Bayesian Neural Network and Discrete Wavelet Transform for Partial Discharge Pattern Classification in High Voltage Equipment

Hui Ma, *Member*, IEEE, Jeffery C. Chan, *Student Member*, IEEE, Tapan K. Saha, *Senior Member*, IEEE
School of Information Technology & Electrical Engineering
The University of Queensland
Brisbane, Australia
huima@itee.uq.edu.au, j.chan@uqconnect.edu.au, saha@itee.uq.edu.au

***Abstract*— Partial discharge (PD) pattern recognition has been applied for identifying the types of insulation defects in high voltage (HV) equipment. It can provide an effective means for condition assessment of the insulation system of HV equipment. This paper proposes a novel Bayesian neural network (BNN) and discrete wavelet transform (DWT) hybrid algorithm for PD pattern recognition. Laboratory experiments on a number of PD models have been conducted for evaluating the performance of the proposed algorithm.**

***Index Terms*-- Bayesian neural network (BNN), discrete wavelet transform (DWT), Partial Discharge (PD), and pattern recognition.**

I. INTRODUCTION

Partial discharge (PD) measurement has been widely adopted for monitoring and diagnosis of high voltage (HV) equipment [1]-[2]. One of the major tasks of PD measurement is the PD pattern recognition for identifying the types of defects that cause discharges in HV equipment. Over the past two decades, a number of intelligent techniques have been developed for automatic PD pattern recognition. Some examples include statistical methods, various artificial neural networks (ANNs), genetic algorithms, expert systems, discrete wavelet transforms, and support vector machines (SVMs) [3]-[10].

However, it is still a non-trivial task to apply the intelligent algorithms for automatically recognizing various types of defects in the insulation system of HV equipment. The two most challenging issues are: (1) extracting representative features from PD measurement data while maintaining lower dimensionality; and (2) choosing appropriate algorithms to attain desirable performance in classifying various PD patterns due to different defects.

This paper proposes a novel algorithm of integrating Bayesian neural network (BNN) and discrete wavelet transform (DWT) for PD pattern recognition. BNN provides a probabilistic treatment of leaning in neural network. Instead of only considering a single set of optimal network parameters (i.e. weights), BNN exploits an entire probability distribution of these parameters and can naturally address the issue of regularization to avoid over-fitting [11], [12].

The original PD measurement data is extremely high dimensional. To deal with this high dimensionality, the traditional approach computes a number of statistic operators on the discharge pulse height and number distributions [3], [4]. These statistic operators form a feature set to represent PD patterns. However, this paper adopts discrete wavelet transform (DWT) for feature extraction. The benefit of using DWT in that it can integrate PD signal de-nosing and feature extraction in a single step. It is expected that the BNN and DWT hybrid algorithm can attain desirable recognition accuracy compared to a number of other algorithms. This will be verified using a PD dataset obtained from laboratory experiments on a variety of artificial PD models.

The remaining part of this paper is organized as follows. Section II describes PD experiments set-up and data acquisition procedure. Section III provides a brief review on DWT approach for feature extraction. A data visualization algorithm is also presented in this section. Section IV details the formulation of BNN. Section V presents the recognition results of BNN and DWT hybrid algorithm as well as a number of other algorithms. Section VI concludes the paper.

II. EXPERIMENT SET-UP AND DATA ACQUISITION

In this paper, a number of artificial PD models are constructed to generate PD dataset for evaluating PD pattern recognition algorithms. These models include: corona, discharge in transformer oil, surface discharge, internal discharge, and discharge due to floating particles (Figure 1). Omicron's MPD600 (complying with IEC60270) was adopted for PD data acquisition. Figure 2 shows the typical three dimensional (3D) phase resolved diagram for each of the above five PD models. The phase resolved diagram depicts the possible correlations among discharge pulse number, discharge pulse magnitude, and the phase angle of the applied AC voltage.

This work was supported by the Australia Research Council (ARC) on Linkage Grant.

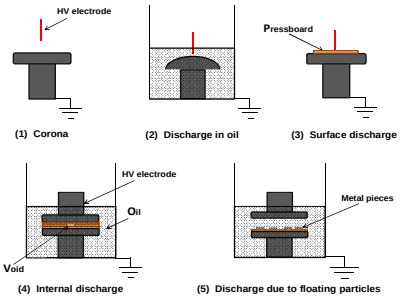

Figure 1. Artificial PD models

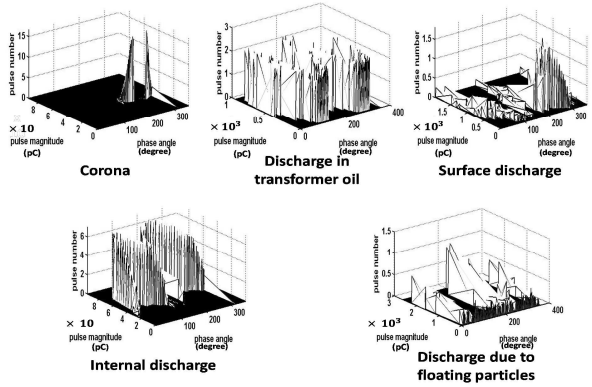

Figure 2. Typical PD patterns of different PD models

For each of the above five PD models, PD measurements were conducted under three different applied voltage levels and with three different noise gating thresholds. At one acquisition, PD pulses of 100 power cycles were recorded. For each of five PD models, 200 acquisitions were obtained. The resultant dataset consists of total 1000 data points. The discrete wavelet transform (DWT) will extract features based on this dataset and construct a new dataset for training and validating Bayesian neural network (BNN).

III. Feature Extraction and Data Visulization

A. *Discrete Wavelet Transform (DWT) Approach*

Discrete wavelet transform (DWT) has been widely adopted as a signal de-noising tool in PD measurement [4]. Recently it has also been applied for extracting representative features of different PD patterns corresponding to various insulation defects in HV equipment [9].

In DWT, the original PD signals are decomposed into a number of approximation and detail coefficients through a series of low pass and high pass filters [13]. Both coefficients are then down-sampled and the approximation coefficients will be further decomposed until reaching a predefined decomposition level. In this paper, the original PD signals are decomposed into nine levels by using the abior1.5 wavelet. Thus, each discharge pulse is represented by nine coefficients. Given a considerable large number of PD pulses obtained in each PD signal acquisition (about several thousand discharge pulses in one acquisition in the experiments), the above DWT decomposition process will introduce considerable high dimensionality.

To combat the high dimensionality, this paper computes the first four moment statistics including mean $\bar{x}$, standard deviation σ, skewness s, and kurtosis δ for each of the nine distributions composed by the detail coefficients:

$$\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x(n) \tag{1}$$

$$\sigma = \left(\frac{1}{N}\sum_{n=1}^{N}[x(n) - \bar{x}]^2\right)^{1/2} \tag{2}$$

$$s = \frac{1}{\sigma^3}\sum_{n=1}^{N}[x(n) - \bar{x}]^3 \tag{3}$$

$$\delta = \frac{1}{\sigma^4}\sum_{n=1}^{N}[x(n) - \bar{x}]^4 \tag{4}$$

where $x(n)$ is the wavelet coefficient at location n and N is the total number of wavelet coefficients at each level. Finally, total 36 features are extracted from the original PD dataset.

B. *PD Data Visualization*

To provide data visualization for high dimensional data, the NeuroScale is adopted in this paper. NeuroScale projects the data points in the original space into a two dimensional space, in which the data points that are close in the original space are kept close while the data points that are significantly separated in the original space are remained well-separated [14]. This is achieved by minimizing the following function E

$$E = \sum_{i}^{N}\sum_{j>1}^{N}\left(d'_{ij} - d_{ij}\right)^2 \tag{5}$$

where d'_{ij} is the Euclidean distance between data point i and j in the original space, and d_{ij} is the Euclidean distance between corresponding data points in the projection space. N is the total number of data points in the dataset. An approach proposed by Lowe and Tipping can be used for solving the above minimization problem [14].

IV. Theory of Bayesian Neural Network

In this paper, the Bayesian neural network (BNN) is constructed by applying Bayesian approach on the conventional multi-layer perceptron (MLP). Unlike classical neural network, in which a single set of network parameters (weights) are sought using maximum likelihood method, the BNN approach considers a probability distribution function over the distribution of these network parameters [11], [12]. BNN can effectively solve the over-fitting problem through the control of model complexity and naturally handle the uncertainties through probabilistic modelling.

Starting with a brief review of MLP, this section will present the Bayesian approach on finding the optimal network parameters of MLP. It is assumed the PD data is an N x d dataset, $\mathbf{X} = [\,\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_N\,]$, where N is the number of data points and d is the size of features of each data point, i.e. $\mathbf{x}_i = \left[x_i^{\mathbf{1}}, \ldots, x_i^{\boldsymbol{d}}\right]$. Moreover, each data point $\mathbf{X}$ belongs to one of the T independent classes, i.e. $C \in \{1, \ldots, t_k, \ldots, T\}$, where each class corresponds to one type of PD models depicted in Figure 1.

A. *Multi-layer Perceptron (MLP)*

Figure 3 depicts the structure of three layers MLP. The first layer is the input layer, which is a set of discriminative features describing the characteristics of different PD patterns. The third layer is the output layer, which includes five different types of PD models. The middle layer is the hidden layer, and the nodes in this layer are connected to all nodes in the input layer and output layer. Each connection carries a weight.

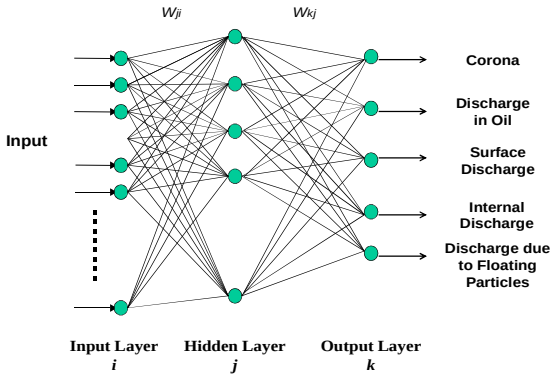

Figure 3. Structure of multi-layer perceptron (MLP)

In the input layer, M linear combinations of inputs are formed to obtain a set of variables associated with hidden nodes as follows:

$$\alpha_i^m = \sum_{j=1}^{d} w_i^{mj} x_i^j + b_i^m \quad i = 1, \dots, N;\ m = 1, \dots, M \qquad (6)$$

where x_i^j is the j-th element of the i-th data point $\mathbf{x}_i$, w_i^{mj} is the weight element regarding the m-th hidden node to the j-th element of data point $\mathbf{x}_i$, b_i^m is the bias term, and M is the total number of hidden nodes. The outputs of the above hidden nodes are determined by a tanh activation function:

$$z_i^m = tanh(\alpha_i^m) \qquad m = 1, \dots, M \qquad (7)$$

The variables z_i^m are transformed by the second sets of weights regarding the k-th output to the m-th hidden node as

$$a_i^k = \sum_{m=1}^{M} w_i^{km} z_i^m + b_i^k \quad i = 1, \dots, N;\ k = 1, \dots, T \qquad (8)$$

Finally these values go through an output activation function to give output value y_k. For the two-class pattern recognition problem, the activation function normally adopts the logistic sigmoid function

$$y_k = g(a) = \frac{1}{1+\exp(-a_i^k)} \qquad (9)$$

For the multi-class pattern recognition problem such as the PD pattern recognition (five PD models in this paper), the activation function normally adopts the softmax function:

$$y_k = g(a) = \frac{\exp(a_i^k)}{\sum_{k'} \exp(a_i^{k'})} \qquad (10)$$

Traditionally, the above MLP network is trained through the back-propagation technique to find an optimal set of values for network weights. It is necessary to evaluate the derivative of an error function, which is the sum of square error between network output y_k and the true class label t_k with respect to the network weights [11].

B. *Bayesian Approach for Multi-layer Perceptron (MLP)*

(1) *Bayesian learning of network weights*

Unlike conventional approach, the Bayesian approach exploits the probability distribution function of network weighs to represent the relative degrees of belief on different values for these weights. By using Bayes' theorem, the posterior probability distribution function of weights with respect to the class is

$$p(\mathbf{w}|C) = \frac{p(C|\mathbf{w})p(\mathbf{w})}{\int p(C|\mathbf{w})p(\mathbf{w})} \qquad (11)$$

where $p(\mathbf{w})$ is the prior probability distribution, $p(C|\mathbf{w})$ is the likelihood function, and $\int p(C|\mathbf{w})p(\mathbf{w})$ is the normalization factor. Upon receiving training data, the posterior probability distribution $p(\mathbf{w}|C)$ can be evaluated.

The prior probability distribution for weights can be chosen as a Gaussian prior with zero means [11]:

$$p(\mathbf{w}) = \frac{1}{Z_w(\alpha)} \exp\left(-\frac{\alpha}{2}\|\mathbf{w}\|^2\right) = \frac{1}{Z_w(\alpha)} \exp(-\alpha E_w) \qquad (12)$$

where α is the inverse variance of the distribution, $Z_w(\alpha)$ is a normalization constant, and the term $\frac{\alpha}{2}\|\mathbf{w}\|^2 = \frac{\alpha}{2}\sum w_i^2$ is equivalent to the weight decay regulation term in the conventional MLP. α is also called the hyperparameter as it is a parameter for the distribution of other parameters. The calculation of α will be discussed later in this section.

For the easier demonstration of key ideas behind the BNN algorithm, two-class classification problem ($t_k = 0$ or 1) is adopted as an example for deriving the likelihood function $p(C|\mathbf{w})$ and the posterior probability distribution $p(\mathbf{w}|C)$. The likelihood function $p(C|\mathbf{w})$ is then in the form of

$$p(C|\mathbf{w}) = \prod_N y(\mathbf{x}_k)^{t_k} \left(1 - y(\mathbf{x}_k)\right)^{1-t_k} = exp(-G) \qquad (13)$$

where

$$G = -\sum_n \{t_k \ln y(\mathbf{x}_k) + (1 - t_k)\ln\left(1 - y(\mathbf{x}_k)\right)\} \qquad (14)$$

By combing Equation (11)-(14), the posterior distribution of weights can be expressed as follows:

$$p(\mathbf{w}|C) = \frac{1}{Z_s}\exp(-G - \alpha E_W) = \frac{1}{Z_s}\exp\left(-S(\mathbf{w})\right) \qquad (15)$$

$S(\mathbf{w}) = G + \alpha E_W$ is the overall error function. The above posterior distribution $p(\mathbf{w}|C)$ can be further approximated by a Gaussian centered on the maximum posterior weight vector $\mathbf{w}_{\mathrm{MP}}$:

$$p(\mathbf{w}|C) = \frac{1}{Z_S^*} \exp\left[-S(\mathbf{w}_{\mathrm{MP}}) - \frac{1}{2}(\mathbf{w} - \mathbf{w}_{\mathrm{MP}})^{\mathrm{T}}\mathbf{A}(\mathbf{w} - \mathbf{w}_{\mathrm{MP}})\right] \qquad (16)$$

where Z_S^* is the normalization constant appropriating to the Gaussian approximation.

(2) Probability distribution of network output

As discussed in the above section, the "trained" MLP network can be represented by the posterior probability distribution of network weights. Upon receiving a new input $\mathbf{x}^*$, the trained network needs to classify it into one of T classes. The probability distribution of network output for the input $\mathbf{x}^*$ can be written as

$$p(t|\mathbf{x}^*, C) = \int p(t|\mathbf{x}^*, \mathbf{w}) p(\mathbf{w}|C)\, d\mathbf{w} \tag{17}$$

where $p(t|\mathbf{x}^*, \mathbf{w}) = y(\mathbf{x}^*; \mathbf{w})$ is the output function (Equation 9 for two-class recognition problem). Since $y(\mathbf{x}^*; \mathbf{w})$ is nonlinear, it is inappropriate to use $y(\mathbf{x}^*; \mathbf{w}_{\mathrm{MP}})$ to approximate $y(\mathbf{x}^*; \mathbf{w})$ as in Equation 16.

To evaluate $y(\mathbf{x}^*; \mathbf{w})$, Mackay introduced a locally linear function of the weights as [12]:

$$a(\mathbf{x}^*; \mathbf{w}) = a_{\mathrm{MP}}(\mathrm{x}) + \mathbf{g}^{\mathrm{T}}(\mathbf{x}^*)(\mathbf{w} - \mathrm{w}_{\mathrm{MP}}) \tag{18}$$

And its probability distribution can be derived as

$$p(a|\mathbf{x}^*, C) = \frac{1}{(2\pi s^2)^{1/2}} \exp\left(-\frac{(a - a_{\mathrm{MP}})^2}{2s^2}\right) \tag{19}$$

where the variance s^2 is given by

$$s^2(\mathbf{x}^*) = \mathbf{g}^{\mathrm{T}} \mathbf{A}^{-1} \mathbf{g} \tag{20}$$

where $\mathbf{A} = \nabla\nabla S|\mathrm{w}_{\mathrm{MP}}$ is the Hessian matrix of the overall error function $S(\mathbf{w})$, $\mathbf{g} = \nabla y|\mathrm{w}_{\mathrm{MP}}$ is the gradient. The value of a_{MP} is obtained by forwarding propagation $\mathbf{x}^*$ through the network with weights w_{MP} and variance s^2. Then the probability distribution of output becomes

$$p(t|\mathbf{x}^*, C) = \int p(t|a) p(a|\mathbf{x}^*, C)\, da \tag{21}$$

where $p(t|a) = g(a)$ is the output function (Equation 9), and $p(a|\mathbf{x}^*, C)$ is given by Equation 19. The following approximation can be used to the above integral in Equation 21 [12]:

$$p(t|\mathbf{x}^*, C) \approx g(\vartheta(s) a_{\mathrm{MP}}) \tag{22}$$

where

$$\vartheta(s) = \left(1 + \frac{\pi s^2}{8}\right)^{-1/2} \tag{23}$$

(3) Iterative process for calculating hyperparameter α

The hyperparameter α can be computed together with the network weights through the following iterations [11],[12]:

(1) Initially, the hyperparameter α is set to a small arbitrary value. Then the network is trained to find the maximum weight vector $\mathbf{w}_{\mathrm{MP}}$ by minimizing the cost function $S(\mathbf{w})$. Training is stopped when the training error falls below a pre-specified value.

(2) The hyperparameter α is updated to

$$\alpha_{new} = \frac{\gamma}{2E_W} \tag{24}$$

where γ is calculated using the value of α at the previous iteration, and it is in the form of

$$\gamma = \sum_{i=1}^{W} \frac{\lambda_i}{\lambda_i + \alpha} \tag{25}$$

λ_i are the eigenvalues of the data Hessian matrix, refer to Equation 20.

(3) Once the hyperparameter α has been updated, the network is trained again from where it halted until a specified lower training error value is attained. The above training process continues until convergence.

V. Results and Analysis

This section presents the results of using the hybrid BNN and DWT algorithm for PD pattern recognition. For the purpose of comparisons, the results of using three commonly adopted pattern recognition algorithms namely *k*-nearest neighbor (*K*NN), radial basis function (RBF) network, and multi-layer perceptron (MLP) are also included in this section. And one more feature extraction approach, the statistic operator is also included in this section. The statistic operator approach forms a feature set of 24 statistic parameters to quantify discharge pulse maximum magnitude distribution, discharge pulse average magnitude distribution, and discharge pulse number distribution with respect to the phase angle of the applied AC voltage [3].

By integrating BNN, *K*NN, RBF, and MLP with DWT and statistic operator, total eight algorithms are implemented in this section. These eight algorithms are: BNN-DWT, *K*NN-DWT, MLP-DWT, RBF-DWT, BNN-Stat, *K*NN-Stat, MLP-Stat, and RBF-Stat (Stat refer to statistic operator).

As mention in Section II, 200 acquisitions (each consists of PD pulses in 100 power cycles) are obtained for each of the five PD models depicted in Figure 1. Thus, the dataset is made up of 1000 data points belonging to five different PD classes. This original dataset will be fed into either DWT or statistic operator for constructing the dimension reduced dataset. The dimension reduced dataset constructed by DWT has the dimension of 1000 x 36 while that constructed by statistic operator has the dimension of 1000 x 24.

Each of the above two dimension-reduced datasets is randomly divided into two parts: a training dataset that comprises 70% samples and a testing dataset that comprises 30% samples. For each pattern recognition algorithms, the optimal values of some parameters need to be found. These include: the number of neighbors in *K*NN, and the number of hidden nodes in RBF, MLP, and BNN. To decide the optimal values of these parameters, ten-fold cross validation is performed on each algorithm using the above training dataset. Once the best parameters are found, the algorithm will be trained with its optimal parameters. Finally, each trained algorithm is presented with the testing dataset and tasked to make recognition of the types of PD sources (PD models) for the samples (data points) in the testing dataset.

The class splitting, tenfold cross validation, and testing are repeated 20 times for each algorithm. Figure 4 shows the NeuroScale visualization with the 2D projection of original training and testing datasets of the five PD models (classes).

Table 1 presents the results of eight algorithms including overall recognition rate (in percentage) and the recognition rate (in percentage) with respect to each type of PD models.

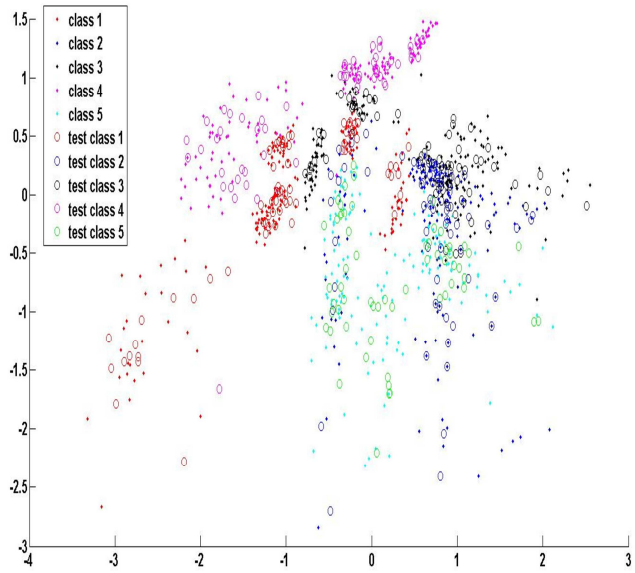

Figure 4. NeuroScale visualization

Note: Class 1- corona; Class 2 – discharge in oil; Class 3 – surface discharge; Class 4- internal discharge; Class 5 – discharge due to floating particle.

Table 1 Recognition rates (in percentage) of various algorithms (averaged over 20 trials)

Algorithm	Overall Rate	Class 1 Rate	Class 2 Rate	Class 3 Rate	Class 4 Rate	Class 5 Rate
*K*NN-Stat	90.8	99.4	83.8	86.7	99.5	85.2
RBF-Stat	92.5	98.5	87.4	90.1	97.8	89.2
MLP-Stat	92.0	98.1	84.8	91.6	99.1	86.6
BNN-Stat	93.8	98.3	88.8	91.4	99.0	92.3
*K*NN-DWT	94.2	99.2	91.4	96.0	98.3	85.4
RBF-DWT	97.3	98.8	97.5	97.2	98.3	94.3
MLP-DWT	98.5	97.9	99.6	99.4	98.2	97.2
BNN -DWT	**99.1**	**99.4**	**98.8**	**98.5**	**99.8**	**98.8**

From Figure 4, it can be seen that the data points of one class mixed up with those of other classes in both training dataset and testing dataset. Especially, the data points belonging to class 2 (in blue color, PD model of discharge in oil) and class 5 (in green color, PD model of discharge due to floating particles) scatter in the data space and blend into each other. This may cause difficulties for some algorithms to make explicit recognition and correctly classify all data points in the testing dataset into their corresponding types of PD models (classes). This can also explains why the recognition rates of *K*NN-Stat, *K*NN-DWT, MLP-Stat, RBF-Stat on Class 2 or Class 5 are relatively low as shown in Table 1.

Table 1 reveals that BNN outperforms *K*NN, MLP, and RBF in both cases of integration with DWT and integration with statistic operator. It can also be seen from Table 1 that the algorithms integrated with DWT attained higher recognition rate than the pattern recognition algorithms integrated with statistic operator. It can also be observed that the proposed hybrid of BNN and DWT achieves the highest recognition rate amongst the all eight algorithms.

VI. Conclusions

This paper proposed a novel Bayesian neural network (BNN) and discrete wavelet transform (DWT) hybrid algorithm for PD pattern recognition. Laboratory experiments on different PD models were conducted for evaluating the proposed algorithm. Results show that the proposed hybrid algorithm can consistently attain higher recognition rate. Future research will extend the hybrid BNN and DWT algorithm for multiple sources PD pattern recognition by conducting laboratory experiments and field measurements.

Acknowledgment

The authors gratefully acknowledge Australian Research Council, Powerlink Queensland, Energex, Ergon Energy, and TransGrid for providing supports for this work.

References

[1] R.Bartnikas, "Partial Discharges: Their Mechanism, Detection and Measurement," *IEEE Trans. Dielectric and Electrical Insulation*, Vol. 9, pp. 763-808, 2002.

[2] IEC International Standard 60270, High-Voltage Test Techniques-Partial Discharge Measurements, IEC, 3rd edition, 2000.

[3] E.Gulski, "Digital Analysis of Partial Discharges," *IEEE Trans. Dielectric and Electrical Insulation*, Vol. 2, pp. 822-837, 1995.

[4] N.C.Sahoo, M.M.A.Salama, and R.Bartnikas, "Trends in Partial Discharge Pattern Classification: A Survey," *IEEE Trans. Dielectric and Electrical Insulation*, Vol. 12, pp. 248-264, 2005.

[5] C.Mazzetti, F.M.F.Mascioli, F.Baldini, M.Panella, R.Risica, and R. Bartnikas, "Partial Discharge Pattern Recognition by Neuro-Fuzzy Networks in Heat-Shrinkable Joints and Terminations of XLPE Insulated Distribution Cables," *IEEE Trans. Power Delivery*, Vol. 21, No.3, pp. 1035-1044, 2006.

[6] A.Rizzi, F.Mascioli, F.Baldini, C.Mazzetti, and R.Bartnikas, "Genetic Optimization of a PD Diagnostic System for Cable Accessories," *IEEE Trans. Power Delivery*, Vol. 24, No.3, pp. 1728-1738, 2009.

[7] S.M.Strachan, S.Rudd, S.D.J.McArthur, and M.D.Judd, "Knowledge-Based Diagnosis of Partial Discharges in Power Transformers," *IEEE Trans. Dielectric and Electrical Insulation*, Vol. 15, pp. 259-268, 2008.

[8] T.Babnik, R.K.Aggarwal, and P.J.Moore, "Principal Component and Hierarchical Cluster Analyses as Applied to Transformer Partial Discharge Data With Particular Reference to Transformer Condition Monitoring," *IEEE Trans. Power Delivery*, Vol. 23, No.4, pp. 2008-2016, 2008.

[9] D.Evagorou, A.Kyprianou, P.L.Lewin, A.Stavrou, V.Efthymiou, A.C.Metaxas, and G.E.Georghiou, "Feature extraction of Partial Discharge Signals Using the Wavelet Packet Transformer and Classification with a Probabilistic Neural Network", *IET Sci. Meas. Technol.*, Vol.4, pp. 177-192, 2010.

[10] L.Hao and P.L.Lewin, "Partial Discharge Source Discrimination Using a Support Vector Machine," *IEEE Trans. Dielectric and Electrical Insulation*, Vol. 17, pp. 189- 197, 2010.

[11] C.M.Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[12] I.T.Nabney, *NETLAB: Algorithms for Pattern Recognition*, Springer, 2004.

[13] D.Salomon, G.Motta, and D.Bryant, *Data compression: the complete reference (4th edition)*, Springer, 2007.

[14] D. Lowe and M. E. Tipping, "Neuroscale: Novel topographic feature extraction with radial basis function networks," *Advances in Neural Information Processing Systems*, vol. 9, pp. 543-549, Cambridge, MIT Press, 1997.

[15] P.Lauret, E.Fock, R.N.Randrianarivony, J.-F Manicom-Ramsamy, "Bayesian neural network approach to short time load forecasting," Energy Conversion and Management, Vol. 49, pp.1156–1166, 2008.