A method of sample enhancement based on partial discharge PRPS spectrum

Qi Tang¹, a, Lihui Wu¹, b, Rongbo Luo¹, c, Guowei Li¹, d, Junbo Wang¹, e, Xiaolong li¹, f, Shisong Yan², *

¹Foshan Power Supply Bureau of Guangdong Power Grid Corporation, Foshan 528000, China
²Huacheng Electric Technology Co., Ltd, Shanghai, 201114, China

*Corresponding author e-mail: 269155508@qq.com, *68843420@qq.com,
b58042788@qq.com, c529671978@qq.com, d1957852@qq.com, 
æhoneymoon651@126.com, f769681051@qq.com

Abstract. In this paper, we summarize the problems existing in the traditional partial discharge PRPS sample enhancement method, and propose a method based on the partial discharge PRPS spectrum sample enhancement, including threshold denoising, noise superposition, data density, data sparsity, phase shift, data enhancement method matrix and so on. It solves the problem that the difficulty of collecting the spectrum of partial discharge cases restricts the application of intelligent algorithms such as deep neural network. The practical application shows that the local partial discharge PRPS spectrum sample enhancement method can effectively enhance the sample size, make the sample distribution uniform and representative, and the deep learning neural network model trained by the sample set has high accuracy and generalization ability for PRPS spectrum diagnosis.

1. Introduction

Partial discharge (PD) is a discharge phenomenon caused by partial area breakdown in insulating medium. Different from breakdown or flashover, partial discharge is a tiny breakdown in the local area of insulation and the initial phenomenon of insulation degradation [1-2]. PD detection is considered to be the most important and effective method to evaluate the insulation state of electrical equipment. It has been widely studied and applied in the diagnosis and evaluation of the insulation state of electrical equipment. UHF (ultra high frequency) detection technology can detect primary equipment such as transformer, GIS, switch cabinet, etc. It has certain universality, and different defect type signals have obvious differences in PRPS spectrum, which is conducive to the diagnosis of different defects. Therefore, UHF technology is often used for PD detection and defect analysis of power equipment.

Electrical equipment insulation defects include corona, air gap, particle, suspension, and surface, and noise, normal and other eight categories. After UHF detection, the typical spectrum is shown in Figure 1.
It can be seen from the typical Atlas of eight kinds of partial discharge in Figure 1 that the phase distribution, periodic distribution and amplitude size of each type of defect are obviously different with great difference. Therefore, the study of using artificial intelligence algorithm to identify the UHF PRPS spectrum has a certain foundation.
At present, the mainstream artificial intelligence algorithm based on deep learning can effectively realize the classification of spectrum, which has been widely used at home and abroad [3]. The recognition accuracy of atlas depends on the granularity of sample classification and the number of samples in each category. The PRPS spectrum data of PD comes from the field disintegration cases of substation, but the probability of case occurrence is small and the distribution is uneven, which leads to the lack of sample data for deep learning algorithm training and the imbalance of classification, which limits the application and development of deep learning algorithm in UHF spectrum recognition of PD.

In this paper, aiming at the problems of small amount and uneven distribution of partial discharge PRPS spectrum data, which cannot meet the application requirements of deep learning algorithm in partial discharge UHF spectrum recognition, a variety of spectrum sample enhancement methods are studied to increase the amount and even distribution of partial discharge PRPS spectrum sample data. So as to establish a sample database and train the deep learning algorithm to realize the intelligent diagnosis of partial discharge PRPS spectrum.

2. Analysis of PD PRPS spectrum data enhancement
In the actual PD detection process, a large number of labeled training samples are needed for PD PRPS pattern recognition, and the cost of collecting labeled training samples is high. For the low frequency PD phenomenon, it is often difficult to quickly accumulate enough samples to support the training. In the traditional methods of PRPS sample enhancement, translation, rotation, scaling and other simple processing are the main methods. These methods do not take into account the complexity of the data in the real detection situation. All the data are processed by a unified standard, which reduces the authenticity of the data.

For the deep learning algorithm, when the number of samples in the data set is small and the sample similarity is high, it is easy to cause over fitting, and the generalization ability of the algorithm is poor. So it cannot be effectively applied to the application of PRPS spectrum diagnosis.

2.1. Insufficient samples
In the field of computer artificial intelligence, the sample data for deep learning training can usually be obtained by actual sample collection, laboratory simulation, artificial synthesis, etc. However, in the application scenario of PD detection, because of the complex environment of the substation, the defects of primary equipment are different, and the environment has various interferences, the typical defects can only be simulated in the laboratory. The diagnosis algorithm trained by the sample of typical defects has low accuracy in practical application. Partial discharge PRPS spectrum is different from natural photos (such as trees, buildings, cats, dogs, etc.) [4], and has a certain degree of professionalism (specifically manifested by the use of three-dimensional spectrum, whose X-axis represents phase, Y-axis represents period, z-axis represents amplitude, and the image should be complete). Therefore, the traditional translation, rotation, scaling and other sample enhancement methods cannot be applied to the local PRPS sample data. It is difficult to solve the problem of insufficient samples.

2.2. Uneven distribution of samples
In the case found in the substation site, the defects found in the switch cabinet are mainly along the surface, the defects found in the transformer are mainly suspended, the defects found in GIS are mainly suspended, and the environment in the substation is mainly noise, normal and other, while the probability of corona, air gap, particles and other defects is small. So it is difficult to accumulate more samples. This results in uneven distribution of PRPS spectrum data.

The effective diagnosis algorithm should be able to correctly identify all kinds of defects, and the final generalization ability and fitting ability of the algorithm are determined by the sample distribution and the number of samples. The uneven distribution of samples seriously affects the application of deep learning algorithm in the PRPS spectrum recognition.
3. Sample enhancement method of PD PRPS spectrum

In this paper, a sample enhancement method based on the matrix of enhancement method is proposed to enhance the PRPS spectrum. The data enhancement methods in the matrix include threshold denoising, noise superposition, data density, data sparseness and phase shift. Each method can be used as an entrance to the matrix. PRPS data can be used for random method and random number of sample enhancement, and the new samples can be used to train the network with stronger generalization ability. The overall process is as follows:

**Figure 2.** PD PRPS spectrum data enhancement matrix

3.1. Threshold denoising
The process and idea of threshold denoising are shown in Figure 3:
Figure 3. Threshold denoising flowchart

The input data of PRPS spectrum is three-dimensional data with phase as X-axis, period as Y-axis and amplitude as Z-axis. The amplitude range of PRPS spectrum is 0-70db, which is divided into five sub ranges. The number of pulses in each sub range is counted respectively. The amplitude histogram statistics are shown in Figure 4:

Figure 4. Amplitude histogram
Through statistical amplitude histogram, the small amplitude data with many values is removed, that is, 27 pulse data in the range of 0-14 amplitude in Figure 4 are deleted. The data distribution of each phase is calculated in the same way, the scattered data is removed, and the threshold denoised data is obtained to form the enhanced spectrum data.

3.2. Noise superposition
The overall idea of noise superposition is that the input data is three-dimensional data with phase as X-axis, period as Y-axis and amplitude as Z-axis. The data is expressed in the form of two-dimensional array (such as \( \text{Image}[x][y]=v \), i.e. when \( x \) phase and \( y \) period, the amplitude is \( v \)). The radar noise, mobile phone noise, microwave sulfur lamp interference and other data features are respectively generated into corresponding three-dimensional data, and the input data and noise data are accumulated to obtain the noise coupled data.

Define the sample data to be enhanced as \( \text{DstImage}[x][y] \), define the enhanced sample data as \( \text{EnhDstImage}[x][y] \), define the added radar noise variable \( rv \) as \( \text{val} \), and calculate \( \text{val} \) by the following formula:

\[
\text{val} = \text{DstImage}[x][y] + rv
\]  
(1)

The range of \( \text{val} \) is modified by the following formula:

\[
\begin{align*}
\text{if}(\text{val} < 0) &\{\text{val} = 0;\} \\
\text{if}(\text{val} > 255) &\{\text{val} = 255;\}
\end{align*}
\]  
(2)

(3)

Redefine the sample data value:

\[
\text{EnhDstImage}[x][y] = \text{val}
\]  
(4)

Through the above formula, all the data in the \( \text{DstImage} \) array are processed to form the enhanced sample data \( \text{EnhDstImage} \).

3.3. Dense data
The idea of data density is to randomly add data points on the phase with data, with an increase range of [10%, 20%]. The data amplitude dynamically takes the average amplitude of all current data points in this phase, and then the data is dense after superposition.

 Traverse each phase and record the number of PD pulses \( p \) on each phase, calculate the average amplitude of all PD pulses on each phase, which is the ratio of the sum of all PD pulse amplitudes to the number of PD pulses; use random number generator to generate an integer \( k \) for each phase randomly, and the output number of random number generator is an integer, with the range of 10% \( p \sim 20\% p \). A random number generator is used to randomly generate \( k \) integers \( Z\alpha \) for each phase, where \( \alpha = 1,\ldots, k; \) \( Z\alpha \) satisfies the following requirements:

\[
0 < Z\alpha < T
\]  
(5)

\( T \) represents the period, the value of integer \( Z\alpha \) represents the position number of the PD pulse to be added in each phase, and the value of the PD pulse corresponding to each position number is set to the average amplitude of the PD pulse in each phase to obtain the densified 3D PRPS spectrum data.
3.4. Data sparseness
The idea of data density is to randomly reduce data points in the phase with data, the reduction range is [10%, 20%], and the proportion of data points reduced in each phase is the same; after reduction, the sparse data can be obtained.

Traverse each phase and record the number \( d \) of PD pulses on each phase; use random number generator to randomly generate integer \( q \) for each phase, and the output number of random number generator is integer, with the range of 10% \( d \) ~ 20% \( d \); use the second random generator to generate \( q \) integer \( H_\beta \) for each phase, where \( \beta = 1, \ldots, q \); \( H_\beta \) meets the following requirements:

\[ 0 < H_\beta < T \]  \hspace{1cm} (6)

\( T \) is the period, the integer \( H_\beta \) is the position number of the PD pulse to be thinned on each phase, and the PD pulse corresponding to each position number is set to 0; the thinned 3D PRPS spectrum data is obtained.

3.5. Phase shift
The idea of phase shift is as follows: firstly, a random integer \( n \) between \([0,360]\) is generated, and the input data is shifted to the left by \( n \) intervals according to the phase. The generation formula of random data \( n \) is as follows:

\[ n = \text{rand}() \% (b - a + 1) + a \]  \hspace{1cm} (7)

Where, \( a \) is the lower limit of the random number, which is 0, and \( b \) is the upper limit of the random number, which is 360.

Shift the input data \( DstImage[x][y] \) \( n \) intervals to the left according to the phase, and get the translated data \( EnhDstImage[x][y] \), as follows:

\[ EnhDstImage[x][y] = DstImage[x - n][y] \]  \hspace{1cm} (8)

Get the data after the phase shift \( EnhDstImage[x][y] \).

3.6. Application of data enhancement method matrix
The input data is the original data of PD cases found and disassembled for acceptance in the substation. The input data is transformed into three-dimensional data with phase as X-axis, period as Y-axis and amplitude as Z-axis.

After the raw data is input through the data enhancement method matrix, one of the methods can be randomly selected as the entry, and can be stopped at any time until the iteration reaches the required sample size.

The new data is transformed into spectrum file, and the enhanced sample data is obtained.

4. Comparative analysis
Take the original sample set \( a \) and the enhanced sample set \( B \) as examples. The information of the sample set is shown in Table 1:
Table 1. Sample set table

| Serial number | Defect types | Number of sample set A | Number of sample set B |
|---------------|-------------|------------------------|------------------------|
| 1             | Corona      | 63                     | 700                    |
| 2             | Air gap     | 51                     | 680                    |
| 3             | Particle    | 38                     | 600                    |
| 4             | Suspension  | 280                    | 700                    |
| 5             | Surface     | 312                    | 600                    |
| 6             | Noise       | 619                    | 610                    |
| 7             | Normal      | 1074                   | 600                    |
| 8             | Other       | 195                    | 600                    |
| **Subtotal**  |             | **2632**               | **5090**               |

As can be seen from Table 1, the sample distribution of original sample set A is extremely uneven. Specifically, the proportion of samples with "normal" defect type is 40.81%, while the proportion of samples with "particle" defect type is only 1.44%, and the total number of samples is only 2632.

Using the data enhancement method proposed in this paper, we get a sample set B, the sample distribution of each defect type is relatively balanced, and the sample problem is 5090, an increase of nearly 93.39% compared with sample set A.

In order to verify the usability of the enhanced samples, GoogLeNet deep learning algorithm was used to build a GPU-based training environment and train the sample set A and B respectively. The process is as follows:

1) Divide the defect data of sample set A and sample set B into 3:7 parts, in which 30% is the test sample set and 70% is the training sample set;
2) Training samples of sample set A and B were used to train the GoogLeNet deep learning algorithm, and model A and model B were obtained;
3) Using Java code, model A and model B are encapsulated to obtain diagnostic interface A and diagnostic interface B;
4) Using the test sample data of sample set A, call the diagnostic interface A and record its accuracy;
5) Using the test sample data of sample set B, call the diagnostic interface B and record its accuracy;
6) Compare the two test results, as shown in figure 5:

![Diagnostic accuracy statistics](image_url)

**Figure 5.** Diagnostic accuracy statistics
By comparing the identification effect of defect types, it was found that the diagnosis accuracy of each defect type was higher in diagnostic model B, and the overall performance was stable, which was significantly better than that of diagnostic model A.

5. Conclusion
The sample enhancement method of PRPS spectrum proposed in this paper can effectively enhance the number of samples while ensuring the quality of samples. It solves the problem that deep learning algorithm is difficult to be effectively applied to the diagnosis of partial discharge PRPS spectrum due to the lack of effective sample set, including the following aspects:

1) A large number of labeled samples are generated with less calculation, which solves the problem of high acquisition cost of labeled samples and insufficient training data.
2) On the basis of the original data samples, the complexity of the data is fully considered, and the authenticity and diversity of the data under various circumstances such as data noise, data phase shift, data attenuation and data collection loss are introduced.
3) It can meet the requirements of training samples for the balance and quantity of each label, the extended training data can be used to train the model to avoid the over-fitting of the model.

Acknowledgments
This work was financially supported by science and technology project fund (No.030600KK52180133) of Guangdong Power Grid Co., Ltd.

References
[1] QIU Changrong, WANG Naiqing. Partial discharge of electrical equipment and its testing technology [M]. Beijing, China: Machinery Industry Press, 1994: 36 - 42.
[2] YAN Zhang, ZHU Deheng. High voltage insulation technology [M]. Beijing, China: China Electric Power Press, 2002: 205 - 206.
[3] Cheng Xiaqi. Research on feature extraction and correlation analysis of feature in PD pattern recognition [D]. Chongqing University, 2003.
[4] Ma Dongmei. Research on image retrieval based on deep learning [D]. Inner Mongolia University, 2014.
[5] Zhang Kun, Zhang Dongping, Yang Li. Research on face recognition algorithm based on sample enhancement. School of information engineering, China University of metrology, 2018.