OODCN: Out-Of-Distribution Detection in Capsule Networks for Fault Identification

Imene Mitiche*, Alireza Salimy*
Falk Werner †, Philip Boreham†, Alan Nesbitt* and Gordon Morison*
* School of Engineering and Built Environment
Glasgow Caledonian University,
Glasgow, United Kingdom
† Innovation Centre for Online Systems
Doble Engineering Company
Bere Regis, United Kingdom

***Abstract*—In order to aid survey engineers identify Partial Discharge (PD) types during their asset diagnostics, we develop an image-based system for PD signal classification and Out-Of-Distribution (OOD) rejection. First, the PD signal is converted to a Phase-Resolved PD (PRPD) image. Then, the image is passed to the system which exploits a Capsule Network in an auto-encoder framework, where the encoder output is used for PD classification and the decoder output is used in the OOD decision. The latter is the main contribution of this work which combines the decoder part with a reconstruction metric evaluating the difference between the original and reconstructed image. A threshold for OOD decision is introduced based on the distribution of reconstruction values from the training data. Most importantly, the OOD data is not exposed to the model during training. Results demonstrate high performance in both PD types classification and OOD detection tasks using synthetic and real data.**

I. Introduction

Early Partial Discharge (PD) detection in High-Voltage (HV) rotating machines is essential to maintain healthy machinery and consistent operation. On-line PD testing is one of the tools used in condition monitoring to prevent high costs associated with equipment failure and repair. PD is an electrical discharge that occurs due to insulation breakdown or deterioration of, for example, stator windings in generators. This is generally caused by various factors including thermal or electrical stress, improper installation and insulation aging [1]. The collected PD measurement can be represented in two main forms: time-resolved and Phase-Resolved PD (PRPD). The latter is widely used in PD diagnosis as it provides information on the PD location and illustrates different patterns for PD types [2]. Survey engineers rely on the analysis and visualisation of PRPD patterns to identify PD presence in generators and its type. However, this process is manual and not practical for continuous diagnostic. Recent advances in Machine Learning (ML) combined with big data collection permit the move towards automated diagnosis that aids the engineers in their survey and provide more information to power station owners on the fault allowing precautionary measures to be taken.

For a ML model to be successful when deployed in real-world problems, the model is required to be able to distinguish data that is similar to the training data from anomalous or significantly different data. This is known as Out-Of-Distribution (OOD) detection [3]. A successful OOD rejection avoids classifying OOD data into the in-distribution classes leading to false alarms.

In this work we propose a model that satisfies the following: 1) Identification and classification of PD types in rotating machines using PRPD image as input to the Capsule Network (CapsNet). 2) Rejection of OOD data to prevent false classification of data that is not similar to the PD types used in training the CapsNet model.

CapsNet is an alternative to traditional Convolutional Neural Networks (CNNs) designed to mimic the biological neurons functionality, and has been demonstrated to perform well in classification problems [4]. The design is based on human vision and its capability to ignore irrelevant details [5]. This makes CapsNet a suitable choice for this work, since we attempt to imitate the engineer's visual analysis of PRPD images. CapsNet provides the ability to communicate through layers while holding information about spatial relationships between features. This is achieved by replacing the pooling functions by a routing algorithm [6]. We exploit the image reconstruction part of the CapsNet model to form our proposed OOD detection method. The reconstruction part is another motivation to choosing CapsNet over the popular CNN based models for image classification tasks such as VGGNet [7], ResNet [8], MobileNet [9] etc. Data detected as OOD by our system can be examined by an experienced engineer and recycled in the model training if identified by the engineers as a new PD type.

PRPD data has been utilised in two main ways with ML classification, statistical feature extraction from the data or image-based feature extraction from the PRPD image. The common statistical measures include mean, variance, skewness, kurtosis and cross-correlation [10] [11]. Image based features extraction tools have been applied to PRPD image such as texture and fractal analysis [12], wavelet image decomposition [13]. Feature extraction is very useful as it reduces data dimension and extracts the relevant information at the same time [14], however this requires knowledge of the

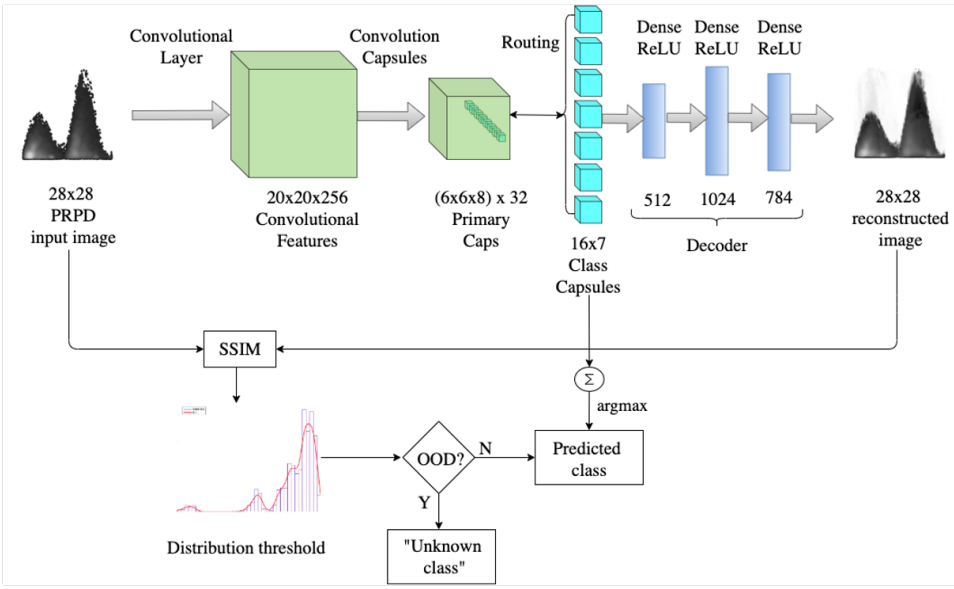

Fig. 1. The proposed classification approach.

relevant feature extraction method to employ. Deep Learning Networks have the ability to extract relevant features through the layers during the learning [15]. Various deep learning architectures were exploited in PRPD classification including Deep Neural Networks [16], Recurrent Neural Networks [17], Auto-Encoders (AE), Deep Belief Networks [18], [19] and a combination of CNN and AE [20].

To the best of our knowledge, CapsNets have not been applied to PRPD images for PD types classification in rotating machines. Furthermore, the proposed OOD detection algorithm combined with CapsNet AE framework has not been utilised in the literature.

The remainder of this paper is organised as follows. Section II defines PRPD patterns along with the proposed framework and the algorithms involved in the analysis. Section III describes the synthetic and real-world data used in training/testing the model. Results of the proposed method are also presented in this section. Section IV concludes this work along with future work.

II. PRPD Pattern Recognition

The proposed approach for the classification of PD types in HV rotating machines is summarised in Fig. 1. First, the PRPD data is represented in a density scatter plot, which is considered as an image. The latter is converted to grey-scale and down-sampled at a lower resolution (28×28) for memory and computation purposes. The prepared image is passed to the CapsNet model for classification and image reconstruction by taking the class capsules output to a decoder network consisting of three fully connected (Dense) layers with ReLU activation. The Structural Similarity Index Measure (SSIM) is calculated between the original and reconstructed images and thresholded for OOD decision. In this section we denote scalars by lower or upper case, vectors by bold lower case and matrices by bold upper case.

A. Phase-Resolved Partial Discharge (PRPD) Pattern

A PRPD plot is a visual representation of PD activity with respect to the AC power cycle where the magnitude is distributed across the $360°$ phase. The PD pulses are acquired based on the AC signal's phase angle (ϕ), charge magnitude (q) and the number of PD pulses (n) over a predetermined time duration [21]. Fig. 3 illustrates two examples of PD types occurring in rotating machines.

B. Capsule Network (CapsNet)

The main building blocks of CapsNet in [5] are convolution, primary capsules, routing process, class capsules and the decoder. These are described in detail as follows.

1) Convolution: Convolutional layers are used to extract feature maps from the input. Equation (1) represents the convolutional operation, where $\mathbf{X}$ is the input, the PRPD image in this work. $\mathbf{K}$ is the kernel filter with size $k \times k$, $n = 0, 1, ..., N-1$ and $m = 0, 1, ..., L-1$ with L and N being the length and width of the image respectively. C represents the number of channels in the input. A bias B is then added and passed to an activation function $f\big(\cdot\big)$ to produce the layer's final output.

$$(\mathbf{X} * \mathbf{K})_{m,n} = f\Big(\sum_{i=0}^{k-1}\sum_{j=0}^{k-1}\sum_{h=1}^{C} \mathbf{X}(i+m, j+n, h)\mathbf{K}(i,j,h) + B\Big) \quad (1)$$

The CapsNet architecture uses convolution in the first convolutional layer and in convolution capsules.

2) Primary Caps: we refer to $\mathbf{s}_a$ as the capsule in the present layer and $\mathbf{s}_b$ as the capsule in the next layer. The latter is obtained using (2) by calculating the weighted sum of all prediction vectors $\hat{\mathbf{u}}_{b|a}$ and a coupling co-efficient c_{ab} between a and b. The prediction vectors are calculated using the capsule output in the previous layer $\mathbf{u}_a$ and a transformation matrix $\mathbf{W}_{ab}$, as shown in (3).

$$\mathbf{s}_b = \sum_a c_{ab} \hat{\mathbf{u}}_{b|a} \tag{2}$$

$$\hat{\mathbf{u}}_{b|a} = \mathbf{W}_{ab} \mathbf{u}_a \tag{3}$$

The coupling coefficients sum to 1 and are refined by the routing algorithm which is 'dynamic routing' [5] in this paper. They are obtained by a Softmax of the logits g_{ab}, equivalent to the log prior probabilities that capsule a should be coupled to capsule b, this relationship is illustrated in (4) where d is the class capsule. The number of coefficients is equal to the number of the current layer Capsules.

$$c_{ab} = \frac{\exp(g_{ab})}{\sum_d \exp(g_{ad})} \tag{4}$$

The original coupling coefficients are updated in every iteration by measuring the agreement between the current capsule b output $\mathbf{v}_b$ and the prediction vector $\hat{\mathbf{u}}_{b|a}$ from capsule a. This agreement is the scalar product of $\mathbf{v}_b$ and $\hat{\mathbf{u}}_{b|a}$, which is added to the logits g_{ab}, outlined in (5).

$$g_{ab} \longleftarrow g_{ab} + \hat{\mathbf{u}}_{b|a} \cdot \mathbf{v}_b \tag{5}$$

Capsules attempt to represent the probability of an entity's presence in the input through the magnitude of an output vector in the range $0 \rightarrow 1$. A vector with a magnitude close to 1 correspond to high probability of an entity's presence in the input, and therefore $\mathbf{s}_b$ agrees with $\mathbf{s}_a$. In contrast, those capsules disagree when a vector with a value close to 0 correspond to an entity's absence in the input. This criterion is satisfied by a 'Squash' operation on s_b, to obtain the output vector $\mathbf{v}_b$ of capsule b as illustrated in (6). This is performed for all capsules in CapsNet.

$$\mathbf{v}_b = \frac{\|\mathbf{s}_b\|^2}{1 + \|\mathbf{s}_b\|^2} \frac{\mathbf{s}_b}{\|\mathbf{s}_b\|} \tag{6}$$

3) Dynamic Routing: Tying this section to II-B2, dynamic routing is the algorithm during which the transformation matrix weights $\mathbf{W}_{ab}$, coupling coefficients C_{ab}, logits are trained/updated and the agreement between capsules occurs. The overall algorithm for dynamic routing is presented in procedure 1 consisting of an inner iteration (3-7) in the main iteration (2-end).

4) CapsNet loss and Reconstruction: two loss functions are implemented in this paper. The first one is the margin loss used in classification. The second one is the Mean Squared Error (MSE) loss used in the image reconstruction part. The margin loss is computed as:

Procedure 1 Dynamic routing

1: **procedure** ROUTING($\hat{\mathbf{u}}_{b|a}, r, l$)
2: for all capsules a in layer l and capsule b in layer $l+1$: $g_{ab} \leftarrow 0$
3: **for** r iterations **do**
4: for all capsule a in layer l: $c_a \leftarrow \text{softmax}(g_a)$
5: for all capsule b in layer $(l+1)$: $\mathbf{s}_b \leftarrow \sum_a c_{ab} \hat{\mathbf{u}}_{b|a}$
6: for all capsule b in layer $(l+1)$: $\mathbf{v}_b \leftarrow \text{squash}(\mathbf{s}_b)$
7: for all capsule a in layer l and capsule b in layer $(l+1)$: $g_{ab} \leftarrow g_{ab} + \hat{\mathbf{u}}_{b|a}.\mathbf{v}_b$
return $\mathbf{v}_b$

$$\mathcal{L} = \sum_{d=1}^{D} \mathbf{t}_d \max(0, p^+ - \|\mathbf{v}_d\|)^2 + \lambda(1 - \mathbf{t}_d) \max(0, \|\mathbf{v}_d\| - p^-)^2 \tag{7}$$

where D is the total number of classes, λ is a constant used for numerical stability and is set to $\lambda = 0.5$ along with the parameters $p^+ = 0.9$ and $p^- = 0.1$ set as recommended in [5] in order to prevent the vector length from reaching the max or collapsing. The λ value is used as a down-weighting in order to avoid the initial learning from shrinking the lengths of the class capsules vectors. The true label $\mathbf{t}_d$ is equal to 1 when an entity of class d is present and it is equal to 0 otherwise.

The MSE reconstruction loss [22] is implemented to fine tune the encoding of the input class originating from the correct class capsule. The output of the latter is passed to the decoder part. The MSE is calculated between the original input image $\mathbf{X}$ to the CapsNet and its reconstructed image $\mathbf{I}$ by the network as:

$$MSE = \frac{1}{NL} \sum_{i=0}^{N} \sum_{j=0}^{L} \Big[\mathbf{X}(i,j) - \mathbf{I}(i,j)\Big]^2 \tag{8}$$

The total CapsNet loss is obtained by (9), with the reconstruction loss being scaled down by $\alpha = 0.0005$ in order to reduce its influence over the margin loss.

$$\mathcal{L}_t = \mathcal{L} + \alpha MSE \tag{9}$$

C. Out-of-Distribution (OOD) Detection

The OOD decision algorithm is created after successfully training the model and is deployed for the testing stage. The trained model is used to produce reconstructed images of the training data. Then, reconstruction metric values are calculated between the original and reconstructed images. This value is used to classify the data as in or out of distribution based on a threshold learned from the training stage. Example histograms of the metrics are shown in Fig. 2. If the test instance's metric is outside the threshold, the instance is considered as OOD and $Unknown$ class is returned, otherwise the predicted class from the CapsNet's classifier is returned. It is observed in

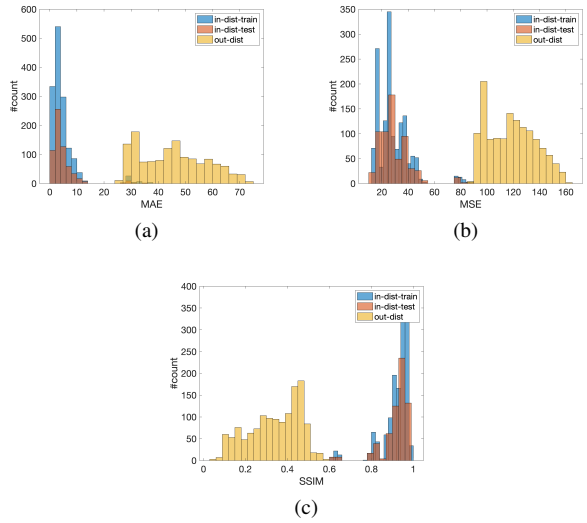

Fig. 2. Histogram of the reconstruction metrics (a) MAE (b) MSE (c) SSIM.

Fig. 3 that the reconstruction for in-distribution PRPD data is similar to the original input image, however the reconstructed OOD data images are closer to the in-distribution images than the original OOD images. This occurs because the CapsNet maps the learned features from the training data to the unseen test data. We investigate three reconstruction metrics employed in image processing including MSE, Mean Absolute Error (MAE) and SSIM. It is observed from Fig. 2 that SSIM provides the best separation between the in-distribution and OOD values. We introduce a threshold for the OOD decision which is the 5th percentile of the in-distribution SSIM values and the 95th percentile of the in-distribution MAE and MSE values. MSE was previously defined in (8), the MAE [23] is obtained by:

$$MAE = \frac{1}{NL}\sum_{i=0}^{N}\sum_{j=0}^{L}\left|\mathbf{X}(i,j) - \mathbf{I}(i,j)\right| \tag{10}$$

Since MAE and MSE attempt to identify the error, a low value is achieved for high similarity between two images and a higher value is obtained otherwise. SSIM identifies the change in the structural information of the image, and quantifies the quality of the second image with respect to the original one [24]. The SSIM is defined as:

$$SSIM = \frac{(2\mu_{\mathbf{X}}\mu_{\mathbf{I}} + z_1)(2\sigma_{\mathbf{XI}+z_2})}{(\mu_{\mathbf{X}}^2 + \mu_{\mathbf{I}}^2 + z_1)(\sigma_{\mathbf{X}}^2 + \sigma_{\mathbf{I}}^2 + z_2)} \tag{11}$$

where $z_1 = (0.01S)^2$ and $z_2 = (0.03S)^2$ are two default stability variables for the division with weak denominator, with S being the dynamic range of the 255 pixel values for 8-bit grayscale images ($2^{\#bits\ per\ pixel} - 1$). $\mu_{\mathbf{X}}$ and $\mu_{\mathbf{I}}$ represent the mean intensity of the original and reconstructed image respectively, $\sigma_{\mathbf{X}}$ and $\sigma_{\mathbf{I}}$ are their standard deviation respectively. The SSIM score lies between 0 and 1, where 1 is obtained for a perfect match between the original and reconstructed images.

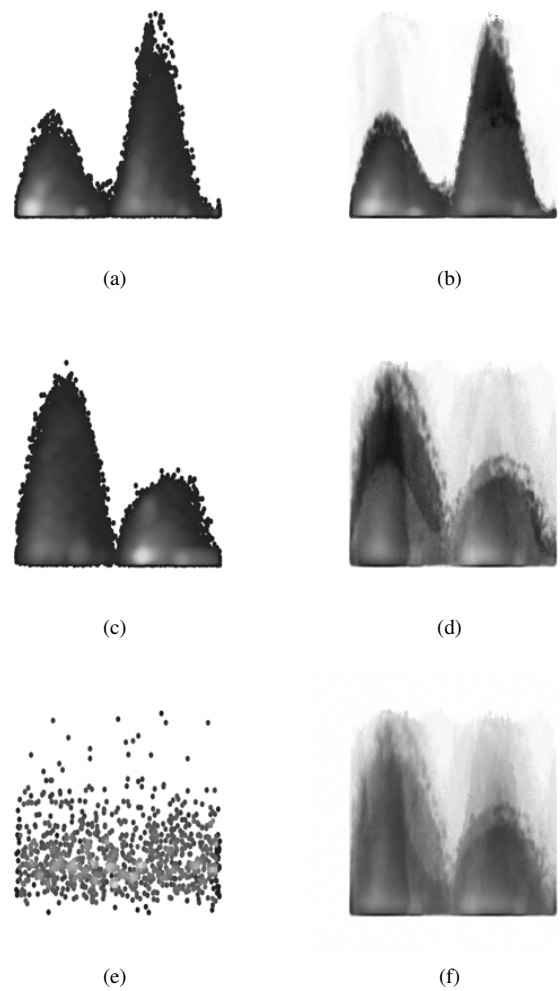

Fig. 3. Reconstructed PRPD images by CapsNet (a) Surface PD in slot original (b) Surface PD in slot reconstructed (c) PD adjacent to copper original (d) PD adjacent to copper reconstructed (e) OOD original (f) OOD reconstructed.

III. Experiment

The CapsNet network was trained on synthetic data to classify seven PD types over 100 epochs with Adam optimization algorithm implemented using a scheduler starting at learning rate of 0.001 which is reduced when a plateau is reached after a patience of 15 epochs without improvement in the loss. The minimum learning rate that can be used is bound to $1e^{-6}$.

A. Synthetic Data-set

The artificial PRPD data was measured in a laboratory set-up of 13.8KV stator bar with stressed line to ground insulation at 8KV. The set-up produced void discharge type of PD which was modified to create the remaining PD types as follows: The pattern's magnitude is varied to narrow and widen the distribution, the phase is shifted, and the probability of pulses is modified which controls the number of pulses. These modifications can be made to each half power cycle and result in the seven PD types. Each PD type is then modified by

TABLE I
CLASSIFICATION RESULTS OF CAPSNET WITH DIFFERENT RECONSTRUCTION MEASURES FOR OOD DETECTION.

Method	Overall acc. %	In-dist acc. %	OOD acc. %
CapsNet-MAE	96.61	93.23	100
CapsNet-MSE	96.67	93.34	100
CapsNet-SSIM	97	94	100

a random 1dB magnitude variance with a Gaussian distribution and a random phase variance of $\pm 2^\circ$ with uniform distribution. 300 samples per PD type were created providing a total of 2100 samples which were shuffled and split into a ratio of 70/30 to train/test the CapsNet. The 7 PD types include: void, surface, end-winding discharge, PD adjacent to copper, surface PD in slot and Phase to phase A and Phase to phase B. The training set is used in a 10-fold cross validation and the classification test results are presented for in-distribution, OOD and overall average accuracy.

B. Real-world Data-set

PD measurement was performed on 660MW, 24kV, hydrogen and water cooled generator made by Parsons in 1974. The obtained 1349 PRPD patterns were analysed by test engineers and none of the seven PD types were identified. Instead, background noise, floating potential, slot discharges and their combinations were observed in the PD data. This is an ideal scenario to evaluate the OOD performance of the proposed method as all the aforementioned observations are different from the classes used to train the model. Note that this data was not used in training the model.

C. Results

Classification results are presented in Table I with a comparison of the different reconstruction metrics. This problem can be seen as a binary classification between in-distribution and OOD data. It is observed that SSIM metric achieved better performance in each category and in the overall performance. The multi-classification results of the 7 PD types achieved 100% accuracy.

IV. CONCLUSION

This work achieved a successful OOD detection and PD types classification in rotating machines using CapsNet where OOD detection method is proposed by using the SSIM metric of the original and the reconstructed PRPD images without relying on OOD data during training. The real-world data in this work was limited to OOD data, however further data containing both in-ditribution and OOD instances will be collected and used to test the OODCN model in the future.

REFERENCES

[1] I. Mitiche, M. D. Jenkins, P. Boreham, A. Nesbitt, and G. Morison, "Deep complex neural network learning for high-voltage insulation fault classification from complex bispectrum representation," in *27th European Signal Processing Conference (EUSIPCO)*, 2019, pp. 1–5.

[2] B. Karthikeyan, S. Gopal, and S. Venkatesh, "Partial discharge pattern classification using composite versions of probabilistic neural network inference engine," *Expert Systems with Applications*, vol. 34, no. 3, pp. 1938 – 1947, 2008.

[3] J. Ren, P. J. Liu, E. Fertig, J. Snoek, R. Poplin, M. Depristo, J. Dillon, and B. Lakshminarayanan, "Likelihood ratios for out-of-distribution detection," in *Advances in Neural Information Processing Systems*, vol. 32, 2019, pp. 14 707–14 718.

[4] A. Salimy, I. Mitiche, P. Boreham, A. Nesbitt, and G. Morison, "Low complexity classification of power asset faults for real time iot-based diagnostics," in *IEEE Global Conference on Artificial Intelligence and Internet of Things*, 2020.

[5] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *International Conference on Neural Information Processing Systems*, 2017.

[6] G. E. Hinton, S. Sabour, and N. Frosst, "Matrix capsules with EM routing," in *International Conference on Learning Representations*, 2018.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[9] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017.

[10] F. H. Kreuger, E. Gulski, and A. Krivda, "Classification of partial discharges," *IEEE Transactions on Electrical Insulation*, vol. 28, no. 6, pp. 917–931, 1993.

[11] E. Gulski, "Computer-aided measurement of partial discharges in hv equipment," *IEEE Transactions on Electrical Insulation*, vol. 28, no. 6, pp. 969–983, 1993.

[12] S. Barrios, D. Buldain, M. Comech, and I. Gilbert, I.and Orue, "Partial discharge classification using deep learning methods—survey of recent progress," *Energies*, vol. 12, p. 2485, 2019.

[13] E. M. Lalitha and L. Satish, "Wavelet analysis for classification of multi-source pd patterns," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 7, no. 1, pp. 40–47, 2000.

[14] W. J. K. Raymond, H. A. Illias, A. H. A. Bakar, and H. Mokhlis, "Partial discharge classifications: Review of recent progress," *Measurement*, vol. 68, pp. 164 – 181, 2015.

[15] M. Jogin, Mohana, M. S. Madhulika, G. D. Divya, R. K. Meghana, and S. Apoorva, "Feature extraction using convolution neural networks (cnn) and deep learning," in *2018 3rd IEEE International Conference on Recent Trends in Electronics, Information Communication Technology (RTEICT)*, 2018, pp. 2319–2323.

[16] V. M. Catterson and B. Sheng, "Deep neural networks for understanding and diagnosing partial discharge data," in *2015 IEEE Electrical Insulation Conference (EIC)*, 2015, pp. 218–221.

[17] M.-T. Nguyen, V.-H. Nguyen, and Y.-H. Yun, S.-J.and Kim, "Recurrent neural network for partial discharge diagnosis in gas-insulated switchgear," *Energies*, vol. 11, p. 1202, 2018.

[18] M. Karimi, M. Majidi, M. Etezadi-Amoli, and M. Oskuoee, "Partial discharge classification using deep belief networks," in *2018 IEEE/PES Transmission and Distribution Conference and Exposition (T D)*, 2018, pp. 1061–1070.

[19] J. Tang, M. Jin, F. Zeng, X. Zhang, and R. Huang, "Assessment of pd severity in gas-insulated switchgear with an ssae," *IET Science, Measurement Technology*, vol. 11, no. 4, pp. 423–430, 2017.

[20] H. Song, J. Dai, G. Sheng, and X. Jiang, "Gis partial discharge pattern recognition via deep convolutional neural network under complex data source," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 25, no. 2, pp. 678–685, 2018.

[21] N. C. Sahoo, M. M. A. Salama, and R. Bartnikas, "Trends in partial discharge pattern classification: a survey," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 12, no. 2, pp. 248–264, 2005.

[22] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? a new look at signal fidelity measures," *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 98–117, 2009.

[23] J. Qi, J. Du, S. M. Siniscalchi, X. Ma, and C.-H. Lee, "On mean absolute error for deep neural network based vector-to-vector regression," *IEEE Signal Processing Letters*, vol. 27, p. 1485–1489, 2020.

[24] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.