Front. Electr. Electron. Eng. 2012, 7(4): 391–398
DOI 10.1007/s11460-012-0208-9

RESEARCH ARTICLE

R. AMBIKAIRAJAH, B. T. PHUNG, J. RAVISHANKAR

The fusion of classifier outputs to improve partial discharge classification

Abstract The detection of partial discharge signals and classifying its patterns is an area of interest in the analysis of defects in high voltage cables. This paper investigates a filter-bank based approach to extract frequency domain based features to represent partial discharge signals. By applying the fast Fourier transform, the sampled partial discharge data are mapped into equivalent discrete frequency bins, which are then grouped into N equal sub-bands and also octave sub-bands, each providing N-dimensional features for partial discharge pattern classification. Two classifiers, namely, the support vector machine and the sparse representation classifier, are implemented and their outputs are fused, in order to improve the accuracy of classifying partial discharge. Classification accuracy is also compared with wavelet domain based octave frequency sub-band features.

Keywords partial discharge, features, fusion, classification

1 Introduction

Electrical insulation is essential in high voltage power equipment and components. The detection of partial discharge (PD) is considered to be a mechanism to predict the aging of equipment insulation and is a precursor to insulation breakdown. Time domain PD signal analysis and phase resolved PD patterns have been widely used as methods to investigate the different types of PD. The PD signals are characterized by a very short rise time and duration, some with frequencies in the range of GHz, at the location where the PD first occurs. When the PD signal propagates such as in a high voltage cable, it is subject to attenuation and distortion due to the frequency-dependent distributed electrical characteristic of the cable. Sensors such as high frequency current transformers (HFCTs) are often used to detect the PD signals. The measured signals are the convolution of the original signals that were generated at the PD source with the combined impulse response of the test object (cable) and the measurement setup. When the output of the HFCT is sampled, time domain, frequency domain, or time-frequency domain based methods can be used to analyze this signal. In situations where there is more than one PD source, frequency domain or wavelet based analysis are effective in locating and identifying the multiple PD sources [1].

A key challenge in PD signal detection is to discriminate the signals from noise. Sriram et al. [2] have compared many de-noising techniques for PD signals. Kyprianou et al. [3] have investigated the use of wavelet packet based de-noising and claimed that their technique successfully recovered the PD signal in many cases. Recently, a de-noising technique employing a signal boosting algorithm in the frequency domain has been effectively used to de-noise a variety of PD signals that includes laboratory data and on-site data [4].

This paper looks at classifying the different types of PD signals that occur. To date, a number of signal processing algorithms that extract features from the recorded PD data have been used. These include wavelet transform, wavelet packet transform, higher order statistical analysis, and fast Fourier transform (FFT) [3,5]. A filter-bank based approach to extract frequency domain based features to represent a PD signal is investigated in this paper. These features are then used as an input to a classifier.

Pattern classification tools such as the probabilistic neural network (PNN), support vector machine (SVM), artificial neural network (ANN) with back propagation, and sparse representation classifier (SRC), have all been used as PD signal classifiers. The fusion of classifier outputs has been used in order to improve the accuracy of the classification [6]. There are many types of fusion techniques, including linear score weighting and artificial

Received August 14, 2012; accepted September 19, 2012

R. AMBIKAIRAJAH (✉), B. T. PHUNG, J. RAVISHANKAR
School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW 2052, Australia
E-mail: r.ambikairajah@student.unsw.edu.au

neural network fusion. In this paper, two particular classifiers are studied, i.e., SVM and SRC, and the outputs of these two classifiers are fused using the linear score weighting technique in order to improve the classification accuracy of PD signals.

2 De-noising and feature extraction

2.1 De-noising

The de-noising of the PD signal is carried out in the frequency domain, where the PD signal is decomposed into N octave frequency sub-bands using the FFT. The energy corresponding to each sub-band is then calculated using the magnitude of the FFT. A gain factor for each sub-band is calculated, as outlined in Ref. [5] and shown in Fig. 1, and is used to suppress the noise and either retain or boost the magnitude components in each sub-band.

This technique is referred to as the signal boosting technique. When the original signal is not corrupted by much noise, then the calculated gain factor is high; thus, providing a boost to the magnitudes in that sub-band. When the noise in a band is much greater than the signal, the gain factor is closer to zero, which results in the magnitude components in that band being brought to zero.

The de-noised sub-band data obtained, as shown in Fig. 1, is used to reconstruct the signal with the use of the inverse FFT. Figure 2 shows that this technique can extract PD pulses that are both embedded in noise or present outside the noise. In Fig. 2, $N = 7$, where the first octave frequency sub-band has a bandwidth ranging from 0 to 0.7812 MHz and the bandwidth of the seventh octave frequency sub-band ranges from 25 MHz to 50 MHz, and the recorded PD data was sampled at 100 MHz. Figure 2(b) shows the reconstructed signal.

2.2 Feature extraction

A pattern that can be recognized and classified has a number of discriminatory attributes and features. In the process of feature extraction, it is important to identify which of these features should be chosen and how they can be extracted. Evagorou et al. [7] used wavelet packet analysis to extract features from a PD signal, where they viewed the wavelet coefficients as a realization of a stochastic process and used statistical moments as representations of PD features. On the other hand, frequency domain based features were proposed and used by Ref. [5], where octave frequency sub-bands were used, with seven sub-bands that matched a 6-level discrete wavelet transform (DWT) decomposition.

In this paper, the authors investigate the use of equal sub-bands, and octave frequency sub-bands in order to extract discriminatory features for PD signals. Figure 3 highlights the steps involved in the feature extraction processes. In Fig. 3(a), the Fourier transform maps the PD signal into 2 million discrete frequency bins. These bins are then grouped to form N sub-bands. Each sub-band is de-noised and the energy corresponding to each sub-band is then calculated. The resulting energy vector (E_1, E_2, …, E_N) is smoothed using the discrete cosine transform to

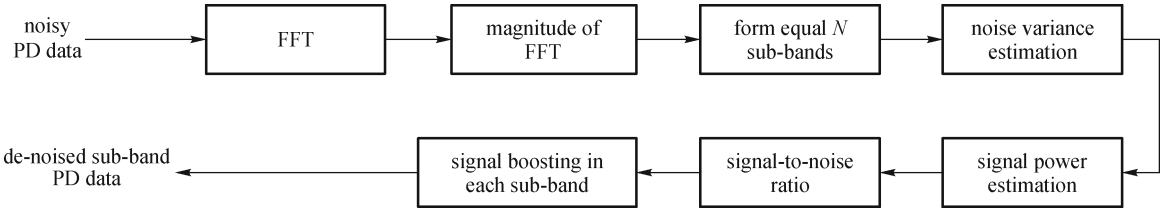

Fig. 1 De-noising PD signals in the frequency domain

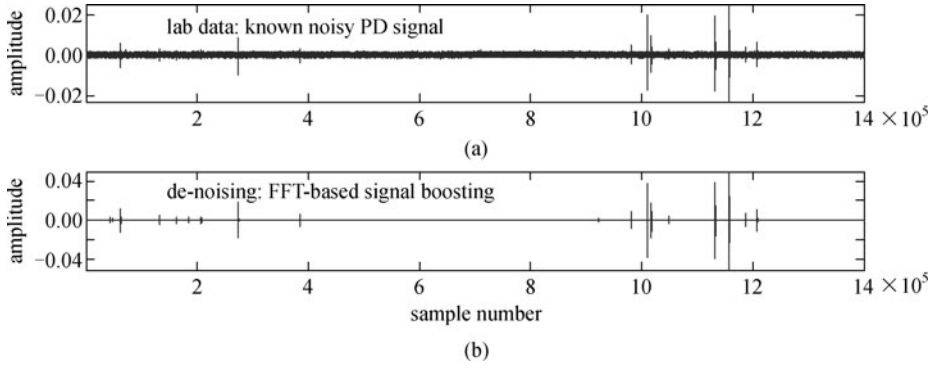

Fig. 2 De-noising of PD data using signal boosting technique

obtain transformed coefficients (C_0, C_1, C_2, …, C_{N-1}) and these are used as features for the classifier.

The PD events were recorded in the laboratory environment and on-site over an entire AC cycle of 20 ms. The bandwidth of each sub-band depends on the value of N, where the frequency bins can be grouped into either equal or octave sub-bands.

Figure 3(b) demonstrates how the N-dimensional features were obtained using the DWT. The DWT used in this paper has a 6-level decomposition (seven octave frequency sub-bands). The authors have used two different noise reduction techniques for the DWT. In Ref. [5], a threshold-based de-noising method was used [8], whereas in this paper a signal boosting de-noising technique was used for DWT, as outlined in Section 2.1 of this paper.

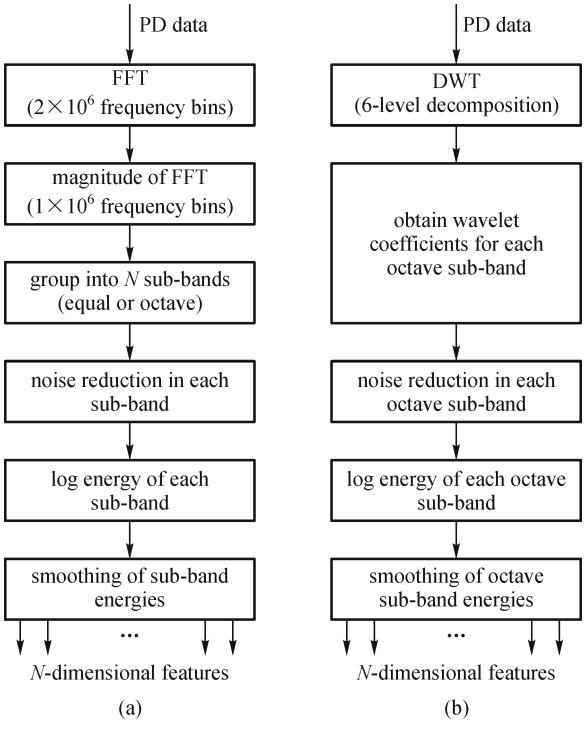

Fig. 3 (a) FFT-based feature extraction process (equal sub-bands); (b) DWT-based feature extraction process (octave frequency sub-bands)

Figure 4 shows the difference between the signal boosting and thresholding de-noising techniques. Figure 4(a) shows that the energy in the first octave frequency sub-band is boosted and the others are suppressed if it is likely that it is noise. Figure 4(b) shows that the thresholding technique loses energy, as any wavelet coefficients below the threshold is removed, regardless of whether it is noise or not.

It should be noted that in the DWT, sub-band 7 is the lowest energy band (0 to 0.7812 MHz); however, for the purposes of illustration in Fig. 4, Band 1 in both Figs. 4(a) and 4(b) represents sub-band 7.

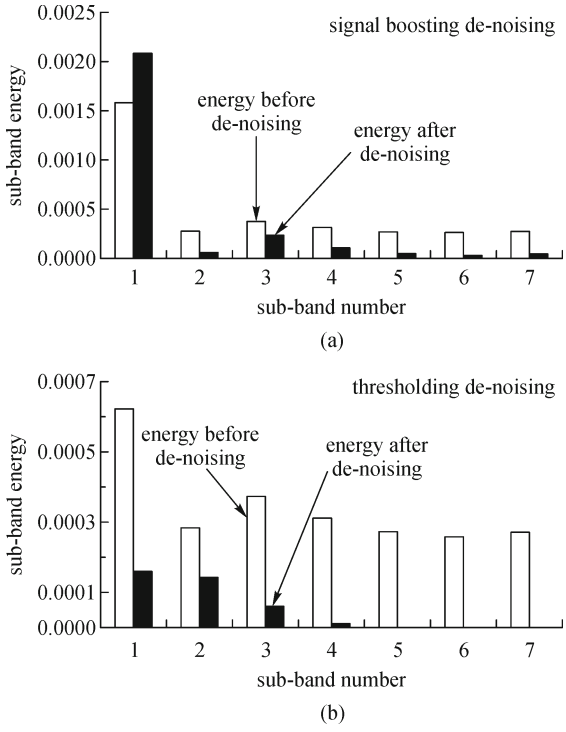

Fig. 4 (a) De-noising of PD data using signal boosting technique; (b) de-noising of PD data using thresholding technique

3 Classifiers

3.1 Types of classifiers

Literature records many classifiers being used for pattern recognition. For the classification of partial discharge signals, the most common classifiers used are PNN and SVM [1,7]. Other classifiers such as neural networks and Bayes classifier have been trialed, but with limited success. This paper will look at SRC in detail, and compare its performance in classifying PD signals with SVM. The SRC has been used successfully for face recognition [9] and recently this classifier has also been used for classifying partial discharge data [5].

Figure 5 shows a PD classification system where there are two distinct phases, namely, the training phase and the test phase. During the training phase, the features are extracted from the training data and that classifier is trained to create a model for each class. In the test phase, the features extracted from the test data are compared with the model created in the training phase and a decision is made based on the pre-determined criteria for each class.

3.2 Classification based on sparse representation

Pattern recognition, in its most basic form, looks at labeled training data from k distinct classes, and then based on this information, identifies the class of a test pattern. Assume that there are N training data available from the ith class.

Each training data (in the case of this paper, this is recorded PD data of 2 million time samples), is mapped to an L-dimensional feature vector (L = 7) using an FFT-based filter bank as outlined in Section 2.2 of this paper.

Let the resultant L-dimensional feature vector be t_{ij}, where i is the class index (i = 1, 2, 3, …, k) and j is the training feature vector index (j = 1, 2, 3, …, N). All training data feature vectors from the ith class are put into a matrix as columns: $A_i = [t_{i1}, t_{i2}, t_{i3}, \ldots, t_{iN}]$. Provided that there are sufficient feature vectors of the training data in the ith class, a test pattern y from the same class will approximately lie on a linear subspace: $y = a_{i1}t_{i1} + a_{i2}t_{i2} + \cdots + a_{iN}t_{iN}$, where a_{ij} are scalar quantities.

A dictionary matrix A is now developed for all k classes by concatenating A_i for each class:

$$A = [A_1, A_2, A_3, \ldots, A_i, \ldots, A_k] = [t_{11}, \ldots, t_{1N} | t_{21}, \ldots, t_{2N} | \cdots | t_{k1}, \ldots, t_{kN}]. \quad (1)$$

Class 1 Class 2 Class k

The number of column vectors in $A_1, A_2, A_3, \ldots$ depends on the number of training data available for each class. However, it is assumed, in this case, that each class has N training data. The test pattern y can now be represented as a linear combination of all training data feature vectors:

$$y = Ax, \quad (2)$$

where $x = [0, 0, 0, \ldots, a_{i1}, a_{i2}, \ldots, a_{iN}, 0, 0, 0]^{\mathrm{T}}$ and this is a coefficient vector whose entries are zero, except those associated with the ith class. y is a single column vector with L rows, the size of matrix A is $L \times kN$, and the single column vector x has kN rows. Let $K = kN$, where K is the total number of training data feature vectors available for all five classes. There are three possible cases that can occur in solving Eq. (2) as shown in Fig. 6.

In Fig. 6(a), the matrix A is known as a complete dictionary, as there are the same number of coefficients (a_{ij}) and equations and there is only one solution. By solving simultaneous equations, the coefficients can be obtained. In Fig. 6(b), the matrix A is known as an over-complete dictionary, and Eq. (2) is known as an over-determined system of linear equations and the correct x can be found as its unique solution.

However, in Fig. 6(c), $L < K$ and the system of linear equations for $y = Ax$ is known as under-determined, and the solution is not unique. The matrix A is known as an under-complete dictionary. Out of the vast number of solutions available, a sparsest solution needs to be found. In the case of PD data, the dimensionality feature vector is much less than the number of columns of training vectors in the under-complete dictionary. The optimization equation in Eq. (3) can be solved, to obtain the sparsest solution:

$$\arg\min \|x\|_0 \text{ subject to } y = Ax, \quad (3)$$

where $\|x\|_0$ denotes the l^0-norm, that counts the number of nonzero coefficients in x. Solving Eq. (3), however, is not an easy task. Alternative approaches have been developed to approximate l^0-norm; however, it takes many iterations to converge. Iterative methods, such as matching pursuit or orthogonal matching pursuits [10], provide a sub-optimal solution. Recently, Ref. [11] approximated the l^0-norm, using a smoothing function that provided good convergence properties.

The basis pursuit (BP) [10] is another approach that is used, where the l^0-norm is replaced by l^1-norm, as in Eq. (4):

$$\arg\min \|x\|_1 \text{ subject to } y = Ax, \quad (4)$$

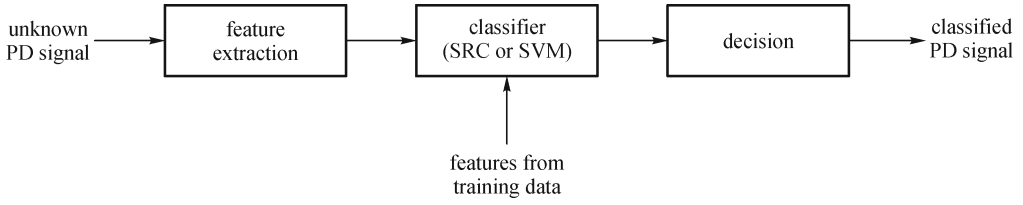

Fig. 5 A PD classification system

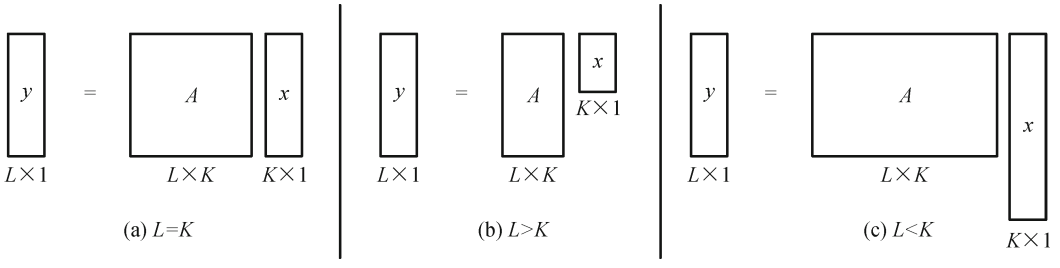

Fig. 6 (a) $L = K$, where there are K coefficients and equations; (b) $L > K$, where there are a larger number of equations compared to the number of coefficients; (c) $L < K$ where there are a larger number of coefficients compared to the number of equations

where $\|x\|_1$ denotes the l^1-norm. This minimization problem can be solved efficiently by the BP using linear programming. As real data are noisy, Eq. (4) can be modified to include a noise term as in Eq. (5):

$$y = Ax + \varepsilon, \tag{5}$$

where ε is white Gaussian noise. The sparsest solution x can still be approximately recovered by solving the minimization problem in Eq. (6):

$$\arg\min \|x\|_1 \text{ subject to } \|y - Ax\|_2 \leqslant \varepsilon. \tag{6}$$

A generalized version of Eq. (6) is to find x such that the function (7) is minimized:

$$\arg\min (\|y - Ax\|_2 + \mu\|x\|_1), \tag{7}$$

where μ is a scalar regularization parameter and $0 \leqslant \mu \leqslant 1$. It balances the trade-off between the reconstruction error and the sparsity. The weight of l^1-norm is controlled and this regularization that imposes the l^1 constraint is called LASSO [12]. The elastic net method [12] imposes a mixture of an l^1-norm and an l^2-norm constraint on x. It is given by

$$\arg\min (\|y - Ax\|_2 + \mu_1\|x\|_1 + \mu_2\|x\|_2^2), \tag{8}$$

where μ_1 and μ_2 are weights that control the l^1-norm and l^2-norm and $\mu_1 + \mu_2 = 1$. The l^1-norm in Eq. (8) enforces the sparsity of the solution and the l^2-norm has a smoothing effect to stabilize the solution that is obtained. When $\mu_1 = 0$, Eq. (8) falls under ridge regression [12] and when $\mu_1 = 1$, Eq. (8) becomes LASSO. μ_1 can be chosen by trial and error for any given problem. In the experiments of this paper, Eq. (8) is implemented to obtain the coefficients of vector x.

Given a test data feature vector y, from one of the classes in the global dictionary A, its coefficient vector x is computed via Eq. (8). Ideally, x should have nonzero values corresponding to a class y and will be associated with the columns of matrix A from that class. The test vector y can easily be assigned to that class; however, a modeling error can lead to small nonzero entries associated with multiple classes. To resolve this, the residual error for the ith class can be computed as follows:

$$r_i = \|y - Ax_i\|_2, \tag{9}$$

where $i = 1, 2, 3, \ldots, k$. Now, the test vector y is assigned to the class that minimizes the residual error r_i that is calculated using Eq. (9).

Equation (8) was solved using the under-determined matrix A, for a 5-class problem in order to calculate the coefficients of vector x for a particular Class 1 test vector. The test vector had 7 dimensions and matrix A comprised 7 rows and 186 columns of the training data from all five classes. The x vector had 186 coefficients. The coefficients are plotted in Fig. 7(a) and boundaries (columns of A) for each class are shown. Figure 7(b) shows the residual error, r, calculated for each class. It can be seen from Fig. 7(b) that Class 1 has the minimal residual error and the test vector was accurately classified.

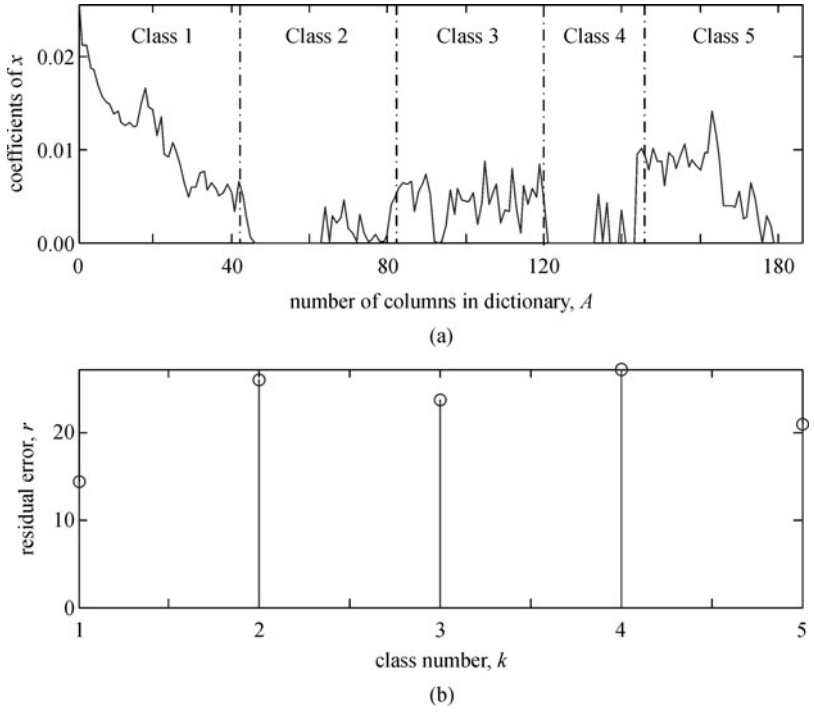

Fig. 7 (a) Coefficients of x for a particular Class 1 test vector; (b) residual error, r, calculated for each class

3.3 Fusion of classifier outputs

The combination of the outputs from a variety of classifiers has been investigated [6]. This process is referred to as classifier fusion. The overall conclusion from classifier fusion is that if the outputs from different classifiers can be fused efficiently; then, the classification accuracy can be better than the individual classifier performances. One of the techniques used for classifier fusion is called the linear weighting technique. The fused output, F_o, can be obtained as follows:

$$F_o = \sum_{k=1}^{N} \alpha_k C_k, \tag{10}$$

where N is the number of classifiers, C_k is the classifier output, and $\alpha_k > 0$ is the fusion weight. The sum of the fusion weights must be equal to 1. These weights can be empirically selected to optimize the overall classification performance. Figure 8 shows the process of classifier fusion for two classifiers (SRC and SVM). The maximum value in F_1 to F_k is calculated and the index corresponding to this maximum value is chosen as the class.

4 Experimental set-up and results

4.1 Data collection

Experiments were set up in a high voltage laboratory to create a 5-class problem on the three types of PD: surface, corona, and internal. A PD measurement circuit was set up and an HFCT together with a digital oscilloscope was used to record the PD signal. For each type of PD, the applied voltage was steadily increased; this data recorded was used to create the dictionary for the SRC and also used to train the SVM. The applied voltage was also steadily decreased for each PD type and this recorded data was used as test data. Variations on this included reversing the polarity of the circuit, and this data (surface and corona discharges) was also collected, thus producing a 5-class problem. Table 1 shows the details of all test and training data collected for five classes.

Table 1 Test and training data for five classes

class type	training data	test data
Class 1: surface discharge	44	22
Class 2: corona discharge	42	25
Class 3: internal discharge	34	21
Class 4: corona discharge with polarity reversal	23	8
Class 5: surface discharge with polarity reversal	43	22

Figure 9 shows examples of single event waveforms extracted from the recorded PD for each of the five classes. Note that the time interval is 10 ns between successive sampled points.

It can be seen from these waveforms that corona discharge has a long decay time, whereas in polarity reversal, its decay is faster. Surface discharge only has a few oscillatory cycles with a fast decay, which is in contrast to its polarity reversal. Surface discharge with polarity reversal also has a high oscillatory frequency. Internal discharge has a low oscillatory frequency with a reasonably quick decay. Overall, it can be seen that visually all five waveforms shown have different characteristics and thus one expects discriminatory features can be extracted.

4.2 Results

The 7-dimensional features were extracted as per the process in Fig. 3 for all the training and test data. These features were extracted using both the frequency domain technique and wavelet based techniques after undergoing the de-noising process. The dictionary A for the SRC was created using the training data features and x was calculated from Eq. (8) using the test data. Table 2 shows the classification accuracy for the two classifiers

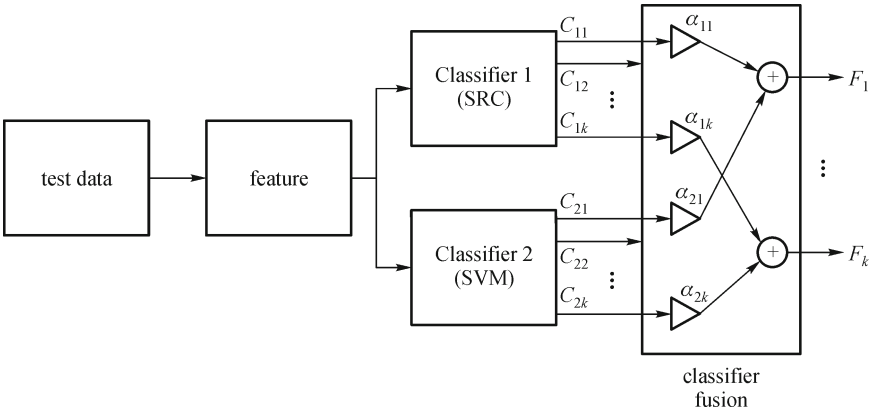

Fig. 8 Classifier fusion process using SRC and SVM classifiers, e.g., $F_1 = \alpha_{11}C_{11} + \alpha_{21}C_{21}$ and $\alpha_{11} + \alpha_{21} = 1$

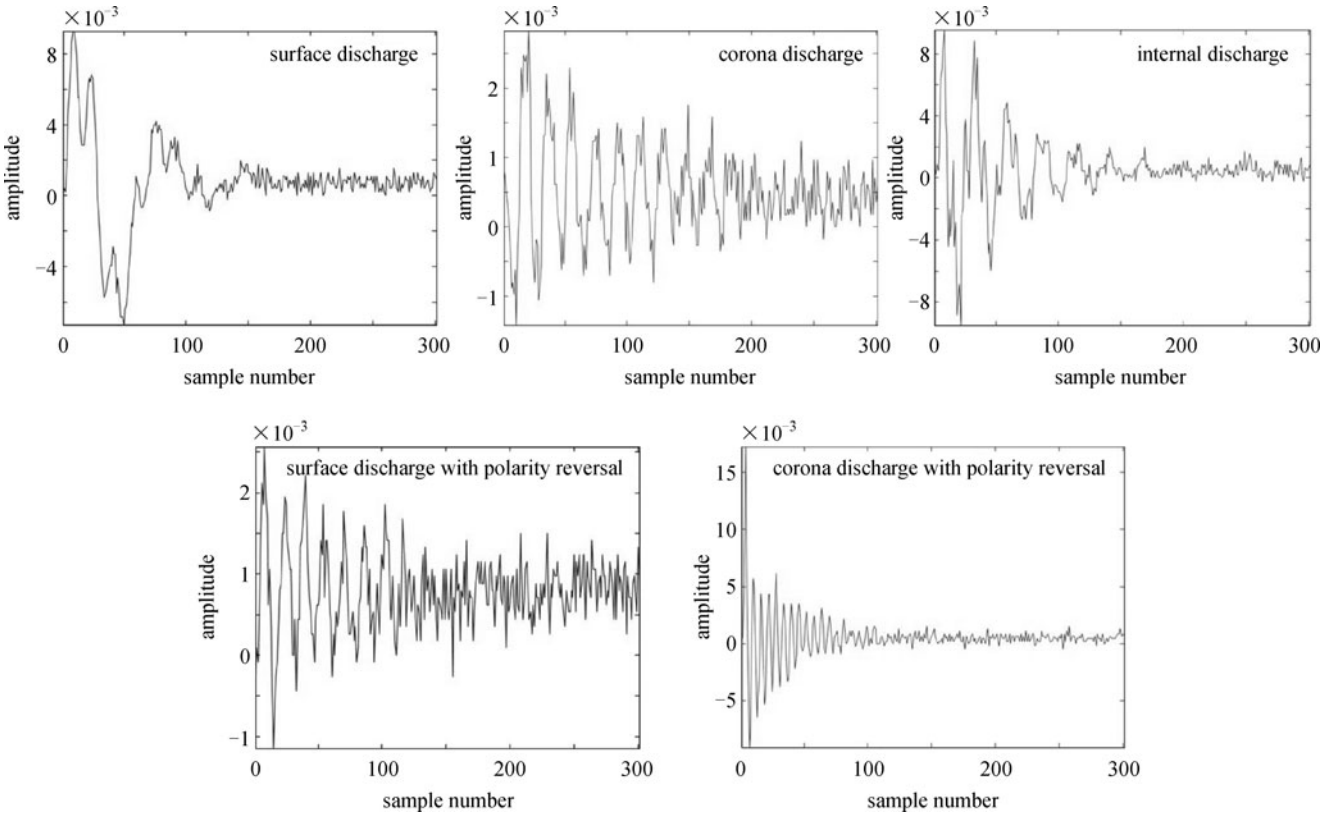

Fig. 9 Waveforms of the five different classes of partial discharge used in this paper

independently and also the fused classifier outputs. It can be seen that the accuracy of the fused classifier output is much higher than the single classifier output.

The confusion matrix for the FFT (7 equal bands) feature, for the SVM, SRC, and the fused classifier outputs, is presented in Table 3. For the SVM classifier, 5 test data in Class 2 have been misclassified as Class 3, and 2 test data in Class 4 have been misclassified as Classes 2 and 5.

In the SRC classifier, 9 test data from Class 3 have been classified inaccurately, along with data from Classes 2 and 4. However, when the outputs of both classifiers are combined with the appropriate weights, the fused classifier output has improved to 97.9% with errors in only 2 test data from Classes 2 and 4. It can be concluded that both the SVM and SRC classifiers have complementary information and when combined, result in a better classification.

5 Conclusion

In this paper, the authors have looked at the fusion of classifier outputs to improve the classification accuracy of PD signals. The de-noising process of PD signals, as well as the extraction of features in the frequency domain is presented with schematics in this paper. The two classifiers used are SVM and SRC. The sparse representation is analyzed in detail in this paper, as it is a relatively new classifier and, to the authors' knowledge, has not yet been used for the purpose of classifying PD data. There are several methods for fusing classifiers; however, this paper uses the linear weighting technique.

Results are presented for two sets of features — one with equal frequency sub-bands using FFT and the other with octave frequency sub-bands using DWT. Both

Table 2 Classification accuracy for SVM, SRC, and the fused classifier outputs

classifier	features (C_0, C_1, C_2, …, C_6)	
	FFT (7 equal bands)	DWT (7 octave bands)
support vector machine (SVM)	96.9%	90.8%
sparse representation classifier (SRC)	82.7%	79.6%
fused classifier outputs	**97.9**%	**96.9**%

Table 3 Confusion matrix for FFT feature with two classifiers and the fused outputs

classes	FFT (7 equal bands) with SVM classifier					FFT (7 equal bands) with SRC classifier					fused classifier outputs				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
1	22	1	0	0	0	22	5	4	2	0	22	1	0	0	0
2	0	24	0	1	0	0	20	5	1	0	0	24	0	0	0
3	0	0	21	0	0	0	0	12	0	0	0	0	21	0	0
4	0	0	0	6	0	0	0	0	5	0	0	0	0	7	0
5	0	0	0	1	22	0	0	0	0	22	0	0	0	1	22
accuracy	96.9%					82.7%					97.9%				

features are analyzed via SVM, SRC, and the fused classifier outputs. The results are compared, and the fused classifier outputs demonstrate a greater accuracy for both the equal and octave frequency sub-bands.

Future research can investigate the minimum number of features required to accurately classify PD signals. To do this, a larger database of PD data needs to be collected, under a wide range of conditions and a variety of applied voltages.

References

1. Hao L, Lewin P L. Partial discharge source discrimination using a support vector machine. IEEE Transactions on Dielectrics and Electrical Insulation, 2010, 17(1): 189–197
2. Sriram S, Nitin S, Prabhu K M M, Bastiaans M J. Signal denoising techniques for partial discharge measurements. IEEE Transactions on Dielectrics and Electrical Insulation, 2005, 12(6): 1182–1191
3. Kyprianou A, Lewin P L, Efthymiou V, Stavrou A, Georghiou G E. Wavelet packet de-noising for online partial discharge detection in cables and its application to experimental field results. Measurement Science and Technology, 2006, 17(9): 2367–2379
4. Ambikairajah R, Phung B T, Ravishankar J, Blackburn T R, Liu Z. Smart sensors and online condition monitoring of high voltage cables for the smart grid. In: Proceedings of the Fourteenth International Middle East Power Systems Conference (MEPCON). 2010, 807–811
5. Ambikairajah R, Phung B T, Ravishankar J, Blackburn T R, Liu Z. Novel frequency domain features for the pattern classification of partial discharge signals. In: Proceedings of the XVII International Symposium on High Voltage Engineering. 2011, Paper F-044
6. Kuncheva L, Bezdek J C, Duin R P W. Decision templates for multiple classifier fusion: an experimental comparison. Pattern Recognition, 2001, 34(2): 299–314
7. Evagorou D, Kyprianou A, Lewin P L, Stavrou A, Efthymiou V, Metaxas A C, Georghiou G E. Feature extraction of partial discharge signals using the wavelet packet transform and classification with a probabilistic neural network. IET Science, Measurement and Technology, 2010, 4(3): 177–192
8. Ma X, Zhou C, Kemp I J. Interpretation of wavelet analysis and its application in partial discharge detection. IEEE Transactions on Dielectrics and Electrical Insulation, 2002, 9(3): 446–457
9. Wright J, Yang A Y, Ganesh A, Sastry S S, Ma Y. Robust face recognition via sparse representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(2): 210–227
10. Elad M. Sparse and Redundant Representations. New York, NY: Springer, 2009
11. Mohimani G H, Babaie-Zadeh M, Jutten C. Fast sparse representation based on smoothed L0 norm. In: Proceedings of the Seventh International Conference on Independent Component Analysis and Signal Separation. 2007, 389–396
12. Kanevsky D, Sainath T N, Ramabhadran B, Nahamoo D. An analysis of sparseness and regularization in exemplar-based methods for speech classification. In: Proceedings of Interspeech 2010. 2010, 2842–2845