_a = \dfrac{\left|\dot{U}_a\right|}{\left|\dot{U}_a^{[0]}\right|} \\ RatioU_b = \dfrac{\left|\dot{U}_b\right|}{\left|\dot{U}_b^{[0]}\right|} \\ RatioU_c = \dfrac{\left|\dot{U}_c\right|}{\left|\dot{U}_c^{[0]}\right|} \end{cases} \quad (2)$$

Where, U_a , U_b , U_c are phase voltage after failure; $U_a^{[0]}$, $U_b^{[0]}$, $U_c^{[0]}$ are phase voltage before failure; $RatioU_a$, $RatioU_b$, $RatioU_c$ are ratio of phase voltage after and before failure.

$$RatioU_0 = \frac{|U_0|}{|U_p|} \tag{3}$$

Where, U_0 is zero-sequence voltage, U_p is system voltage, $RatioU_0$ is the ratio of the zero sequence voltage amplitude to the system voltage amplitude.

V. MACHINE LEARNING MODEL

A. Model Selection

There are a variety of algorithms to choose when solving the same problem. Therefore, when using machine learning algorithms to solve problems, model selection is required. Since the generalization error cannot be directly obtained and the training error is not suitable as a standard due to the over-fitting phenomenon, it is necessary to evaluate the model and then select it. The test set is used to test the learner's ability to discriminate against new samples, and then the test error is used as an approximation of the generalization error. The dataset D needs to be appropriately processed to obtain the training set S and the test set T. Cross-validation is used to avoid over-fitting, and data set D is divided into mutually exclusive subsets of similar size, $D = D_1 \cup D_2 \cup ... \cup D_k$, $D_i \cap D_j = \varnothing (i \neq j)$. Each subset maintains consistency in data distribution. The union of k-1 subsets is used as the training set, and the remaining subsets are used as the test sets. The k-group training set and test set are used to perform k training and testing and return the mean of all test results finally. Fig.7 shows the schematic of the 5-fold cross validation.

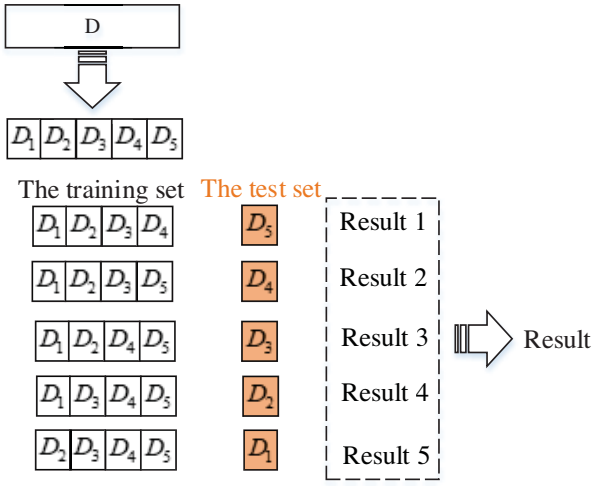


Fig. 7 5-fold cross validation

A variety of machine learning models are implemented in MATLAB, and the classification accuracy of each algorithm is shown in Table 2. When using machine learning for fault classification, use as few fault features as possible. When the multiple factors of the phase current amplitude change and the ratio of the zero-sequence voltage amplitude to the system voltage are selected, the accuracy rate one is obtained. When all seven feature quantities are selected, the accuracy rate two is achieved. It can be seen that in both cases, the accuracy of ensemble classifiers is the highest in all algorithms. In all of the individual algorithms, the decision tree has the highest accuracy of 98.2% when using four features. The next section will detail the application of decision trees in fault classification.

TABLE II. ACCURACY OF MACHINE LEARNING ALGORITHMS

No.	Algorithm	Accuracy rate 1	Accuracy rate 2
1	**Decision Tree**	**98.5%**	99.6%
2	Discriminant Analysis	97.5%	**100%**
3	Support Vector Classifiers	94.2%	98.3%
4	KNN	96.5%	99.4%
5	Ensemble Classifiers	**99.6%**	**100%**

B. Decision Tree

The difference between decision trees is the method of attribute selection. The Gini index is used to select the partitioning attribute, and the purity of the dataset D can be measured by the Gini index.

$$Gini(D) = \sum_{k=1}^{|y|} \sum_{k' \neq k} p_k p_{k'} = 1 - \sum_{k=1}^{|y|} p_k^{\,2} \tag{4}$$

The proportion of the k-th sample in the sample set D is $p_k (k = 1,2,...,|y|)$. The smaller the value of $Gini(D)$, the higher the purity of dataset D. Discrete attribute a has V possible values $\{a^1, a^2, ..., a^V\}$, and the dataset D is divided using a to generate V branch nodes. D^v is the set of samples with value a^v on attribute a. The branch nodes are given weights $|D^v| / |D|$ in consideration of the number of samples of different branch nodes. The Gini index of attribute a is defined as:

$$Gini\_index(D,a) = \sum_{v=1}^{V} \frac{|D^v|}{|D|} Gini(D^v) \tag{5}$$

In the candidate attribute set A, the attribute that minimizes the Gini index is selected as the optimal partition attribute.

$$a_* = \arg\min_{a \in A} Gini\_index(D,a) \tag{6}$$

Attribute selection is performed using the methods mentioned in the previous section. Fig.8 is the decision-tree formed using training samples.

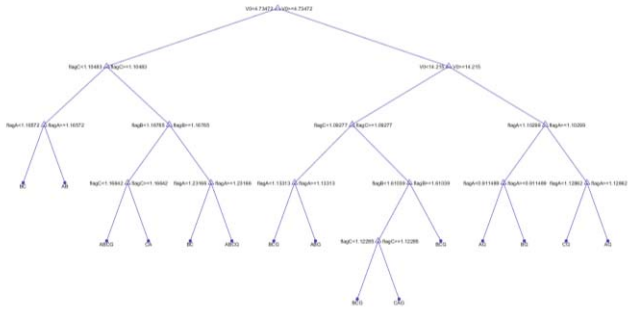


Fig. 8 Decision-tree formed using training samples

The confusion matrix of the decision tree is drawn in Fig.9. The error rate of the fault type ABCG is 2%, and the error rate of the fault type CA is 6%, and the error rate of the fault type CAG is 2%. The accuracy of the remaining fault types is 100%. The identification of various fault types can be visually seen through the confusion matrix.

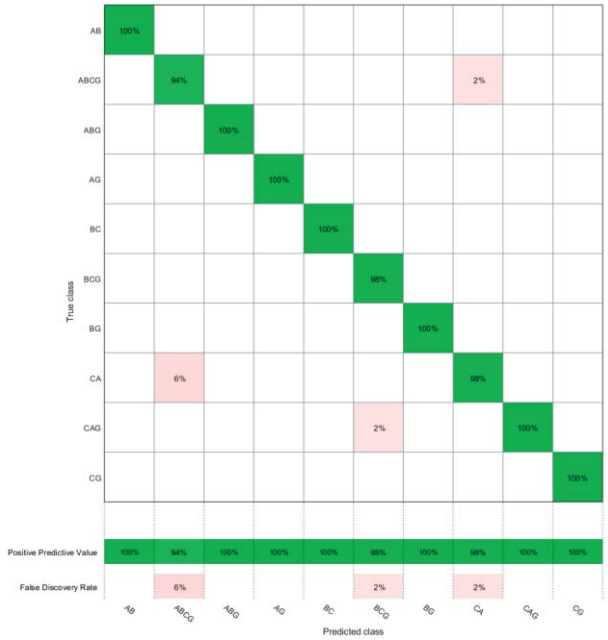


Fig. 9 confusion matrix of decision tree

VI. CONCLUSIONS

This paper uses machine learning models to complete the identification of fault types in the active distribution networks. The main conclusions of this work are as follows:

1) A method for batch generation of fault scenarios in PSCAD is proposed, which provides data preparation for machine learning. This method avoids manual modification of parameters and improves the efficiency of data preparation.

2) Fault features that can describe different fault types are selected. Variation multiples of voltage and current are extracted as the features to represent the fault type.

3) Various machine learning models are used to identify fault types, and the identification results of the decision tree are analyzed. Tree graph and confusion matrix are used to show the classification results of decision tree intuitively.

The test result shows that the method based on machine learning is able to classify faults under various fault conditions. It can be concluded that the proposed fault classification technique is simple and doesn't require a specific threshold value.

ACKNOWLEDGMENT

The paper is supported by National Key R&D Program of China (2017YFB0902800).

REFERENCES

[1] W. Lin, C. Yang, J. Lin, and M. Tsay, "A fault classification method by RBF neural network with OLS learning procedure," IEEE Transactions on Power Delivery, vol. 16, pp. 473-477, 2001.

[2] A. Ferrero, S. Sangiovanni and E. Zappitelli, "A fuzzy-set approach to fault-type identification in digital relaying," IEEE Transactions on Power Delivery, vol. 10, pp. 169-175, 1995.

[3] A. A. Girgis and M. B. Johns, "A hybrid expert system for faulted section identification, fault type classification and selection of fault location algorithms," IEEE Transactions on Power Delivery, vol. 4, pp. 978-985, 1989.

[4] F. S. Wen and C. S. Chang, "Probabilistic approach for fault-section estimation in power systems based on a refined genetic algorithm," IEE Proceedings-Generation, Transmission and Distribution, vol. 144, pp. 160-168, 1997.

[5] Y. Zhang, Y. Zhang, F. Wen, C. Y. Chung, C. Tseng, X. Zhang, F. Zeng, and Y. Yuan, "A fuzzy Petri net based approach for fault diagnosis in power systems considering temporal constraints," International Journal of Electrical Power & Energy Systems, vol. 78, pp. 215-224, 2016.

[6] M. Togami, N. Abe, T. Kitahashi, and H. Ogawa, "On the application of a machine learning technique to fault diagnosis of power distribution lines," IEEE transactions on power delivery, vol. 10, pp. 1927-1936, 1995.

[7] B. Das, "Fuzzy logic-based fault-type identification in unbalanced radial power distribution system," IEEE Transactions on Power Delivery, vol. 21, pp. 278-285, 2006.

[8] S. Hong-chun, S. Xiang-fei and S. I. Da-jun, "A study of fault diagnosis in distribution line based on rough set theory," PROCEEDINGS-CHINESE SOCIETY OF ELECTRICAL ENGINEERING, vol. 21, pp. 73-77, 2001.

[9] A. M. El-Zonkoly, "Fault diagnosis in distribution networks with distributed generation," Electric Power Systems Research, vol. 81, pp. 1482-1490, 2011.

[10] J. Korbicz, J. M. Koscielny, Z. Kowalczuk, and W. Cholewa, Fault diagnosis: models, artificial intelligence, applications: Springer Science & Business Media, 2012.

[11] M. H. Center, "PSCAD/EMTDC User's Manual," Manitoba HVDC Center, Winnipeg, Canada, 1998.

[12] J. Upendar, C. P. Gupta and G. K. Singh, "Statistical decision-tree based fault classification scheme for protection of power transmission lines," International Journal of Electrical Power & Energy Systems, vol. 36, pp. 1-12, 2012.