

Received August 8, 2020, accepted August 20, 2020, date of publication August 25, 2020, date of current version September 11, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3019377

Feature Selection Based on Random Forest for Partial Discharges Characteristic Set

RUI YAO¹, JUN LI², MENG HUI¹ , (Member, IEEE), LIN BAI¹, AND QISHENG WU¹

¹School of Electronic Control, Chang'an University, Xi'an 710064, China

²State Grid Shaanxi Electric Power Research Institute, Xi'an 710049, China

Corresponding author: Jun Li (18629635935@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 51407012 and Grant 61601059, in part by the Natural Science Foundation of Shaanxi Province under Grant 2020JM-256, in part by the Technology Innovation Leading Program of Shaanxi under Grant 2020QFY03-01, in part by the Fundamental Research Funds for the Central Universities, CHD, under Grant 300102329102 and Grant 300102329104, in part by the Natural Science Basis Research Plan in Shaanxi Province of China under Grant 2020GY-060, and in part by the Xi'an Science and Technology Plan Project under Grant 2020KJRC0126.

ABSTRACT Since the dimension of combined feature set for partial discharge (PD) pattern recognition is higher, the corresponding sample size increases, as does the required amount of storage space and calculation, and there are features with less category-related characteristics in the feature parameters, which may contain redundant information between them. To solve the problem of higher feature dimension and complicated classification model required for the identification of partial discharge insulation defect type in this paper. Random forest sequential forward selection method based on variance analysis (RF-VA) is proposed for the optimal subset selection. This method is improved in two aspects. Firstly, a method based on variance analysis is proposed, which measures feature differences between categories, and obtains a modified arrangement displacement scheme to guide rearrangement of the order of values taken on data sample out of bag. Secondly, the sequence forward search method used to do feature selection could get iteration evaluation results, which solves randomness to determine the size of feature subset and instability of the results existing in the original algorithm. The results show RF-VA can obtain a better subset of features. It is feasible to reduce the dimension of partial discharge characteristic set, and effectively improve the identification rate of partial discharge defect type.


INDEX TERMS Partial discharge, feature selection, variance analysis, random forest.

I. INTRODUCTION

Feature space or feature sets required for partial discharge (PD) pattern recognition is obtained through different feature extraction methods. Feature parameters of different feature sets are both different and complementary. Therefore, different feature sets are combined to obtain a combined feature set. Since the dimensions of combined feature set are higher, the corresponding sample size increases, as does the required amount of storage space and calculation, and there are features with less category-related characteristics in the feature parameters, which may contain redundant information between them. As a result, to save storage space, reduce calculation time, obtain essential characteristics, and reduce the difficulty of the classification model, optimal subset

selection of feature sets exerts a tremendous fascination on us [1].

Features that need to be removed are usually divided into two categories: one is redundant features, that is, features are duplicated; the other is irrelevant features. This subset transformation from d -dimensional features to d' -dimensional features is called feature selection, and finally these d' -dimensional features are used for model training. The common feature selection methods are roughly divided into three categories: Filter approach, Wrapper approach, and Embedding approach. These feature selection methods have been applied in various fields and achieved good recognition results. In [2], 9 representative feature parameters in horizontal and vertical directions extracted by two-dimensional principal component analysis based on partial discharge grayscale image decomposition were used. To get further improve recognition performance, feature selection

The associate editor coordinating the review of this manuscript and approving it for publication was Youqing Wang .

based on non-dominated genetic algorithm techniques is used to reduce feature dimensions. Reference [3] proposed a new method based on the random forest for partial discharge characteristic optimization for construction and optimization of partial discharge characteristics of high-voltage cables. In [4], to reduce the dimension of partial discharge identification parameters, the feature vector separability evaluation criteria were defined, and 9 sets of feature parameters with the best separability were selected using a floating forward search algorithm. Reference [5] uses an improved maximum correlation minimum redundancy algorithm to select the optimal feature subset of partial discharge.

Random Forest (RF) is a non-linear model that uses a decision tree as a base learner and uses bagging, also known as bootstrap aggregating, to process training data sets. At present, due to its low computational complexity, the randomness of sample, and feature selection, it has achieved good recognition results in various fields [6].

Random forest model calculates and evaluates the importance of each feature factor through the feature division process. For feature i , if the performance after replacing it with 'random' values of $x_{n,i}$ is worse than before, it indicates that feature i is more important and should take a larger weight and cannot be replaced with 'random' values. Conversely, if the performance of 'random' values replacement is not much different, it indicates that feature i is less important and optional. Therefore, by comparing the performance of a feature before and after it is replaced by 'random' values, the weight and importance of the feature can be inferred. The more common method for selecting random values in the random forest is to use a permutation test. That is when calculating the importance of the i -th feature, the original i -th feature distribution of all N samples are disrupted. Then compare the differences in the performance of features before and after permutation. If the difference is large, the i -th feature is important. How to measure the difference of feature performance before and after permutation? To simplify computational complexity, the permutation test on the original training set is moved to the feature vector of out-of-bag validation set [7].

Since the arrangement of features on out-of-bag set is random when using the out-of-bag data to measure the importance of feature to achieve feature rank in existing random forest feature selection algorithms. It cannot guarantee the feature with a strong correlation with class label get a good importance score. To solve this problem, this paper proposes a random forest Sequential Forward Selection method based on variance analysis, using for partial discharge feature optimization of GIS. Moreover, it is verified by experimental data. The results show this method can obtain a good feature subset and effectively improve recognition rate on partial discharge defect types identification.

II. EXPERIMENTAL SETUP

A. TEST SAMPLE

The UHF test system in this paper is performed on a section of an actual 220kV GIS in Xi'an XD Switchgear Electric Co.,



FIGURE 1. 220kV GIS test sample.

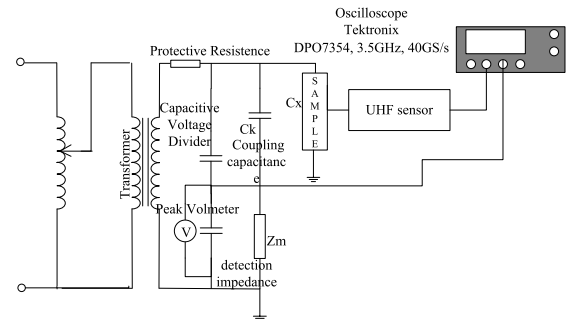


FIGURE 2. Measuring system.

Ltd. Experimental setup is shown in Figure 1. 220kV GIS sample is composed of a high-voltage conductor, metal enclosure, and two basin-type insulators. The total length of this test cavity is 2700mm, the outer diameter of the high-voltage conductor is 106mm, and the inner diameter of the enclosure is 320mm. UHF sensor is installed in a pre-designed mounting hole, and the physical model is installed in the designed hand hole. The basin-shaped spacer insulator material is epoxy resin, the relative dielectric constant is 4, and it is 520mm in outer diameter, 40mm in thickness. The GIS cavity is filled with SF₆ gas at 0.4Mpa.

B. MEASURING SYSTEM

As shown in Figure 2. The measurement system is composed of a voltage regulator (Rated Capacity 200kVA, Frequency 50Hz, Rated Voltage 10kV/1kV, Rated Current 20A/200A), a transformer, and a capacitor voltage divider (Voltage Division Ratio 70kV/3.5V). The applied voltage is provided by a corona-free power frequency high voltage test transformer (rated voltage 750kV, rated capacity 375kVA). The test voltage is adjustable from 0 to 750kV. C_x represents the physical model of partial discharge. The reference voltage signal is obtained from the peak voltmeter of the low-voltage arm of the capacitive voltage divider. When measuring the UHF signal, an oscilloscope (oscilloscope Agilent, DSO9404A, 4GHz bandwidth across all 4 analog channels, 20GS/s max. sample rate) is used.

C. PHYSICAL MODELS

(1) Suspending electrode. As shown in Figure 3 (a), this type of defect is achieved by placing a gasket with a diameter of

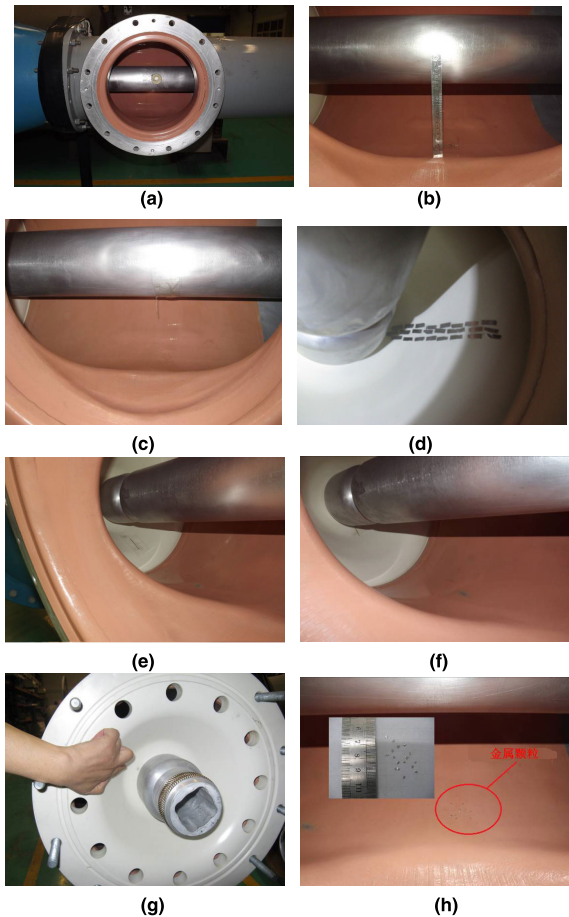


FIGURE 3. Physical models (before the test) (a) Suspending electrode (b) A metal protrusion on the enclosure (c) A metal protrusion on the high-voltage conductor (d) Metal particles contamination on the spacer (e) Single wire contamination on the spacer I (f) Single wire contamination on the spacer II (g) A void in the insulator (h) Free metal particles on the enclosure.

33mm on a high-voltage conductor wound with an insulation tape in this paper. Gasket and high-voltage conductors are separated by an insulating tape and are not in direct contact.

(2) A metal protrusion on the enclosure. As shown in Figure 3 (b), single metal wire with a length of 17mm is fixed on the enclosure to simulate corona discharge caused by metal protrusions on the GIS enclosure.

(3) A metal protrusion on the high-voltage conductor. As shown in Figure 3 (c), a metal protrusion with a length of 21 mm is fixed on the high-voltage conductor to simulate corona discharge.

(4) Metal particles contamination on the spacer. As shown in Figure 3 (d), the aluminum foil is cut into rectangular pieces of 16mm in length and 4mm in width. And these pieces are arranged radially according to the insulator. It is a total of 3 rows, 8 pieces in each row, of which every two pieces of strip-shaped tin foil are spaced 2 to 4 mm in the radial direction and 3 mm in the lateral direction.

(5) Single wire contamination on the spacer I. As shown in Figure 3 (e), a 50mm long wire was fixed on the surface

of spacer to simulate surface discharge. Its top end is 40mm away from the high-voltage conductor, and its end is 60mm away from the GIS enclosure.

(6) Single wire contamination on the spacer II. As shown in Figure 3 (f). A 53mm long metal wire is fixed on the surface of the insulation spacer to simulate surface discharge. The metal wire is as close as possible to the high-voltage conductor but does not contact the conductor. The other end of the metal wire is directed to the cylinder wall in a radial direction.

(7) A void in the insulator. The defect manufacturing process is as follows: firstly punching a hole with a diameter of 3mm and a depth of 15mm on the spacer. The hole is at a distance of 10mm away from the high-voltage conductor. Then the same epoxy resin as the insulator is used to seal the hole, only one of the bubbles remaining in the hole bottom after coagulation. The physical model of this type of defect is shown in Figure 3 (g).

(8) Free metal particles on the enclosure. As shown in Figure 3 (h), free metal particles with a width of 0.5mm ~ 1mm and a length of 1mm ~ 2mm are placed on the enclosure of the GIS cylinder, and they are fixed in a transparent glass cover to prevent spreading everywhere when GIS is inflated.

III. FEATURE EXTRACTION BASED ON MULTIPLE PD PATTERNS

To identify the PD insulation defect, it is indispensable to extract distinguishing feature parameters based on the obtained PD patterns. Feature parameters as input to the classification model are the prerequisite and basis for ensuring successful identification of defects. So feature parameters are respectively obtained for partial discharge insulation based on the phase resolved pulse sequence analysis (PRPSA) pattern, partial discharge (PRPD) pattern, and the polar coordinate phase resolved analysis (PCPRA) pattern in this paper. It laid the foundation for the identification of partial discharge insulation defects.

A. PRPS DATA ACQUISITION

First, the initial discharge voltage and breakdown voltage of the defect can be determined by applying the voltage step by step. Then the data of the defect from initial discharge to discharge breakdown can be obtained in a short time through increasing voltage method after several tests. As shown in Table 1, 25 samples are obtained at the same test voltage level for each defect and a total of 150 samples for 6 voltage levels through multiple experiments.

The partial discharge data recording format is a partial discharge pulse sequence $q_s(t_s, u(t_s))$. It is a PD activity of M PD pulses, $\forall s = 1, \dots, M$ over a measuring time t_m with discharge amplitude of PD pulses q_s , time of PD pulses occurrence t_s , and the applied test voltage $u(t_s)$ [8], [9]. Such data is called Phase Resolved Pulse Sequence (PRPS) data. Various partial discharge patterns for each defect can be obtained from PRPS data.

TABLE 1. Test voltage levels and sample number for each defect.

Label	Defect type	Voltage level	Sample number
1	Suspending electrode	296,320,350,370,390,400	150
2	A metal protrusion on enclosure	210,252,271,332,370,383	150
3	A metal protrusion on high-voltage conductor	73,90,106,140,183,204	150
4	Metal particles contamination on spacer	20,23,35,72,90,120	150
5	Single wire contamination on spacer I	80,110,149,171,198,240	150
6	Single wire contamination on spacer II	46,53,58,61,64,69	150
7	A void in insulator	247,280,320,340,360,385	150
8	Free metal particles on enclosure	502,510,520,540,560,580	150

B. PD PATTERNS ACQUISITION

After obtaining PRPS data, various PD patterns should be get to carry out effective pattern recognition of partial discharge.

In this paper, three kinds of partial discharge patterns are drawn. It takes an insulation defect of metal particles contamination on spacer as an example at 120kV test voltage level, which is shown in Figure 4.

Firstly, the phase resolved pulse sequence analysis (PRPSA) pattern is obtained based on the PRPS data [10], [11]. As shown in Figure 4 (a). This pattern contains complete information of discharge pulse. This paper obtains the PRPSA pattern through plotting the first 10 power frequency cycles of each sample together.

Secondly, the phase resolved partial discharge (PRPD) pattern is obtained based on the PRPS data, which is shown in Figure 4 (b). This pattern is also called $\varphi-q-n$ pattern, which is obtained according to the number of partial discharge pulses N , discharge phase φ , and amplitude of discharge pulse q . The PRPD pattern is obtained according to 100 power frequency cycles of the PRPS data sample.

Finally, the polar coordinate phase resolved analysis (PCPRA) pattern is obtained [12], as shown in Figure 4 (c). By converting the angle φ of discharge occurrence into radians φ_{norm} . PD data are plotted in the polar coordinate, taking the normalized discharge amplitude q_{norm} as polar radius and φ_{norm} as polar angle, using a 50Hz power frequency signal as a reference signal. This pattern is obtained through plotting 100 power frequency cycles of the PRPS data sample.

C. FEATURE EXTRACTION BASED ON PD PATTERNS

The 15 basic feature parameters extracted in this paper based on PRPSA pattern include: discharge start phase φ_s , center of gravity of discharge phase (φ^+ , φ^-), discharge width (Phw^+ , Phw^-), average value of the time interval $E(\Delta t)$, standard deviation of the time interval $S(\Delta t)$, the quantity of

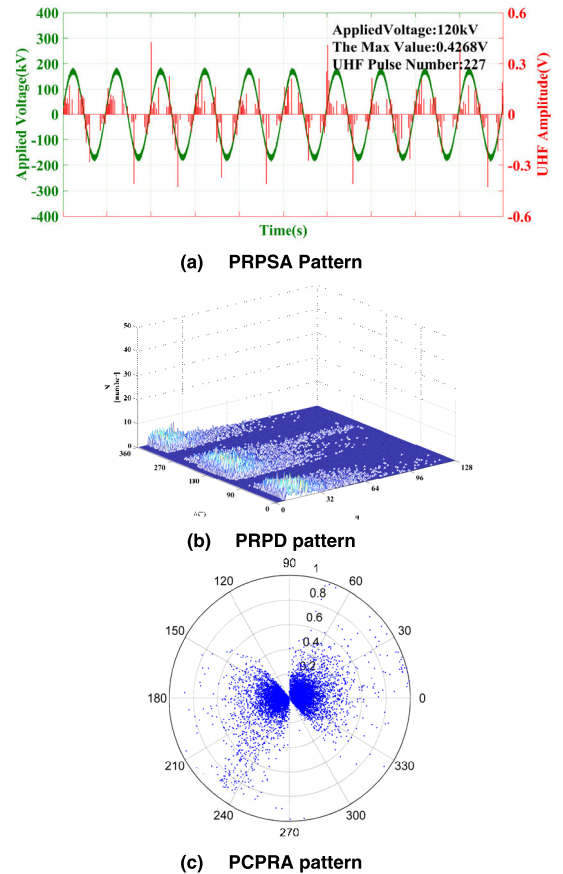


FIGURE 4. Three discharge patterns of metal particle contamination defect on spacer with the applied voltage at 120kV.

the time interval between two consecutive discharge pulses $N(\Delta t)$, the center of gravity of discharge amplitude (q^+ , q^-), the average value of discharge amplitude Eq , the standard deviation of discharge amplitude Sq , the number of discharge pulses Nq , the maximum ratio of adjacent partial discharge pulse amplitudes Rat_{max} , and the square ratio of discharge amplitude D , which constituted the input feature vector F_1 (1200×15) to identify the PD types

Maximum pulse height distribution $H_{q_{max}}(\varphi)$, mean pulse height distribution $H_{qn}(\varphi)$, pulse count distribution $H_n(\varphi)$, and discharge power-phase distribution $H_p(\varphi)$ can be obtained from the PRPD pattern. These four distributions are divided into two types for the positive and negative half cycles of test voltage. Besides the distribution of discharge amplitude $H(q)$ can also be obtained. Then the selected statistical operators are calculated to describe the shape features of these distributions for positive and negative half cycles of the test voltage [13]–[16]. 33 statistical operators contained skewness (S_k), kurtosis (K_u), number of peaks (P_e), discharge asymmetry Q , cross-correlation factor (cc), and modified correlation factor (mcc), which constituted the input feature vector F_2 (1200×33) to identify the PD types

The steps of extracting feature parameters based on the PCPRA pattern are as follows: First, the PCPRA spectrum is classified into two cluster discharge data sets using

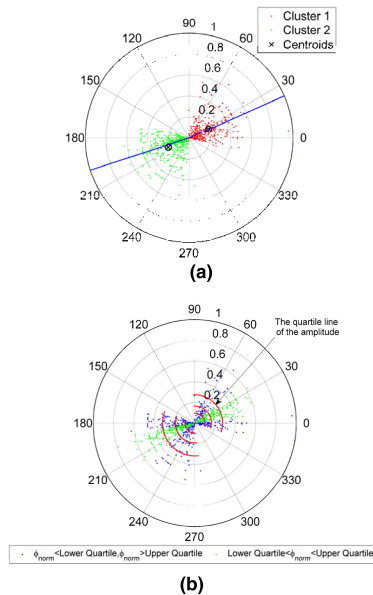


FIGURE 5. Characteristic parameters examples of PCPRA pattern (a) Cluster results and the phase median line in PCPRA pattern (b) Phase quartile and amplitude quartile in PCPRA pattern.

the K-means clustering algorithm. Then 10 basic discharge parameters of each cluster are calculated. That is centroid (c_φ, c_q), discharge width W_p , number of discharges N_c , median phase Q_{p2} , median amplitude Q_{a2} , phase quartile Q_{p1}, Q_{p3} , and amplitude quartile Q_{a1}, Q_{a3} of each discharge cluster. A total of 20 feature parameters are achieved.

Centroid describes the average discharge level of each discharge cluster, as shown in Figure 5 (a). The discharge width is calculated based on the maximum discharge angle and the minimum discharge angle of each discharge cluster. The number of discharges is the number of points for each discharge cluster. Phase and amplitude are described by quartile. Lower quartile (0.25 quartile), upper quartile (0.75 quartile) and middle (0.5 quartile) are calculated to evaluate the dispersion of PD data set, and analyze PD characteristics, as shown in Figure 5 (b).

Finally, quadrant-based parameters $D_i(i = 1, 2, 3, 4)$, cosine similarity of the centroid vector $Cossim(c_1, c_2)$, ratios of amplitude quartiles A_{ratio1} and A_{ratio2} , combined with parameters based on statistical theory, constituted the input feature vector $\mathbf{F}_3 (1200 \times 34)$ to identify the PD types [12].

IV. RF-VA

A. RANDOM FOREST

Random Forest (RF) uses a decision tree as a base learner and uses bagging, also known as bootstrap aggregating, to process training data sets [17]. Specifically, randomly selecting N samples from the original training set and putting them back. That is for this N sample, N times random samplings are performed, and one is selected from the samples in each sampling, and then “copy” out. The sample set is still N when the next sampling. Because the sampling process is

replaced, some samples may be selected multiple times and appear multiple times in the same training set, while others may not be selected at once. The ignored samples are called “Out of Bag (OOB)”.

In this paper, RF is used for feature selection. This feature selection method is performed based on the ranking result of the feature set by adopting the importance measure. The classification and regression tree (CART) algorithm is used to construct the decision tree [18].

There are two ways to measure the importance of features in RF [19]: One is to use the Gini Index as the partition function and calculate the “Gini Importance” of the feature to indicate its importance. Gini value is a concept similar to entropy, defined as:

$$Gini(D) = 1 - \sum_{i=1}^{|C|} [p_i]^2 \tag{1}$$

where D is the sample set, p_i is the probability belonging to category i in D , and C is the category set.

Gini index of D is defined under the condition of a known feature A :

$$Giniindex(D, A) = \sum_{v=1}^V \frac{|D^v|}{|D|} Gini(D^v) \tag{2}$$

where V is the number of feature A . D is divided into V subsets $\{D^1, D^2, \dots, D^V\}$ according to feature A values, and the samples in each subset have the same value on A . When selecting properties, the property that minimizes the Gini index after partitioning is selected as the optimal partitioning property. When all the samples in a node belongs to the same class, Gini index is zero.

The other method to measure the importance of feature vectors is using OOB data observations. The samples that are not selected after bootstrap aggregating is called “Out of Bag (OOB)”. For the sample (x_n, y_n) , the size of the OOB sample for the decision tree is about $(1 - 1/N)^N$. In general, if N is large enough, this probability converges to 0.368. OOB has the characteristics of a validation set, so OOB error is used as a generalization error for validating random forests. As shown below:

$$E_{\text{OOB}}(G) = \frac{1}{N} \sum_{n=1}^N \text{err}(y_n, G_n^-(x_n)) \tag{3}$$

where $G_n^-(x_n)$ refers to the decision tree that OOB only concludes x_n .

The method of using the out-of-bag data to measure the importance of feature vectors is to reorder the features on the OOB samples, and calculate the “Permutation Importance” of the features to achieve feature ranking. “Permutation Importance” is called the importance of random arrangement. It is defined according to the statistical “Permutation test”. This test replaces $\{x_{n,j}\}$ in the original OOB sample with random rearrangement $\{x_{n,j}\}_{n=1}^N$ to form a new OOB sample. Recorded $VI^{(t)}(x_j)$ as the importance of the characteristic variable x_j obtained by the decision tree g_t , then there

are:

$$VI^{(t)}(\mathbf{x}_j) = E_{\text{ooB}}(G) - E_{\text{ooB}}^{(P)}(G) \quad (4)$$

where $E_{\text{ooB}}^{(P)}(G)$ is the error of the recorded random rearrangement OOB samples.

The final importance score for each feature parameter is the average of all trees:

$$VI(\mathbf{x}_j) = \frac{\sum_{t=1}^{\text{ntree}} VI^{(t)}(\mathbf{x}_j)}{\text{ntree}} \quad (5)$$

B. RANDOM FOREST SEQUENTIAL FORWARD SELECTION ALGORITHM BASED ON ANALYSIS OF VARIANCE (RF-VA)

The importance of random permutation ‘‘Permutation Importance’’ is based on each decision tree in the existing random forest algorithm. In order to calculate the importance measurement result of each decision tree for a feature parameter, it is necessary to ensure that the value distribution of the OOB samples of the decision tree on the feature is rearranged without changing the other features. By randomly arranging the feature variable \mathbf{x}_j in the OOB sample, its original association with the label \mathbf{y} is destroyed. The ranked variables \mathbf{x}_j and the remaining unranked variables are used to estimate the error of OOB samples. For feature \mathbf{x}_j , the decision tree predicts the original OOB samples and the rearranged OOB samples respectively. The difference between the two prediction errors is the measure of the importance of feature \mathbf{x}_j by the decision tree. The final score of the feature variable is the result of all decision trees working together.

If the original variable \mathbf{x}_j is not related to \mathbf{y} , the error rate on the new OOB sample will not change, theoretically $VI(\mathbf{x}_j) = 0$. If the original variable \mathbf{x}_j is associated with \mathbf{y} , and \mathbf{x}_j is a good feature with discrimination, the discrimination decreases and the OOB error increases after the rearrangement, then $VI(\mathbf{x}_j) < 0$. If \mathbf{x}_j is a bad feature, the discrimination may increase, the OOB error may decrease after the rearrangement, then $VI(\mathbf{x}_j) > 0$. Therefore, the rearrangement method determines the OOB error. If the random arrangement method in the existing random forest feature selection algorithm is adopted, for good features with discrimination, the discrimination may not be increased. It is a problem of that how to arrange or replace to ensure that the importance score of good features (strong correlation with class labels) takes a higher value, and the importance score of corresponding bad features (weak correlation with class labels) takes a lower value.

In order to solve this problem, this paper proposes a method based on analysis of variance to measure the difference of features in different categories, and a modified permutation scheme is obtained, which is used to guide the rearrangement of a feature order on the OOB sample.

Analysis of variance (ANOVA), also known as ‘‘variance analysis’’ or ‘‘F test’’, was invented by R.A. Fisher and used for the significance test of the difference between two or more samples [20]. The difference of feature \mathbf{x}_j in different

categories is measured by analysis of variance:

$$F(\mathbf{x}_j) = \left[\sum_k n_k \left(E(\mathbf{x}_j^k) - E(\mathbf{x}_j) \right) / (K-1) \right] / \sigma^2 \quad (6)$$

$$\sigma^2 = \left[\sum_k (n_k - 1) \sigma_k^2 \right] / (n - K) \quad (7)$$

$E(\mathbf{x}_j)$ represents the mean value of OOB samples on feature \mathbf{x}_j , $E(\mathbf{x}_j^k)$ represents the mean value of OOB samples of category k on feature \mathbf{x}_j , σ^2 represents the combined variance of OOB samples in each category, K is the number of categories of OOB samples, n_k is the OOB samples number of category k , n is the total number of OOB samples.

For distinguishing features, there are significant differences in different categories, and the order of values cannot be ‘‘swapped’’. On the contrary, for irrelevant or redundant features, the order of values can be ‘‘swapped’’. Based on this idea, a modified permutation scheme can be obtained to check the distinguishing ability of features.

The guide to the replacement method of a certain feature on the OOB sample is made through comparing the value of $F(\mathbf{x}_j)$ with threshold values $F_\alpha(K-1, n-K)$ (checked through the F distribution table) at the given significance level α ($\alpha = 0.05$). The larger the $F(\mathbf{x}_j)$ value ($F(\mathbf{x}_j) > F_\alpha$), the greater the feature difference of different samples in different categories, and the smaller the feature difference of different samples in the same category, the more important the feature is, then feature \mathbf{x}_j is rearranged on OOB samples using inter-class replacement method. If the $F(\mathbf{x}_j)$ value is small ($F(\mathbf{x}_j) < F_\alpha$), the feature \mathbf{x}_j is randomly arranged on the OOB sample.

Figure 7 is a schematic diagram of OOB samples for decision tree g_t . Each sample is represented as a pair (\mathbf{x}_i, y_i) ($i = 1, 2, \dots, N$), where $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{im})^T$, corresponds to the attribute set of the i -th sample. $y_i \in \{1, 8\}$ presents its class label. The light shaded part in Figure 7 corresponds to the OOB sample of the decision tree g_t is (\mathbf{x}_i, y_i) , $i = 2, 3, 210, 211, 239, \dots$. For the feature \mathbf{x}_j , if the $F(\mathbf{x}_j)$ value is large, then $x_{2j}, x_{3j}, x_{210j}, x_{211j}, x_{239j}$ using the inter-class substitution method to obtain a new OOB sample, that is, take any two values replacement of x_{2j}, x_{3j} and $x_{210j}, x_{211j}, x_{239j}$. If the $F(\mathbf{x}_j)$ value is small, the arrangement scheme is: $x_{2j}, x_{3j}, x_{210j}, x_{211j}, x_{239j}$ can be arbitrarily replaced. In this way, a new OOB sample is obtained. Then calculate the error of the rearranged OOB samples, and finally the importance of the feature \mathbf{x}_j for the decision tree g_t is obtained.

After the features in descending order of importance is obtained by the existing random forest algorithm, the feature subset can be selected according to the preset feature dimension to be retained. Generally, there are the following ways:

- 1) Retaining features with importance scores greater than 0.
- 2) Retaining the top k features.
- 3) Retaining the top 10% of features.

$(\mathbf{x}_1, \mathcal{Y}_1)$	x_{11}	x_{12}	...	x_{1j}	...	x_{1m}	\mathcal{Y}_1	Category 1
$(\mathbf{x}_2, \mathcal{Y}_2)$	x_{21}	x_{22}	...	x_{2j}	...	x_{2m}	\mathcal{Y}_2	
$(\mathbf{x}_3, \mathcal{Y}_3)$	x_{31}	x_{32}	...	x_{3j}	...	x_{3m}	\mathcal{Y}_3	
...	
$(\mathbf{x}_{120}, \mathcal{Y}_{120})$	x_{1201}	x_{1202}	...	x_{120j}	...	x_{120m}	\mathcal{Y}_{120}	Category 2
$(\mathbf{x}_{121}, \mathcal{Y}_{121})$	x_{1211}	x_{1212}	...	x_{121j}	...	x_{121m}	\mathcal{Y}_{121}	
...	
$(\mathbf{x}_{210}, \mathcal{Y}_{210})$	x_{2101}	x_{2102}	...	x_{210j}	...	x_{210m}	\mathcal{Y}_{210}	
$(\mathbf{x}_{211}, \mathcal{Y}_{211})$	x_{2111}	x_{2112}	...	x_{211j}	...	x_{211m}	\mathcal{Y}_{211}	
...	
$(\mathbf{x}_{239}, \mathcal{Y}_{239})$	x_{2391}	x_{2392}	...	x_{239j}	...	x_{239m}	\mathcal{Y}_{239}	
$(\mathbf{x}_{240}, \mathcal{Y}_{240})$	x_{2401}	x_{2402}	...	x_{240j}	...	x_{240m}	\mathcal{Y}_{240}	
...	

FIGURE 6. OOB sample diagram for the decision tree g_t .

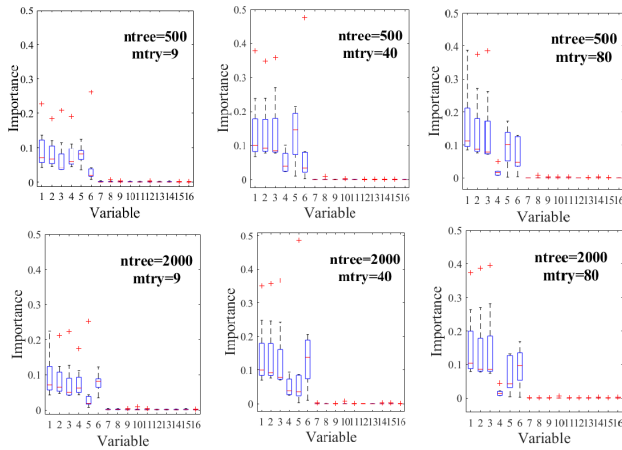


FIGURE 7. Variable importance sensitivity to m_{try} and $ntree$.

These methods are greatly influenced by human factors and features cannot be selected accurately and reliably. In this paper, Sequential Forward Selection (SFS) is used for feature selection to obtain iterative feature evaluation results, which are used to solve the arbitrariness of existing algorithms in determining the size of feature subsets and the instability of results. That is, according to the result of descending order of feature importance, the next feature is added from the first feature, and then each feature vector combination is input to the classifier, and the classification error probability is calculated until all features are used. The feature vector combination with the smallest error probability is selected as the final feature selection result. In this paper, the barrier feature

grouping strategy is used to divide the original feature data set into a training set and a test set. Every 4 samples are taken as the test set samples, and the remaining part is the training set samples, that is, the training set and test set sample ratio is 4:1. A 10-fold algorithm is used for cross-validation. The data in the training set is divided into 10 groups, 9 of which are used to train the model, and the remaining one is used to verify the model. The test data is replaced each time until each group of data is tested. The process of cross-validation is repeated 10 times in sequence, and the average value of the obtained results is used as the final classification result. Support vector machine (SVM) with Gaussian kernel used in this paper to calculate the classification error probability [21], [22].

From the above, the random forest sequential forward selection based on analysis of variance (RF-VA) is proposed to reduce feature dimension in this paper. The original sample number is N and the feature dimension is m . The process of RF-VA is as follows:

Input: a set of training samples $\{(\mathbf{x}_1, \mathcal{Y}_1), \dots, (\mathbf{x}_N, \mathcal{Y}_N)\}$

Step 1: Initialization parameters $t = 1, ntree = 500$, significance level $\alpha = 0.05$, maximum classification correct rate $Acc_{max} = 0$.

Step 2: Generate a bootstrap aggregation sample set D_t of size N .

Step 3: Use the algorithm of classification and regression tree(CART) on the bootstrap aggregation sample set D_t .

Step 4: Sampling columns (features) for each tree, randomly selecting $m_{try} = \sqrt{m}$ features, and selecting a feature with the highest variable importance in m_{try} to perform node splitting.

Step 5: Calculate the error $E_{oob}(G)$ on the original OOB sample of the decision tree g_t .

$$E_{oob}(G) = \frac{1}{N} \sum_{n=1}^N \text{err}(y_n, G_n^-(x_n)) \quad (8)$$

Step 6: Measure the difference between the different categories by calculating the analysis of variance $F(x_j)$ of the feature x_j on the OOB sample.

$$F(x_j) = \left[\sum_k n_k (E(x_j^k) - E(x_j)) / (K - 1) \right] / \sigma^2 \quad (9)$$

Step 7: Determine whether $F(x_j)$ satisfies $F(x_j) > F_\alpha(K - 1, n - K)$. If it meets the requirement, the feature x_j will be replaced by inter-class replacement method on the OOB sample. If it is not satisfied, the feature x_j will be randomly arranged on the OOB sample. That is, the OOB samples of the original feature $\{x_{n,j}\}$ is replaced with randomly rearranged $\{x_{n,j}\}_{n=1}^N$ on the OOB samples to form a new OOB sample.

Step 8: Calculate the error $E_{oob}^{(P)}(G)$ of the decision tree g_t in rearranging OOB samples.

Step 9: Obtain the importance of feature variables x_j for decision tree g_t .

$$VI^{(t)}(x_j) = E_{oob}(G) - E_{oob}^{(P)}(G) \quad (10)$$

Step 10: Determine whether the tree number meets $t \leq ntree$. Repeat steps 2-9 if it is satisfied, end the loop if it is not satisfied.

Step 11: Calculate the importance score of each feature:

$$VI(x_j) = \frac{\sum_{t=1}^{ntree} VI^{(t)}(x_j)}{ntree} \quad (11)$$

Step 12: Sort the features by importance to get **FeaSort**.

Step 13: Take the first feature from **FeaSort** to obtain **Fnew**, and obtain the classification correct rate Acc by using the support vector machine of Gaussian kernel function.

Step 14: Determine whether the current Acc satisfies $Acc_{max} \leq Acc$. If it is satisfied, then **FinalFea** = **Fnew**.

Step 15: Determine whether the current loop variable is less than or equal to the feature dimension m . If it is satisfied, add the feature of the next rank in **FeaSort** to **Fnew**, and repeat steps 13-14. If it is not satisfied, stop the loop.

Output: The highest classification accuracy rate and corresponding feature set **FeaSort** on the test set.

C. ALGORITHM ANALYSIS

The feature selection algorithm proposed in this paper is based on the random forest algorithm. First, the feature vectors are sorted by measuring the importance of feature vectors. After the features in descending order of importance are obtained by the random forest algorithm, Sequential Forward Selection (SFS) is used to generate candidate feature subsets, which are used to solve the arbitrariness of existing algorithms in determining the size of feature subsets and the instability of results. Then, the support vector machine is used as the classifier. The classification accuracy rate of the classifier is used as the feature subset evaluation criterion.

In the process of screening features, the classifier is trained based on the importance metric and the sequence forward selection method to form a feature subset, and the pros and cons of the feature set are evaluated based on the performance of the classifier on the test set.

This method is a wrapper feature selection strategy. It is slower in speed than the filtering method, but the optimized feature subset selected by it is relatively small, which is very conducive to the identification of key features. At the same time, its accuracy is relatively high, but its generalization ability is poor and the time complexity is high. Therefore, the algorithm proposed in this paper is suitable for occasions that require high accuracy (MSE for regression and classification rate for classification).

In the random forest feature selection method proposed in this paper, the base classifier selects the CART algorithm. Assuming that the feature dimension of the training data set is m and the number of training samples is n , the time complexity of the CART algorithm is $O(mn(\log n)^2)$. In the process of constructing the CART tree, the random forest randomly selects features from the m features to calculate the Gini Index and does not prun the tree growth. Therefore, the calculation time for training each base classifier is less than $O(mn(\log n)^2)$. Assuming that the number of base classifiers

in a random forest is k , the time complexity of the random forest algorithm can be approximated as $O(kmn(\log n)^2)$. Compared with the existing RF feature selection method, the time complexity of measuring the difference between the different categories by calculating the analysis of variance of the feature on the OOB sample of RF-VA is $O(n)$. So the time complexity of the RF-VA algorithm can be approximated as $O(kmn(\log n)^2)$. Besides, the Sequential Forward Selection strategy for feature selection requires a cycle of $m - 1$ times. In each cycle, 10-fold cross-validation is used, and the random forest algorithm needs to be run 10 times. The total time complexity of the algorithm can be approximately expressed as:

$$\begin{aligned} O\left((m-1) * \left(10 * O\left(kmn(\log n)^2\right)\right)\right) \\ \approx O\left(km^2n(\log n)^2\right) \end{aligned} \quad (12)$$

It can be seen from equation (12) that the time complexity of the RF-VA algorithm has an approximate square relationship with the feature dimension m , and an approximate cubic relationship with the number of samples in the data set.

D. SENSITIVITY TO m_{try} AND $ntree$

The parameters choice of m_{try} and $ntree$ is important for calculating the Variable Importance. Figure 7 illustrates the behavior of variable importance for several values of m_{try} and $ntree$. Boxplots are based on the RF-VA algorithm, variable importance for a few variables are plotted. The first plot on each row is taken as the reference for observing the relative importance of the initial variables.

In the experiment, $N = 960$, $m = 82$. Three values of m_{try} (9 the default, 40 and 80) and two values of $ntree$ (500 the default, and 2000) are setting separately and variable importance obtained with the setting is plotted.

The effect of taking a larger value for m_{try} is obvious. Indeed the magnitude of Variable Impaortance is more than 1.5 times starting from $m_{try} = 9$ to $m_{try} = 40$, and it remains stable from $m_{try} = 40$ to $m_{try} = 80$. The effect of $ntree$ is less visible, but taking $ntree = 2000$ leads to better stability.

V. EXPERIMENTAL RESULTS AND ANALYSIS

A. EXPERIMENTAL RESULTS

For the single feature sets F_1 (1200×15), F_2 (1200×33), and F_3 (1200×34), the existing random forest and RF-VA proposed in this paper are used to do feature selection. To obtain the relationship between classification accuracy and the number of feature vectors, the Sequential Forward Selection method is used for feature selection after obtaining the feature importance ranking through the existing RF.

The results are shown in Figures 8, 9, and 10. It can be seen from the figure that for the existing RF feature selection method, when the classification accuracy reaches the highest values of 84.58%, 75%, and 84.58%, the corresponding feature dimensions are 12, 21, and 14, respectively. For the RF-VA feature selection method, when the classification accuracy reaches the highest values of 85.00%, 76.25%, and

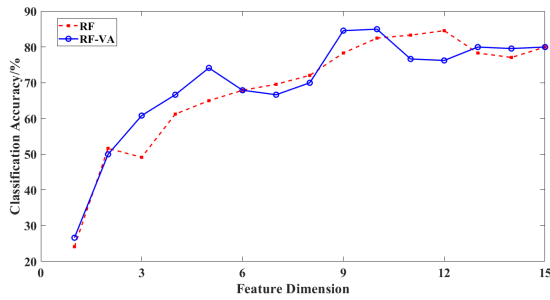


FIGURE 8. Relationship between classification accuracy and feature dimension for F_1 (1200×15).

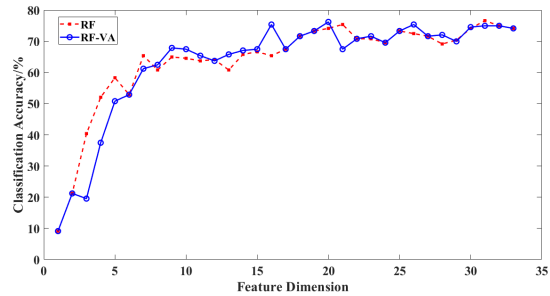


FIGURE 9. Relationship between classification accuracy and feature dimension for F_2 (1200×33).

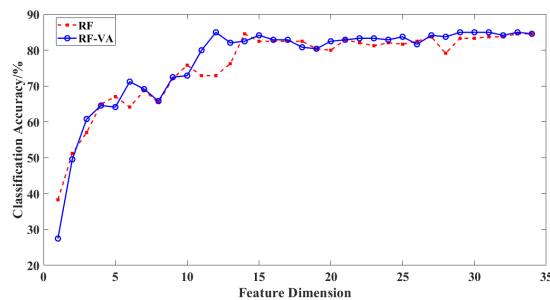


FIGURE 10. Relationship between classification accuracy and feature dimension for F_3 (1200×34).

85.00%, the corresponding feature dimensions are 10, 20, and 12, respectively. It can be seen that for the single feature set F_1 (1200×15), F_2 (1200×33), and F_3 (1200×34), the corresponding feature dimension to high classification accuracy get by the RF-VA feature selection method proposed in this paper is lower than the existing RF feature selection method. This shows that better classification performance and lower-dimensional feature subsets can be obtained by the RF-VA method. The permutation and the replacement scheme based on the analysis of variance is superior to the completely random method.

Inputting the dimensionality reduction results of the single feature set using the existing RF and the RF-VA feature selection method to the classifier separately. As shown in Table 2, The test results show that the RF-VA used to reduce the dimensionality of the single feature set of partial discharge can obtain a higher classification accuracy rate.

TABLE 2. Dimension reduction results and corresponding classification accuracy of the single feature set.

Feature Dimension Reduction Method	Original feature set	F_1 (1200×15)	F_2 (1200×33)	F_3 (1200×34)
RF	Feature Dimension	12	21	14
	Classification Accuracy /%	84.58	75.00	84.58
RF-VA	Feature Dimension	10	20	12
	Classification Accuracy /%	85.00	76.25	85.00

TABLE 3. Dimension reduction results and corresponding classification accuracy of the combined feature set.

Method	Original feature set	(F_1, F_2) (1200×48)	(F_1, F_3) (1200×49)	(F_2, F_3) (1200×67)	(F_1, F_2, F_3) (1200×82)
RF	Feature Dimension	17	18	14	20
	classification accuracy /%	85.00	85.00	84.17	88.33
RF-VA	Feature Dimension	13	13	11	15
	classification accuracy /%	85.42	86.25	84.17	88.33

As shown in Table 3, the three single feature sets combined in pairs or together to obtain the combined feature sets (F_1, F_2) , (F_1, F_3) , (F_2, F_3) and (F_1, F_2, F_3) . For the combined feature sets, the existing RF and RF-VA feature selection methods is used for dimension reduction. For existing RF feature selection methods, when the classification accuracy reaches the highest values of 85.00%, 85.00%, 84.17%, and 88.33%, the corresponding feature dimensions are 17, 18, 14, and 20, respectively. For the RF-VA, when the classification accuracy reaches the highest values of 85.42%, 86.25%, 84.17%, and 88.33%, the corresponding feature dimensions are 13, 13, 11 and 15, respectively. Under the condition that the classification accuracy rate remains high, the corresponding feature dimension obtained by RF-VA is lower than the RF feature selection method.

B. EXPERIMENTAL RESULTS ANALYSIS

From the experimental results, for the single feature sets F_1 , F_2 , and F_3 , the corresponding feature dimension to high classification accuracy get by the RF-VA feature selection method proposed is reduced by 16%, 5% and 14%. And for the combined feature sets (F_1, F_2) , (F_1, F_3) , (F_2, F_3) and (F_1, F_2, F_3) , the feature dimension is reduced more than 20%. This shows that, the effect of RF-VA is better than that of RF for high-dimensional feature variables, which expresses that the permutation and replacement scheme based on analysis of variance is superior to the completely random method.

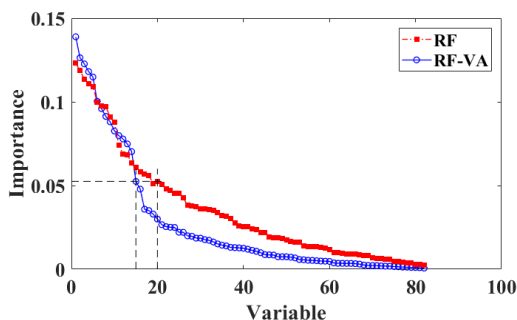


FIGURE 11. Variable importance of the combined feature set (F_1, F_2, F_3) by the two feature selection method.

When the RF-VA feature selection algorithm uses out-of-bag data observations to measure the importance of feature vectors, the analysis of variance is used to measure the differences of features in different categories. The algorithm thought of RF-VA is that, for distinguishing features, there are significant differences in different categories, and the order of values cannot be “swapped”. On the contrary, for irrelevant or redundant features, the order of values can be “swapped”. Based on this idea, a modified permutation scheme can be obtained to check the distinguishing ability of features.

Based on this idea, a modified permutation scheme can be obtained to test the distinguishing ability of features. From the variable importance of the combined feature set (F_1, F_2, F_3) accessed by the two feature selection method RF and RF-VA in Figure 11, it shows that analysis of variance makes the importance score of the feature with a strong correlation with the class label obtains a higher value, and the importance score of the corresponding feature with weak correlation with the class label obtains a lower value. The feature dimension is reduced by 25% at the intersection of the dotted line and the curve in the figure.

VI. CONCLUSION

In the assessment of the insulation status of partial discharges, the amount of original partial discharge data collected through ultra-high frequency methods is quite large. It will be difficult to directly identify the type of discharge and evaluate its severity. To effectively realize partial discharge recognition, the original data must be transformed to obtain various discharge patterns. Each type of partial discharge pattern has its characteristics. Therefore, the study of feature set generation methods suitable for different partial discharge patterns, can not only make full use of the discharge data information, and obtain multiple types of feature sets from different aspects to characterize the partial discharge, which also provides a guarantee to obtain discriminative feature sets.

Aiming at the problem of feature dimensionality reduction method in insulation defect pattern recognition of partial discharge, a random forest Sequential Forward Selection method based on analysis of variance (RF-VA) is proposed in this paper. The method has been improved in two aspects: Firstly, a method based on analysis of variance is used to

measure the difference of features in different categories, and a modified permutation and replacement scheme is obtained to guide rearrangement of the order of values belongs to OOB data sample of a feature. Secondly, the Sequential Forward Selection method is used to obtain iterative feature evaluation results, which solves the problems of the randomness of the existing algorithm to determine the size of the feature subset and the instability of the results.

This paper designs eight kinds of artificial insulation defects according to the characteristics of the insulation defects and the experienced supervisors. Through the partial discharge Ultra High Frequency(UHF) detection system, the phase resolved pulse sequence data sets of each defect is obtained when the discharge is stable at different test voltages. This paper takes the experimental original data as the basic starting to study effective feature parameter extraction methods and feature selection algorithms.

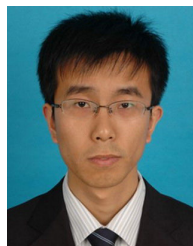
The experimental results show that the better feature subset can be obtained by the proposed RF-VA feature selection method. The classification accuracy rate remains high, the corresponding feature dimension obtained by RF-VA is lower than the RF feature selection method. It is concluded by comparing the classification results of the RF and RF-VA method to do dimensionality reduction on the single feature set and the combined feature set. The RF-VA feature selection method actually works for feature dimensionality reduction, which effectively improves the partial discharge defect type recognition rate.

In this paper, the proposed RF-VA feature selection method is used to optimize the partial discharge characteristics and realize the PD identification in GIS. It shows its effectiveness with a higher recognition rate and more reliable prediction. Moreover, this is an effective solution to the conflict of recognition results under different single feature sets. Last but not least, for field applications that do not have sufficient prior knowledge, this is also a promising method of identifying unknown defects, even if they are not sufficiently trained in the original database.

REFERENCES

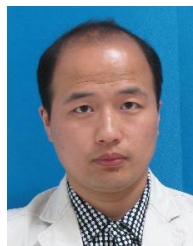
- [1] B. Chen, J. Hong, and Y. Wang, “The problem of finding optimal subset of features,” *Chin. J. Comput.*, vol. 20, no. 2, pp. 133–138, Feb. 1997.
- [2] S. Zhang, C. Li, K. Wang, J. Li, R. Liao, T. Zhou, and Y. Zhang, “Improving recognition accuracy of partial discharge patterns by image-oriented feature extraction and selection technique,” *IEEE Trans. Dielectr. Electr. Insul.*, vol. 23, no. 2, pp. 1076–1087, Apr. 2016.
- [3] X. Peng, J. Li, G. Wang, Y. Wu, L. Li, Z. Li, A. A. Bhatti, C. Zhou, D. M. Hepburn, A. J. Reid, and M. D. Judd, “Random forest based optimal feature selection for partial discharge pattern recognition in HV cables,” *Power Syst. Technol.*, vol. 43, no. 4, pp. 1229–1335, 2019.
- [4] W. Shiqiang, “Pattern recognition of partial discharge based on the feature parameter optimization selection and multi-algorithm combined methods,” *High Voltage App.*, vol. 54, no. 10, pp. 118–125, 2018.
- [5] F. Yang, Y. Xu, Y. Qian, Z. Li, G. Sheng, and X. Jiang, “Application of correlation analysis techniques in feature extraction and selection for DC partial discharge signals of XLPE cables,” *Power Syst. Technol.*, vol. 42, no. 5, pp. 1653–1660, 2018.
- [6] R. Genauer, J.-M. Poggi, and C. Tuleau-Malot, “Variable selection using random forests,” *Pattern Recognit. Lett.*, vol. 31, no. 14, pp. 2225–2236, Oct. 2010.

- [7] D.-J. Yao, J. Yang, and X. J. Zhan, "Feature selection algorithm based on random forest," *J. Jilin Univ.*, vol. 44, no. 1, pp. 137–141, 2014, doi: 10.13229/j.cnki.jdxbgxb201401024.
- [8] H.-G. Kranz, "PD pulse sequence analysis and its relevance for on-site PD defect identification and evaluation," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 12, no. 2, pp. 276–284, Apr. 2005.
- [9] A. Lapp and H.-G. Kranz, "The use of the CIGRE data format for PD diagnosis applications," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 7, no. 1, pp. 102–112, Feb. 2000.
- [10] H.-G. Kranz, "Fundamentals in computer aided PD processing, PD pattern recognition and automated diagnosis in GIS," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 7, no. 1, pp. 12–20, Feb. 2000.
- [11] H. Suzuki and T. Endoh, "Pattern recognition of partial discharge in XLPE cables using a neural network," *IEEE Trans. Electr. Insul.*, vol. 27, no. 3, pp. 543–549, Jun. 1992.
- [12] R. Yao, M. Hui, J. Li, L. Bai, and Q. Wu, "A new discharge pattern for the characterization and identification of insulation defects in GIS," *Energies*, vol. 11, no. 4, pp. 971–990, Nov. 2018.
- [13] F. H. Kreuger, E. Gulski, and A. Krivda, "Classification of partial discharges," *IEEE Trans. Electr. Insul.*, vol. 28, no. 6, pp. 917–931, Dec. 1993.
- [14] E. Gulski and F. H. Kreuger, "Computer-aided analysis of discharge patterns," *J. Phys. D, Appl. Phys.*, vol. 23, no. 12, pp. 1569–1575, Dec. 1990.
- [15] E. Gulski, "Computer-aided measurement of partial discharges in HV equipment," *IEEE Trans. Electr. Insul.*, vol. 28, no. 6, pp. 969–983, Dec. 1993.
- [16] R. Yao, Y. Zhang, G. Si, Y. Yuan, and Q. Xie, "Statistical operators calculation of partial discharge on floating electrode defect in GIS," in *Proc. Int. Conf. Condition Monitor. Diagnosis (CMD)*, Xi'an, China, Sep. 2016, pp. 585–590.
- [17] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
- [18] B. Li, J. H. Friedman, C. J. Stone, and R. A. Olshen, "Classification and regression trees (CART)," *Biometrics*, vol. 40, no. 3, pp. 358–361, 1984.
- [19] T. Bylander, "Estimating Generalization error on two-class datasets using out-of-bag estimates," *Mach. Learn.*, vol. 48, no. 1, pp. 287–297, 2002.
- [20] A. Gelman, "Analysis of variance," *Qual. Control Appl. Stats*, vol. 20, no. 1, pp. 295–300, 2005.
- [21] V. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed. New York, NY, USA: Springer, 1995, pp. 139–147.
- [22] P. N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, 1st ed. Reading, MA, USA: Addison-Wesley, 2005, pp. 256–270.



JUN LI received the B.S. and master's degrees in computer science and technology from North China Electric Power University in 2008 and 2011, respectively.

He is currently working with the State Grid Shaanxi Electric Power Research Institute. He is also a Senior Engineer with the Power Grid Research Institute. His research interests include the areas of partial discharge of high-voltage electrical equipment and fault diagnosis of electrical equipment based on machine learning.



MENG HUI (Member, IEEE) received the bachelor's and master's degrees in mechanical engineering and the Ph.D. degree in electrical engineering from Xi'an Jiaotong University in 2004, 2007, and 2011, respectively.

He is currently an Associate Professor with the School of Electronic and Control Engineering, Chang'an University. His research interests include the areas of fault diagnosis of electrical equipment based on machine learning and nonlinear dynamics in complex networks.



LIN BAI received the bachelor's degree in electronic information science and technology from Northwest University in 2003, the master's degree in electronic science and technology in 2006, and the Ph.D. degree in signal and information processing from Northwestern Polytechnical University in 2010.

He is currently an Associate Professor with the School of Electronic and Control Engineering, Chang'an University. His research interests include the areas of fault diagnosis of electrical equipment based on deep learning and information processing.



RUI YAO received the B.S. and M.S. degrees in control theory and control engineering from the North China Electric Power University in 2008 and 2011, respectively, and the Ph.D. degree in electrical engineering from Xi'an Jiaotong University in 2017.

She is currently a Lecturer with the School of Electronic and Control Engineering, Chang'an University. Her research interests include the areas of partial discharge of high-voltage electrical equipment and fault diagnosis of electrical equipment based on machine learning.



QISHENG WU received the B.S. degree from Dalian University of Technology in 1984 and the master's degree from Chang'an University in 1996. His current research interests include artificial intelligence and automobile control.

...