Condition Monitoring Based on Partial Discharge Diagnostics Using Machine Learning Methods: A Comprehensive State-of-the-Art Review

Shibo Lu, Hua Chai, Animesh Sahoo and **B. T. Phung**
School of Electrical Engineering and Telecommunications
University of New South Wales, Sydney, Australia

ABSTRACT

This paper presents a state-of-the-art review on machine learning (ML) based intelligent diagnostics that have been applied for partial discharge (PD) detection, localization, and pattern recognition. ML techniques, particularly those developed in the last five years, are examined and classified as conventional ML or deep learning (DL). Important features of each method, such as types of input signal, sampling rate, core methodology, and accuracy, are summarized and compared in detail. Advantages and disadvantages of different ML algorithms are discussed. Moreover, technical roadblocks preventing intelligent PD diagnostics from being applied to industry are identified, such as insufficient/imbalanced dataset, data inconsistency, and difficulties in cost-effective real-time deployment. Finally, potential solutions are proposed, and future research directions are suggested.

Index Terms — **condition monitoring, deep learning, partial discharge, intelligent fault diagnostics, machine learning**

1 INTRODUCTION

CONDITION monitoring plays a vital role in the maintenance of the electricity supply networks. Internal structures of bulk operating power apparatus may be susceptible to excessive electrical, thermal, and mechanical stresses, along with environmental impacts, which may lead to severe insulation defects. The health management of insulation requires a fast-track of these defects in terms of their occurrence, location and pattern. To assess the lifespan of the electrical insulation in power system equipment, the level of partial discharge (PD) should be continuously monitored to avoid any unexpected breakdowns because PD activity serves as a primary indicator in monitoring the weakness of insulation [1]. PD, as defined by IEC 60270 Standard, is "a localized electrical discharge that only partially bridges the insulation between conductors and which may or may not occur adjacent to a conductor" [2]. Consequently, PD occurrence can accelerate the degradation of electrical insulation in high voltage (HV) equipment and cause catastrophic power outage if the associated defects are not treated at an early stage. In general, a PD monitoring system needs to achieve three objectives, which are detection, localization and recognition. PD events can occur stochastically with various locations and source patterns. The severity of the defects is not only dependent on the magnitude of the PD signal, which is related to the quantity of the discharge, but also the location of the weak point with respect to the insulation material. For example, in power transformer, void defects involved in solid insulation (e.g. windings) are more severe than bubbles immersed in oil, even though the PD level is significantly lower. In this regard, the former type of defect should be detected and tracked at an earlier stage. Therefore, information on PD location and type should be taken into account before any further preventative maintenance.

The development of the PD monitoring system requires a high level of accuracy, sensitivity, and robustness, which has been discussed over the decades by power grid operators and installation manufacturers. Traditionally, PD diagnostics are mainly based on handcrafted features extracted by conventional feature extraction techniques such as statistical, fractal, time, frequency features, etc. Simple threshold values are calculated for making decisions [3]. Then, some advanced signal processing techniques, such as discrete wavelet transform (DWT), are developed to extract more sophisticated and powerful features, and conventional machine learning (ML) methods are gradually employed for classification and regression tasks, including artificial neural network (ANN), support vector machine (SVM), and fuzzy inference system (FIS). [3]. Recently, with the advances in computing and information technologies, deep learning (DL), as a subset of ML, have been receiving increasing attention from both

Manuscript received on 30 April 2020, in final form 7 July 2020, accepted 7 July 2020. Corresponding author B. T. Phung.

DOI: 10.1109/TDEI.2020.009070

industry and academia for intelligent PD diagnostics. With deeper structures, the DL models are capable of extracting hierarchical features from the input data and achieving more reliable and accurate results. Furthermore, to handle the growing amount of data, end-to-end methodologies based on DL methods are proposed, which frees the human labor from feature engineering. The general procedures for ML based PD diagnostics are shown in Figure 1.

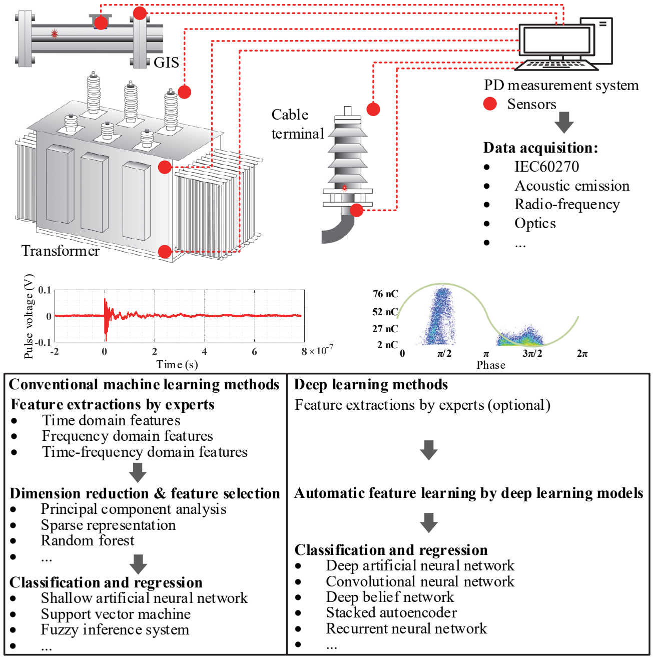

GIS
Cable terminal
Transformer
PD measurement system
Sensors
Data acquisition:
IEC60270
Acoustic emission
Radio-frequency
Optics
...
Pulse voltage (V)
0.1
0
-0.1
-2
0
2
4
6
8×10⁻⁷
Time (s)
76 nC
52 nC
27 nC
2 nC
0
π/2
π
3π/2
2π
Phase
Conventional machine learning methods
Feature extractions by experts
Time domain features
Frequency domain features
Time-frequency domain features
Dimension reduction & feature selection
Principal component analysis
Sparse representation
Random forest
...
Classification and regression
Shallow artificial neural network
Support vector machine
Fuzzy inference system
...
Deep learning methods
Feature extractions by experts (optional)
Automatic feature learning by deep learning models
Classification and regression
Deep artificial neural network
Convolutional neural network
Deep belief network
Stacked autoencoder
Recurrent neural network
...

Figure 1. General procedures for ML based PD diagnostics.

Given the importance of PD diagnosis, there is a large volume of published studies describing the role of methods regarding PD detection, localization and pattern recognition. [4] is an early survey, up to 2003, on PD recognition employing ANNs in terms of different fundamental algorithms. Raymond *et al* [3] conducted a comprehensive review in 2015 of PD classification, introducing and summarizing the methodologies proposed in signal denoising, feature extraction as well as classification. Mas'ud *et al* conducted another review in 2016, focusing on conventional ML based PD recognition using ANN [5]. However, those studies are limited to the traditional methods and the conventional ML techniques for PD classification. Furthermore, some other conventional ML methods such as decision tree (DT), k-nearest neighbors (kNN), and random forest (RF) are not adequately included and discussed. Most recently, Barrios *et al* [6] carried out a survey of recent progress using DL methods for PD classifications. However, the survey does not present different DL methods extensively and systematically, and it rarely provides potential guidelines for future directions. Therefore, it is essential to comprehensively review the latest research work, introduce the recent developments in PD diagnostics (detection, classification, localization, etc.) using ML algorithms, and discuss future trends and potential solutions to challenges.

In this regard, state-of-the-art PD diagnostics using ML methods applied in condition monitoring systems will be comprehensively presented in this review paper. This paper takes the form of six sections, including this introductory section. It will then go on to the background knowledge of PD and its detection regarding data collection methods in Section 2. The classical ML-based methods, along with performance evaluation, are discussed in Section 3. Section 4 presents a comparative discussion of DL-based application for intelligent PD monitoring techniques. Finally, the Section 5 and Section 6 provide a summary and a conclusion respectively on the implications of the development with respect to future research recommendations into PD diagnostics using ML approaches with the emphasis on DL methods.

2 PARTIAL DISCHARGE BACKGROUND

To detect PD activities, various sensing modalities have been explored such as electrical current impulse, by-products from chemical reactions, acoustic (pressure waves) emission, electromagnetic wave radiation. These products generated from the PD process can be measured using a range of sensing devices and analyzed using signal processing techniques. Consequently, a range of methods has been developed and applied to PD testing, including the electrical method, based on the IEC 60270 Standard, along with other non-conventional methods to achieve satisfactory monitoring performance. Given the measurement approach, methods can be classified into off-line and on-line detection methods. On-line monitoring is gaining more popularity due to the fewer power outages as well as fewer disturbances to the system operation with comparable performance. These detection methods regarding the PD data collection are introduced and summarized in the following sections.

2.1 IEC 60270 METHOD

PD measurement based on IEC 60270 Standard is well-developed due to its accuracy and potential to detect PD level at the off-line condition. The testing circuit for this method is shown in Figure 2, which includes the testing object C_x, the coupling capacitor C_b and the measurement impedance Z as the major circuit components. Current impulse below 1 MHz resulting from PD activity in the test object can be captured by the coupling capacitor and coupled with the measuring impedance. The signal can be presented in both the time-domain and phase-domain to illustrate the feature of the PD events [3]. Phase-resolved PD (PRPD) and phase-resolved pulse sequence (PRPS) are two forms of presenting the pattern of PD activities, which show the relation among discharge amplitude

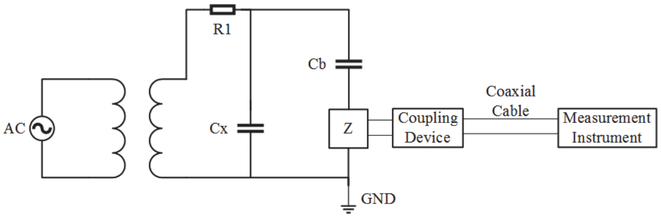


Figure 2. IEC 60270 PD testing circuit.

(q), cycle number (n) with respect to the phase position (φ). They are firstly proposed in [7] in 1990 and mainly based on the statistical features to illustrate the unique pattern for different types of PD defects. There are three different types of PD pattern, namely: surface discharge, void discharge and corona discharge. Considering the previous contributions [3], it can be noted that these patterns can be classified based on the information from the $q-n-\varphi$ representation. Unfortunately, a main shortfall of PRPD and PRPS is that it cannot separate the source types if multiple types of defect are involved, and the overlaps of the phase domain information can adversely affect the performance of PD classification. Another form to represent PD is time-resolved PD (TRPD), which enables the time-domain, frequency domain, and frequency-time domain analysis of PD pulses.

2.2 ACOUSTIC METHOD

Acoustic emission due to the excessive vibration of electrons is typically involved during the discharge process, and the frequency range of acoustic emission detection is 20 kHz–1MHz. Therefore, the acoustic method is practically applied for PD online monitoring due to its advantages such as less susceptible to electrical interference and capability to locate the discharge site [8]. Piezoelectric acoustic sensors can be mounted externally of the equipment, which provides a simple way of on-line monitoring of PD activities without a power outage. However, the sensitivity of acoustic measurement is not only dependent on the amount of energy of the PD source, but also the traveling path of the signal. Internal components of the operating apparatus, including inner insulation, solid and liquid insulation, and outer structure can substantially impact the characteristic of the propagation. Meanwhile, the acoustic attenuation can be significant with respect to distance between the PD source and sensor.

2.3 RADIO-FREQUENCY METHOD

Radio-frequency (R-F) method employs proper sensing devices to detect and capture the induced electromagnetic wave when PD occurs. Based on the frequency range of the R-F signals, this method can be divided into the high frequency (HF) method, very high frequency (VHF) method, and ultra-high frequency (UHF) method. HF signals (in 3-30 MHz band) can be measured by a high-frequency current transformer (HFCT) clamped over the ground terminal of the HV equipment. This provides accessibility and flexibility for the sensor installation if on-site testing is needed. However, the result can only indicate the existence of PD defects without showing an accurate location. VHF method is rare in practical applications because the size of the sensors is physically large, and the internal installation can be hard and dangerous. While UHF method is implemented widely for the on-line monitoring of PD sources due to its effectiveness of noise immunity and localization [9]. UHF method shows a high signal-to-noise ratio because the measurement frequency ranges from 300 MHz to 3 GHz [10], which is higher than the electromagnetic interference from the corona discharge in the surrounding environment. To capture UHF signals, UHF sensors (antennas) are inserted into the equipment via oil drain valves or dielectric windows.

2.4 OTHER METHODS

The optical method is applied by mounting optical sensors in the power apparatus. The competitive feature of this method is the immunity to electromagnetic interference (EMI). However, a major limitation of this method is that the sensitivity is excessively affected by the internal barriers within the equipment due to reflection, scattering, and attenuation of the light [11]. Another drawback of this method is the high cost of optical sensors, which requires further improvement.

Dissolved gas analysis (DGA) is generally used for PD detection based on chemical products due to the discharge. By dealing with the oil sample extracted from the transformer, PD activity can be indicated by specific analysis. However, the location cannot be estimated only by examining the chemical composition, and this can only be manipulated periodically, so DGA cannot be commonly applied in emergency failure detection in operation. This paper will not cover DGA based intelligent fault diagnostics since there is a well-organized paper published recently [12].

3 CLASSICAL MACHINE LEARNING BASED METHODS

3.1 ARTIFICIAL NEURAL NETWORK

ANN is inspired by biological neural networks and has been used for PD diagnostics for decades. For example, multilayer perception (MLP) is a class of feedforward neural networks consisting of at least three fully connected layers (one input and one output with one or more hidden layers) of non-linearly activating nodes. Given a dataset, $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^{n}$, of n samples, the corresponding label vector, and a k-layer MLP (the number of hidden layer is $k-2$), the mathematical representation of the output for j^{th} layer, $f^j(\boldsymbol{x}_i^j)$, is shown as follows:

$$f^j(\boldsymbol{x}_i^j) = \sigma^j({\boldsymbol{w}^j}^T \boldsymbol{x}_i^j + b^j) \tag{1}$$

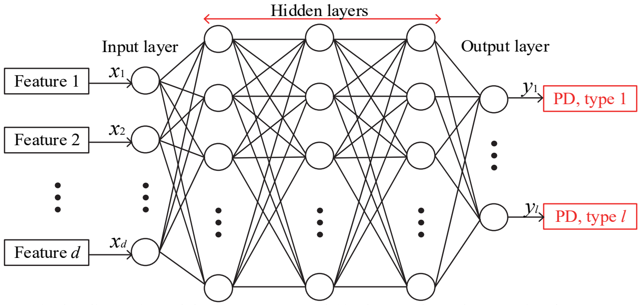
where σ^j is the activation function, $\boldsymbol{w}^j \in \boldsymbol{w}$ is the weight matrix, $b^j \in \boldsymbol{b}$ is the bias coefficient, $\boldsymbol{x}_i^j$ is the input of j^{th} layer, and $j = 2, \ldots, k$. Note that the number of neurons for hidden layers are flexible, while that of the input layer and output layer are identical to the dimension of input data and label vector, respectively. The backpropagation algorithm is widely used for training feedforward neural networks for supervised learning. Therefore, any ANN trained using backpropagation algorithm is also known as the backpropagation neural network (BPNN). Given a portion of the dataset $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^{m} \in \{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^{n}$ for training, the main objective of BPNN is to minimize the error, E (e.g. mean-square error), between the predicted output and the label vector, as shown in Equation (2).

$$\min_{\boldsymbol{w},\boldsymbol{b}} E(\boldsymbol{w},\boldsymbol{b}) = \frac{1}{m}\sum_{i=1}^{m}[f^k(\boldsymbol{x}_i^k) - \boldsymbol{y}_i]^2 \tag{2}$$

After feeding the training dataset, the calculated error will be propagated backward all the way to the input layer to update the parameters of the BPNN using the gradient descent with the learning rate of δ shown as follows:

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \delta\frac{E(\boldsymbol{w},\boldsymbol{b})}{\boldsymbol{w}}\,,\ \ \boldsymbol{b} \leftarrow \boldsymbol{b} - \delta\frac{E(\boldsymbol{w},\boldsymbol{b})}{\boldsymbol{w}} \tag{3}$$

In general, BPNN is the most selected types of ANN for PD diagnostics, and the general structure is shown in Figure 3. In [13], Lumba *et al* investigate the BPNN based PD recognition method for HV assets with statistical features of PRPD using different sensors. In [14], Sukma *et al* carry out a comprehensive study on PD classification using BPNN with different types of sensors, including transient earth voltage (TEV) sensor, surface current sensor (SCS) and HFCT, and different input patterns (waveform parameters and PRPD). BPNN with statistical features of PRPD is found to achieve better classification results in the experimental study. In [15], Soltani *et al* develop a denoising algorithm using ANN, aiming to suppress white noise in PD diagnostics applications using R-F signals. Compared to the conventional wavelet transform-based denoising method, the proposed method does not require prior knowledge of the signals and can achieve better performance. In [16], Kainaga *et al* develop a ML based PD classification method in the high-voltage direct current (HVDC) system. Three features, including statistical features, PD raw data (discharge magnitude and the time difference between two subsequent PD of the sequenced data), and normalized difference star (NoDi*), are used as the input of DT and ANN based classifiers, respectively. Based on the experimental results, statistical features or NoDi* based feature combined with ML based classifiers can achieve classification rate above 95%. In [17], Luo *et al* combine the wavelet-entropy vector extracted from TEV time-domain signal and BPNN to detect and classify PD. In [18], Wang *et al* apply DWT to extract distinguishable features from ultrasonic signals of PD, and then employ BPNN based on wavelet neural network to achieve PD classification in oil-filled submarine cable terminal. In [19], Li *et al* propose a two-stage UHF PD localization method for a substation. Received signal strength indicator (RSSI) of the sensor array is captured as the input of a BPNN to produce preliminary localization results. This step mitigates the impacts of Gaussian noise in the substation and measurement error caused by the heterogeneity of wireless UHF sensors by reducing the size of the fingerprint map. After that, compressed-sensing based localization is performed with a smaller fingerprint map to achieve satisfactory localization accuracy. In [20], Iorkyase *et al* use the received signal strength (RSS) vector from PD measurements by R-F sensors is used as the input of kNN/BPNN to solve a regression problem. BPNN shows a slightly better localization accuracy because of the lower error standard deviation and mean errors. In [21], Zahed *et al* design three Hilbert fractal antennas for PD measurement in transformer and evaluate those sensors via signal-to-noise ratio (SNR) analysis and PD classification. For PD classification, BPNN based classifier with fast Fourier transform (FFT) based features extracted from UHF signal is employed. In [22], Polisetty *et al* develop a PD classification method based on BPNN and harmonics components of acoustic emission signals for outdoor insulator.



d is the dimension of the input array and *l* is the number of categories for output

Figure 3. Structure of a back propagation neural network.

Some researchers apply feature selection or dimension reduction methods prior to the ML classifiers, aiming to reduce the training time and improve diagnostic accuracy. In [23], Khan adopts BPNN for PD recognition in gas-insulated switchgear (GIS) with statistical features of PRPD as input. The impact of principal component analysis (PCA) is also investigated, and it is found that PCA can significantly reduce the execution time of ANN, while maintaining the diagnosis accuracy approximately the same. In [24], Dobrzycki *et al* apply BPNN and SVM for PD detection in epoxy resin with statistical features of acoustic emission signals as input. PCA is applied to achieve dimension reduction to increase the training speed, while the accuracy of each network dropped by a value that did not exceed 0.5%. In [25], Anjum *et al* propose a BPNN based PD classification technique for ceramic insulators in overhead (OH) transmission lines. After applying wavelet packet decomposition (WPD) to the R-F time series signal, statistical parameters (entropy, energy, skewness, and kurtosis) of each decomposition level are calculated. A feature selection method based on the computation of within-class scatter and between classes scatter is used to select the top two features from each statistical parameter, resulting in an input vector with eight elements.

Besides, other ANN methodologies such as radial basis function network (RBFN), probabilistic neural network (PNN), generalized regression neural network (GRNN), extension neural network, extreme learning machine (ELM) are also applied and investigated. In [26], Wang *et al* apply extension neural network to detect PD in the power capacitor, and chaos synchronization detection method is used to pre-processing the HFCT signal to greatly reduce the data size. In [27], Iorkyase *et al* describe a supervisory system for PD monitoring in a substation. Authors use signal strength ratios (SSRs), which is the ratio of RSS components captured by the multiple receivers, combined with ML classifier for PD localization. kNN and ANN based classifiers are investigated through experimental study, MLP and GRNN have better performance over kNN. GRNN employs single-pass learning (no backpropagation required), which makes it faster to train, and it can be generalized well with less training data. As a result, GRNN fits well with Internet-of-things (IoT) towards edge computing, since it can achieve PD localization with adequate accuracy at a reduced computational cost. However, there is no optimal method to improve GRNN, and sometimes GRNN can be oversized in order to achieve satisfactory performance, which makes it computationally expensive (require more memory space to store the model). In [28], Hou *et al* develop a new algorithm based on the spectrum reconstruction of the UHF signal and RBFN for separating multiple PD sources. The signal identification rate is higher than 75% for SNR>10dB in simulation verification results and 70% in substation onsite test results. In [29], Zhou *et al* propose an enhanced UHF PD localization method for substation based on time difference method and multiple RBFNs. In the calibration stage, RBFNs study and simulate the error distribution of the system. Then, during the testing, the error of actual measurement data can be corrected. Instead of using one RBFN to correct the time delay and position error for the whole localization range, three RBFNs separately handle a portion of the localization range. The effectiveness of this improvement is validated through theoretical analysis and experimental study. In [30], He *et al* proposes a two-stage ML based transformer PD recognition method using the PRPS map. Since the PRPS map consists of a significant amount of information, sparse autoencoder (SAE) is applied at the first stage to obtain the essential components of the PRPS map. Then ELM algorithm is used to achieve PD classification. Compared to SVM and BPNN, ELM significantly improves the training speed with higher recognition accuracy. In [31], Zhang *et al* propose a PD pattern recognition method for transformer based on statistical features of UHF PRPS pattern and online-sequential ELM (OS-ELM). The diagnosis performance of this method is better than SVM and BPNN. It is also reported that OS-ELM has better generalization capabilities compared to SVM and BPNN, since the recognition accuracy only decreases slightly when reducing the size of the training sample from 800 to 200. OS-ELM also has fast training speed and thus shorter training time. In [32], Gianoglio *et al* propose a PD recognition method for the rotating machine based on the statistical features of PRPD and ELM. Authors also develop a hardware-friendly ELM digital implementation and validate in complex programmable logic devices (CPLDs) and field-programmable gate arrays (FPGAs). Although ELM can dramatically increase the training speed, the randomness mechanism of ELM (also the key contributor for increasing the training speed) can cause an additional uncertainty problem and degradation phenomenon also exists with improper choose of ELM based on theoretical analysis [33]. Therefore, more investigations need to be carried out in order to use ELM properly.

Instead of using only one ANN classifier to make the decision, the ensemble neural network (ENN) combines multiple ANN models using two major kinds of ensemble learning techniques: bagging and boosting. In [34], Mas'ud *et al* use bagging-ENN with statistical PRPD features for PD classification. Based on the experimental results, when there are resolution changes (amplitude and phase) between training dataset and testing dataset, ENN shows a better property of generalization. In [35], another type of ENN using the boosting algorithm is used by the same authors for PD classification with PRPD patterns. Similarly, ENN is verified to be more effective than single ANN with a 10% improvement of classification rate. Boosting algorithm attempts to reduce the bias, and trained models are weighted by their performance. It is more suitable when the classifier is more stable (less variance in prediction). In contrast, the bagging algorithm tries to decrease the variance, and trained models are averaged. It is more suitable for mitigating the over-fitting problems. The advantages of ensemble methods are: 1) usually it can improve the performance over any single network; 2) less probability to overfit and more stable. The disadvantages of ensemble methods are: 1) it does not perform well on simple dataset (e.g. data come from a linear process); 2) they are usually computationally expensive, and therefore, it adds additional learning time and more memory constrains to the application; 3) reduction in model interpretability. Other techniques also apply to fuse the outputs of multiple ANNs: In [36], Li *et al* develop a decision-making system with a fusion of both time domain and PRPD information for PD recognition in GIS. Two BPNNs are trained for PD recognition under the TRPD and PRPD model, respectively. Then, Dempster-Shafer theory (DST) is applied to fuse the results from two BPNNs, aiming to minimize the possibility of misclassification.

Data augmentation techniques are also helpful for training ANN models in the scenarios of small dataset and imbalanced dataset. In [37], Tra *et al* propose an ML based transformer fault classification using DGA (non-code ratios as input). MLP, SVM, and kNN are investigated, and MLP demonstrates slightly better fault diagnosis performance. Under imbalanced data condition, ML based classifier has lower generalization performance since a bias has given during the training process. Adaptive synthetic minority oversampling technique (ASMOTE) is then adapted to make each category balanced to improve the performance of ML classifiers.

ANNs offer several advantages, such as easily solving multi-classification problems and able to manage large amount of data and input variables. However, ANNs have low interpretability because of their black-box nature. In addition, they cannot always find the global optimal for relatively large ANN models, which makes them prone to be overfitted.

3.2 SUPPORT VECTOR MACHINE

Take the non-kernel SVM (or linear SVM) as an example: consider a binary classification problem, given a dataset, $\{\boldsymbol{x}_i, y_i\}_{i=1}^{n}$, of n samples and the corresponding label, where $\boldsymbol{x}_i \in \boldsymbol{x}$ and $y_i \in \{-1, 1\}$, a hyperplane $f(\boldsymbol{x}) = 0$ is chosen to separate the data into a positive and a negative group, shown as follows:

$$f(\boldsymbol{x}_i) = \boldsymbol{w}^T \boldsymbol{x}_i + b = 0 \tag{4}$$

where $\boldsymbol{w}$ and b are the parameters to determine the hyperplane. Then, a positive boundary and negative boundary can be determined by the closest point from the hyperplane in each group. Any point above the positive boundary is of one class with label 1, while any point below the negative boundary is of one class with label -1. After rescaling the distance of the closest point from the hyperplane in each group to be 1, the chosen hyperplane is subject to the following condition to separate the dataset:

$$y_i f(\boldsymbol{x}_\mathrm{i}) = y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1, i = 1, 2, \dots, n \tag{5}$$

As shown in Figure 4, in order to find the optimal hyperplane to achieve perfect separation, the margin $\gamma = 2/\|\boldsymbol{w}\|$ (the distance between the positive boundary and negative boundary) is expected to be maximized, where $\|\cdot\|$ is the norm operator. As a result, the following optimization problem is formulated for non-kernel SVM [38]:

$$\begin{gathered} \min_{\boldsymbol{w},b} \; Cost(\boldsymbol{w}, b) = \frac{\|\boldsymbol{w}\|^2}{2} \\ s.t. \; y_i(\boldsymbol{w}^T\boldsymbol{x}_\mathrm{i} + b) \geq 1, i = 1, 2, \dots, n \end{gathered} \tag{6}$$

Note that the square term in Equation (6) is for computation optimization purpose. For the dataset that is not linearly separable, hyperplane with a soft margin can be found by adding regularization terms in the cost function in Equation (6).

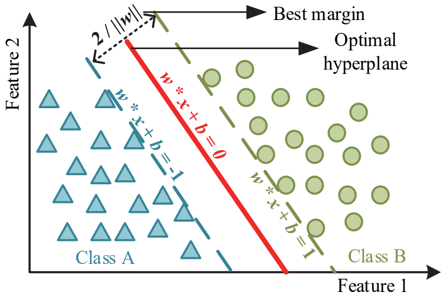


Figure 4. A simple linear SVM for classification.

Generally, the basic SVM only supports binary classification. However, PD classification is a multiclass classification problem, which requires the use of more than two classes. Therefore, it can be reduced to multiple binary classification problems through strategies such as one-against-one (OAO) and one-against-all (OAA). The types of kernels used have a great impact on the performance of an SVM classifier. Common kernels are linear-kernel, polynomial kernel (e. g. cubic and quadratic), Gaussian or radial basis function (RBF) kernel, Kullback-Leibler (KL) divergence based kernel. In [39], Herath *et al* conduct a comparison study on PD classification in generator insulation using supervised ML based classifiers with statistical features of PRPD. Based on the experimental study, most of ML techniques can reach over 95% classification accuracy. SVM (linear kernel) and MLP have the best performance compared to others, and SVM has a slightly higher receiver operating characteristic (ROC) area over MLP. In [40], Firuzi *et al* propose a PD recognition method based on SVM with local binary pattern (LBP) and histogram of oriented gradient (HOG) features of PRPD in transformer applications. Selection of kernels is very important for achieving the desired accuracy, and different SVMs with polynomial and Gaussian kernels are investigated. In the study, the best accuracy is HOG-SVM with cubic kernel (99.3%), and statistical features with SVM has a fairly good accuracy (89.3%). The same authors further use the proposed HOG-SVM as a post-processing step to assist external PD source separation in transformer applications [41]. In [42], Li *et al* develop a PD localization method for GIS based on image edge detection and SVM, where the edge features are extracted using the Canny algorithm from short-time Fourier transform (STFT) based energy density distribution of UHF signals. In [43], Kunicki *et al* develop a two-stage ML based classification method for select defects (different types of PD as well as typical defects) in power transformers using acoustic emission signals. Energy features extracted by DWT are used as the input and SVM (cubic kernerl) achieves the best performance compared to kNN, DT, and ensemble bagged DT. This kind of methodology can effectively eliminate the impacts from on-site acoustic emission disturbances and play as a supplementary part of the condition monitoring system to detect other types of defects in a power transformer. In [44], Desai *et al* propose an SVM (quadratic kernel) based PD classification method for transformer with features extracted from the UHF signal by time-frequency analysis. In [45], Yao *et al* develop the polar coordinate pattern obtained from the UHF PRPS dataset to classify PD types in GIS. K-means clustering is applied to characterize the parameters in the polar coordinate pattern, and the calculated parameter vector is used as the input of SVM (Gaussian kernel) to achieve PD classification. The proposed method is compared to the traditional method, where statistical features of PRPD is used as input to the SVM classifier, and an average improvement of 4.4% is obtained. In [46], Wang *et al* apply advanced signal processing techniques to extract chromatic parameters from UHF signals, and use SVM (Gaussian kernel) to achieve PD recognition in GIS. In [47], Jineeth *et al* employ ANN and SVM (Gaussian kernel) with statistical features of PRPD to achieve PD classification in cross-linked polyethylene (XLPE) cables. Compared to ANN, SVM significantly improves the PD diagnosis performance. In [48], Robles *et al* use SVM with power spectrum density (PSD) of HFCT signal to identify different PD types in the transformer. The application of KL-based kernel demonstrates superior performance compared to Gaussian Radial Basis Function kernel, because Gaussian kernel will give the same importance to all the frequencies of the spectrum, while the KL-based kernel only takes important frequencies into consideration.

Some feature selection or dimension reduction methods are also used to improves the diagnosis and computation performance of SVM. In [49], Xie proposes a PD-based aging state classification method for the transformer using statistical features of PRPD and SVM. It is reported that the SVM with PCA (10 principle component factors) performs better with about a 6% increase in prediction then SVM with the original 27 characteristics factors. In [50], Zang *et al* use SVM, kNN, and linear discriminant analysis (LDA) to achieve PD recognition in gas-insulated transmission lines (GIL). With the help of a newly designed photoelectric fusion based feature using optical PRPD pattern and UHF PRPD pattern, the diagnosis accuracy of ML based classifiers improves significantly. PCA is applied to reduce the feature parameters before the classification step, and the SVM classifier performs the best based on the experimental study. In [51], Tang *et al* introduce a PD severity assessment method using an SVM classifier and a feature selection method based on minimum-redundancy and maximum relevance (mRMR). Statistical PRPD features are extracted from the UHF signal as the input of the proposed algorithm. In [52], Raymond *et al* develop a PD classification method for XLPE cable based on PCA and SVM/ANN. The PRPD features extracted by PCA shows high noise tolerance capability compared to conventional statistical and fractal features, and SVM reaches better diagnosis accuracy. In [53], Morette *et al* develop an SVM-based PD recognition method for HVDC cable. Statistical features and DWT based features of pulse signals are used as input to SVM for investigation, and DWT based features are proven to be more robust than others. Furthermore, Gram-Schmidt orthogonalization procedure is adopted to rank the features. It has been shown that the most relevant feature is sufficiently informative for nonlinear SVM to achieve satisfactory recognition rate, and nonlinear SVM is better than linear SVM. In [54], Mitiche *et al* use a general linear chirplet transform (GLCT) to extract enhanced PD features (spectrogram) from a time-domain CT signal. Compared to STFT and LCT, GLCT provides higher energy concentration and better resolution on the time-frequency domain with less cross-term interference. LBP approach is used to keep the useful information and further reduce the input dimension. Then, an SVM classifier is proposed to achieve accurate EMI transient classification (PD, corona, minor PD, and other EMI transients). Grid search method is applied to find the optimum kernel function.

Some researchers apply modified-SVM to obtain better results, such as least-square SVM (LSSVM), one-class SVM (OCSVM), fuzzy SVM, relevance vector machine (RVM), and hypersphere multiclass SVM (HMSVM). In [55], Iorkyase *et al* develop an LSSVM based method for localization of PD sources in a substation. The RSS vector from three R-F sensors is used as the input of LSSVM to solve a regression problem. LSSVM slightly improves the PD localization accuracy compared to SVM and produces the lowest location estimate error among all methods. Another advantage of LSSVM over SVM is that it only requires a solution for a linear problem instead of a quadratic problem, which makes the training of LSSVM more computationally efficient. In [56], Janani *et al* propose a hybrid PD classification method for the scenarios where multiple PD sources are active at the same time in GIS. In the first stage of the algorithm, the dimension of the features extracted from PRPD is reduced using PCA algorithm. Then, OCSVM is applied to classify the input into single-source PD and non-single-source PD (multiple-source PD). In the second stage, a two-step logistic regression (LR) model is applied to estimate the probability of each PRPD pattern from multi-source PD to achieve PD classification. OCSVM is perfectly suitable for some applications that have an asymmetry in the nature of the classes: the samples of target class are easy to collect and label, while the rest of the classes are poorly defined. One of the main drawbacks of OCSVM is that it requires a well-defined class. In other words, one needs to include all types of target PDs into the training process for a similar application in [56]. Some experiments also reveal this drawback of OCSVM when using PD as the known class in [57]. To mitigate this problem, Parrado-Hernandez *et al* propose a novel PD detection method based on OCSVM (KL based kernel) with the PSD based feature of the environmental noise as input. For PD detection applications, different types of PDs have different characteristics, while the background noise is more homogeneous in the similar experimental environment, which makes it a perfect candidate as the input of OCSVM. Then the goal of OCSVM is to separate background noise and non-noise (PD signal). However, when the applied environment changes, the noise distribution can be different. This can be easily solved using the domain adaptation technique by fine-tuning the OCSVM with the background noise from the current setup. Based on the experimental results, the best PD detection accuracy improves from 90.17% to 99.67 after applying domain adaptation. The main advantages of the proposed method are: 1) PD dataset is not required to train the OCSVM; 2) training set is completely unlabeled (need to ensure that the data is recorded under PD-free conditions). It can be easily applied to real industrial applications. In [58], Janani *et al* extensively investigate the automated statistical PD classification method based on ML. Based on the experimental results from different combinations of extracted features of PRPD and ML classifiers, the fuzzy classifier (F-SVM and F-kNN) show a high recognition rate. Such methods can calculate the 'degree of membership' of a sample to a class of data, which allows probabilistic interpretation of an unknown PRPD pattern that is being classified. In [59], Shang *et al* develop a PD recognition for transformer based on RVM. A new method based on ensemble empirical mode decomposition (EEMD) and sample entropy is applied to extract UHF PD features as input to RVM classifier. RVM shows better performance compared to BPNN, PNN, and SVM. In [60], Shang *et al* apply HMSVM for PD recognition, where the input feature is extracted using variational mode decomposition (VMD) and multi-scale dispersion entropy (MDE) from the UHF signal. HMSVM can be directly used for multi-class classification problems since the samples from the same class are assigned to a hypersphere (input data are mapped to several hyperspheres). Therefore, HMSVM can overcome some disadvantages of OAO and OAA, and the performance of HMSVM is better than conventional SVM as expected according to the PD recognition results.

Cross validation and grid search are generally used to find the optimized parameter set of SVM. Another improvement can be made is tuning the parameters (e. g. penalty factor, slack variable, and kernel function parameters) using enhanced optimization techniques. In [61], Zhang *et al* apply two-dimensional PCA to compress the gray-scale PRPD images to find optimal features. Then, SVM optimized by particle swarm optimization (PSO) is used as a classifier to achieve PD classification in the transformer. The recognition accuracy of with proposed feature extraction method can achieve a 7% average improvement over the traditional PRPD based method using statistical features. PSO is also adapted to select optimal parameters in HMSVM in [60]. In [62], Duan *et al* develop a parameter-optimized SVM based PD recognition method for XLPE cable, where fractal features extracted from PRPD is used as input. The genetic algorithm (GA) is employed to optimize the parameters of SVM. It significantly improves the diagnosis accuracy by 4.7% compared to SVM without optimization.

SVMs have a solid mathematical foundation in statistical learning theory. The solution for a typical SVM is a convex optimization problem, which can always find global optimal. Therefore, SVMs are less prone to overfitting problems compared to ANNs. However, the main disadvantage is SVMs are sensitive to the optimal choice of kernel. Additionally, they are computationally inefficient with a large dataset and does not work well when the number of input features is greater than the number of samples. Consequently, it is required more works in feature engineering for SVMs than ANNs.

3.3 K-NEAREST NEIGHBORS

Given a training dataset, $\{\boldsymbol{x}_i, y_i\}_{i=1}^{n}$, of n samples, the corresponding label, an unlabeled testing dataset $\{\boldsymbol{x}_j\}_{j=1}^{m}$. The kNN classification algorithm firstly computes distance $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ to every training example $\boldsymbol{x}_i$. One of the most common distances, Euclidean distance, is shown in Equation (7).

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|^2 \tag{7}$$

Then, it selects k closest instances $\{\boldsymbol{x}_{i1}, \dots, \boldsymbol{x}_{ik}\}$ and their labels $\{y_{i1}, \dots, y_{ik}\}$. Finally, it can output the most frequent label in $\{y_{i1}, \dots, y_{ik}\}$ as $y_{predict}$ for the input sample. For the kNN regression algorithm, the last step is modified: it computes the mean of $\{y_{i1}, \dots, y_{ik}\}$ as follows:

$$y_{predict} = \frac{1}{k}\sum_{n=1}^{k} y_{in} \tag{8}$$

kNN and its variants (e.g. weighted kNN, fuzzy kNN) are applied to condition monitoring using PD diagnostics such as PD localization [20, 27, 63], PD classification [50, 58, 61, 64, 65] and PD detection [37]. In [64], Harbaji *et al* develop a PD classification method using acoustic emission measurement based on PCA and kNN. Similar to the observations from [66], different PD source locations, oil temperatures, and the presence of obstacle severely affect the classification accuracy. By training with all acoustic emission samples at different conditions, the proposed classifier can achieve a level of 90% for different cases. About 4% reduction is observed compared to the case where the classifier is trained with different PD types measured in the same condition. Some of the other referenced papers are discussed in the previous session, and most of kNN based methods are worse than other types of ML methods such as SVM and ANN.

There are two main advantages for kNN: firstly, it is a very simple ML model and easy to implement; secondly, it has fewer hyperparameters to tune. However, the computational cost increases significantly when the sample size is large, and the value of *k* is hard to select. Additionally, Goodfellow *et al* state the output of kNN on small training sets will essentially be random [67]. All those negative factors make kNN less popular in this field.

3.4 FUZZY INFERENCE SYSTEM

The FIS system is a knowledge-based system based on fuzzy logic (FL) theory, which is simply a nonlinear mapping from the input space to the output space. A FIS system typically consists of five parts: fuzzy set, fuzzifier, fuzzy rule, inference engine and de-fuzzifier. Given the crisp input set, a pre-determined membership function is applied to convert the input set to the fuzzy set. Then, Inference is performed based on the pre-constructed “IF-THEN” rules where several fuzzy operators (OR, AND, and NOT) are applied in “IF” conditions. After all rules are evaluated, a crisp output value can be obtained by defuzzification according to the membership function.

In [68], Faiz *et al* comprehensively compare the application of ML based method and conventional methods in transformer fault detection using DGA. Both ANN and FIS demonstrate improved fault detection ability. In [69], Lumba *et al* analyze the characteristic of PD using FIS with statistical features of PRPD. In [70], Mas'ud *et al* developed a PD recognition method using ML based classifier and statistical features of PRPD. Comparative studies are carried out between ANN and FIS. The main advantage of FIS over ANN is that FIS does not require large data for correct interpretation. However, one interesting phenomenon can be obtained from the experimental results: FL can only provide classification probability of the PDs as either correct or incorrect, while ANN is possible to classify the PDs with similar geometry. This suggests that FL can hardly understand the similarity of the PD defects and whether they have the same geometrical arrangement or not. As a result, the authors conclude that ANN is more suitable for PD recognition in practical conditions despite FIS has some advantages over ANN. In [71], Zeng *et al* propose a PD severity assessment method using two-level FIS and statistical features of UHF PRPD data. FIS is more convenient for training compared to SVM in this study, since the specific label is needed for SVM, which means the severity degree of each PD sample should be given clearly.

FIS requires the expert knowledge, and the incorrect and incomplete knowledge could dramatically affect the performance of FIS. Adaptive neuro FIS (ANFIS) combines the advantages of ANN and FIS, which does not require expert knowledge. In [72], Kari *et al* propose a DGA based transformer fault diagnosis method based on integrated ANFIS and DST. DST is introduced to handle conflicting outputs to avoid confusion in the decision-making process. The proposed method outperforms conventional methods, with PD diagnosis accuracy of 76.7% compared to 43.9%. Algorithm optimization can be used to optimize the performance of ANFIS, especially in the scenarios of complex input samples. In [73], Wang *et al* develop a PD detection method based on improved ANFIS with simplified structure and better model accuracy. The number of pulses detected by the optical sensor, temperature, and humidity are used as input to the improved ANFIS. Firstly, the subtractive clustering method (SCM) and fuzzy C-means (FCM) algorithm are used to initialize ANFIS instead of using the default grid partition method for the traditional ANFIS. For a complex nonlinear input, a FIS needs to achieve high prediction accuracy through dividing the fuzzy subset of each input into small elements (more fuzzy rules), which increase the computation complexity exponentially. The proposed SCM-FCM can efficiently achieve the purpose with reduced computation complexity. Then, an improved ANFIS is introduced by optimizing the traditional ANFIS with Fletcher-Reeves conjugate gradient method, and it reduces the model error by 2% based on the experimental study. However, ANFIS is not quite suitable for some applications under certain conditions. For example, in [74], Raymond *et al* comprehensively investigate the PD classification method using three types of ML classifiers under different levels of noise contamination conditions. Statistical, fractal, and PCA based features of PRPD are considered as input to ANN, ANFIS, and SVM classifiers. PCA technique demonstrates the superior performance in extracting distinguishable features from PD signals. ANN and SVM performs better than ANFIS when using PCA features as input. The main cause is that ANFIS requires normalization of the input data in the training process, which change the relative significance between each PCA component (PCA components have different weighting) [75]. ANFIS also suffers from other limitations such as the type of membership functions, the number of the membership functions, less interpretability, and difficulty of implementation in big data paradigm.

3.5 DECISION TREE AND RANDOM FOREST

DT is a supervised learning algorithm widely used in classification problems, which breaks the input space into regions and has separate parameters for each region. It is a decision-making process by establishing the relationship between the attributes and the class using a flowchart-like structure. A simple example of PD classification using DT is visualized in Figure 5.

DT based classifier is applied to PD classification [16, 66, 76], and fewer cases are reported about the superior performance of DT. This is mainly because they are usually unstable, and a tiny change in the data can cause a significant change in the optimal structure of DT. On contrast, DTs are simple to interpret and easy to achieve good performance with simple input data, which are suitable for some specific applications. For instance, in [77], aiming to find the optimum decomposition level for reducing noise in PD signals, Soltani *et al* develop a multi-level DT based method with energy spectral density of HFCT PD signal as input. Seven trees are designed to classify specific parameters, respectively. In this way, the complexity of the input to each DT can reduce greatly, which ensures the accuracy of each DT classifier.

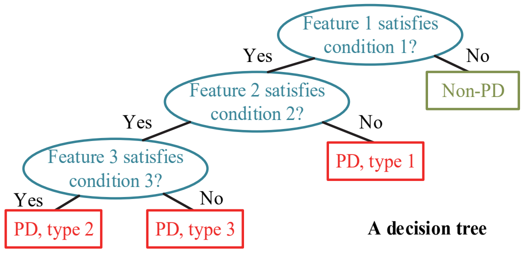

Feature 1 satisfies condition 1?
Yes
No
Non-PD
Feature 2 satisfies condition 2?
Yes
No
PD, type 1
Feature 3 satisfies condition 3?
Yes
No
PD, type 2
PD, type 3
A decision tree

Figure 5. A simple illustration of decision tree for PD classification.

RF is one particular tree-based model that mitigate overfitting in DT by integrating multiple DT-based classifiers, and RF based methods gain some achievements compared to DT. In [76], Wang *et al* develop a ML based transformer PD recognition method using statistical features of PRPD as input. RF based method hits the best diagnosis accuracy compared to other classifiers such as SVM (polynomial/RBF kernels). Similar method and results can be obtained in [78]. In [79], Iorkyase *et al* introduce a PD localization method for substation using WPD to extract location dependent features from PD signals captured by three R-F sensors and RF regression to localize PD. Compared to DT regression and boostrap aggregating DT models (bagging as ensemble learning method), RF regression demonstrate lower mean localization error and lower variance between training and testing predictions. In [65], Hussein *et al* develop a FFT based feature and employ different ML classifier to achieve acoustic PD classification in HV bushing system. Among all ML classifiers, RF classifier achieves superior performance even though the signal is corrupted with noise. In [66], Woon *et al* develop a PD classification method for oil-paper insulation systems based on PSD histogram of acoustic emission signals and ML based classifiers. Different ML techniques are comprehensively investigated under different experimental conditions. It is reported that variations in oil temperature and the presence of barriers between the sensors and PDs have a relatively smaller impact on PD recognition accuracy. Furthermore, the changes in the position of the sensor relative to the PDs demonstrate a more significant influence. Although SVM demonstrates better performance during the cross-validation, it performs not quite well when variations existed between training and testing dataset. By contrast, RF and gradient boosting (GB) show the ability to maintain relatively satisfactory classification accuracies in such a condition. Those examples mentioned above demonstrate the good property of generalization of RF based methods. In [80], Si *et al* propose an effective solution to recognize PD for UHVDC converter transformer based on ML classifier (RF or LSSVM). Single PD pulse waveform, as well as its phase information, are combined as the input to the classifier. Since the data point of each PD pulse is not consistent, zero-padding method is applied to make sure the input vector maintains the same dimension. From experimental results, the proposed feature is more suitable for PD recognition compared to the conventional features (lower input dimension) in [29]. LSSVM and RF based classifiers show the same recognition rate and processing time when conventional features as input. However, RF based classifier demonstrates its superior efficiency in processing large dataset with larger input dimension. Another way to improve the performance of RF, similar to other ML based method, is to introduce proper feature extraction and feature selection methods. In [81], Bag *et al* develop an optical signal based PD localization technique for oil-immersed equipment (e.g. transformer and breaker). S-transform is applied to extract features from captured optical signals from five optical sensors. Higher ranked features are selected using the recursive feature elimination method, and RF is used for localization. The trade-off between the number of input features and computation time is investigated and discuss in detail. In [82], Wang *et al* use isolation forest (IF) combined with linear prediction cepstrum and PCA to separate multi-source PD signals in GIS and transformer.

Besides act as a classifier, RF is also useful in extracting informative features from the input to achieve dimension reduction. In [83], Peng *et al* propose an RF based optimal feature selection for PD classification in HV cables. 1235 features are generated from a single PD current pulse for investigation. The RF based feature selection method is then validated by BPNN and SVM classification results. By selecting informative features, a good balance can be achieved between complexity and recognition accuracy. The main advantages of RF based feature selection method are less probability of overfitting with tuned hyperparameters and explicit ranking results of all features.

3.6 OTHERS

Besides the popular methods reviewed above, other methods, such as tensor-based classifier, multiple linear regression (MLR), Gaussian mixture model (GMM), blind signal separation (BSS), sparse representation classifier (SRC), support vector data description (SVDD), Bayesian network (BN), variable predictive model-based class discrimination (VPMCD), and rough set (RS) theory, are applied for different applications in PD diagnostics.

In [84], Gianoglio *et al* use a tensor-based classifier with PRPD data for PD classification in the electric motor. In [85], Goncalves Junior *et al* develop a PD localization method in transformer windings using MLR model and statistical features extracted from PD high-frequency current pulses. MLR model has slightly higher localization accuracy compared to ANN model, but MLR is easier to be implemented because of its simplicity. However, the ANN used in this paper is too simple, which might introduce an unfair comparison between ANN and MLR. In [86], Yan *et al* propose a framework based on GMM for the probabilistic PD classification in electronic equipment. Semi-supervised learning is adapted for training, and GMM based model can classify new types of PDs with at least 86.3% accuracy. In [87], Boya *et al* apply the BSS technique to recover every single source acoustic emission PD signal from mixed PD signals. The recovered signal is validated through a comparison between the recovered signal and the original signal using time-domain visualization and energy fingerprint from DWT. The same authors further apply BSS using independent component analysis (ICA) to separate multiple UHF PD pulses [88]. In [89], Majidi *et al* propose a PD pattern recognition algorithm based on signal norms extracted from PRPD, SRC and ANN classifier. Although these two classifiers demonstrate similar performance, one main advantage of SRC is that it does not require parameters tuning. The authors further suggest using SRC as a feature selection or a dimension reduction method, and using ANN or SVM as the classifier. Such a combination might have better performance. In [90], Gao *et al* proposed a recognition method of unknown PD using an improved SVDD algorithm. A feature vector based on VMD and sample entropy is developed as the input of the proposed algorithm. Firstly, the tri-training algorithm is applied to train the SVDD. To further refine the boundary in the hypersphere of SVDD to reduce the misclassification rate, the Otsu algorithm is applied to set double thresholds. An improved FCM clustering based classifier is employed to further identify the specific PD types if the PD sample is classified as known type. The proposed method demonstrates the ability to recognize the unknown PD, and the potential in recognizing multi-PDs in electrical equipment. SVDD, similar to OCSVM, belongs to the one-class classification method, and it has also been proven that the solutions of SVDD and OCSVM are equivalent to normalized data representations in the kernel space [91]. The experimental results also show that SVDD and OCSVM have similar accuracy (91.67% and 90.64%), and the proposed improved SVDD significantly increases the recognition rate to 99.58%. In [92], Aizpurua *et al* apply Gaussian-BN with Duval's gas values to detect PD in the transformer. In [93], Jia *et al* apply VPMCD improved by kernel partial least square (KPLS) regression for transformer PD recognition. In [94], Peng *et al* develop a rough set theory based PD recognition method using statistical features of the HFCT PD pulse as input for ethylene-propylene rubber (EPR) cable. The main disadvantage of RS based method is that the required training and testing time is more significant than BPNN, especially with large training samples. However, it is a white box method, which allows the user to understand the decision making process. In [95], Zhu *et al* employ k-means clustering of TDOA arrays based method for separating multiple UHF PD signals in air-insulated substation. In [96], Misak *et al* introduce a PD detection technique in OH lines using singular value decomposition (SVD) and PSO. Although authors announce that proposed PSO-SVD is better than SVM, it needs further validation since the SVM uses statistical features as input instead of SVD based features. A summary of conventional ML based PD diagnosis methods is presented in Table 1.

Table 1. Summary of conventional ML based PD diagnostics.

Ref.	Application	Objective	Input (sensor type)	Sampling rate	Methodology	Key result
[13]	Elec. equip.	Classification	PRPD (HFCT, impedance, loop antenna)	N/A	Statistical features + BPNN	83.3%
[14]	Elec. equip.	Classification	Time-series data & PRPD (TEV/SCS/HFCT)	100 MS/s	Statistical features + BPNN	89.45% (waveform pattern) 95.63% (PRPD pattern)
[15]	Elec. equip.	Denoising	Time-series data (R-F)	2 GS/s	Signal segments + ANN	Superior for denoising
[16]	HVDC	Identification	Sequence PD data	N/A	Statistical features + ANN/DT NoDi* pattern + ANN	All greater than 95%
[17]	Elec. equip.	Detection Classification	Time-series data (TEV)	50 MS/s	Wavelet entropy + BPNN	97.35 detection rate 86.75% recognition rate
[18]	Cable	Classification	Time-series data (ultrasonic)	1 MS/s	DWT features + BPNN	96%
[19]	Substation	Localization	Time-series data (UHF)	2.7 MS/s	RSSI + BPNN	0.89 mean error
[20]	Substation	Localization	Time-series data (R-F)	2 GS/s	RSS + kNN/BPNN	1.72/1.68 m mean error
[21]	Transformer	Classification	Time-series data (UHF)	N/A	FFT + BPNN	Up to 97%
[22]	OH insulator	Classification	Time-series data (acoustic)	80 kS/s	Envelop + FFT + BPNN	More than 85%
[23]	GIS	Classification	PRPD (impedance)	N/A	Statistical features + PCA + BPNN	92% without PCA 88% with PCA
[24]	Epoxy resin	Detection	Time-series data (acoustic)	1 MS/s	Statistical features + PCA + BPNN/SVM	96.6%/ 96.2%
[25]	OH insulator	Classification	Time-series data (R-F)	N/A	WPD + feature selection + BPNN	Higher than 95%
[26]	Capacitor	Detection	Time-series data (HFCT)	20 MS/s	Chaos theory + extension neural network	90%
[27]	Substation	Localization	Time-series data (R-F)	2 GS/s	RSS + kNN/WkNN/MLP/GRNN	2.12/2.06/2.07/1.81 m mean error
[28]	Substation	Separation	Time-series data (UHF)	5 GS/s	Reconstructed spectra + RBFN	Separation rate > 75% (SNR > 10dB)
[29]	Substation	Localization	Time-series data (UHF)	N/A	Time delay vector + RBFNNs	0.5 m & less than 5°
[30]	Transformer	Classification	PRPS	N/A	SAE + ELM/SVM/BPNN	94.08/89.91/82.9%
[31]	Transformer	Classification	PRPS (UHF)	N/A	Statistical features + OS-ELM	91.5%
[32]	Elec. machine	Classification	PRPD (impedance)	500 MS/s	Statistical features + ELM/SVM	ELM is better
[34]	Elec. equip.	Classification	PRPD (impedance)	N/A	Statistical features + ENN	About 67-96%
[35]	Elec. equip.	Classification	PRPD (HFCT)	200 MS/s	PRPD raw data + EBA/ANN	Around 90/80%
[36]	GIS	Classification	Time-series data (UHF) PRPD (UHF)	10 GS/s 50 MS/s	Time-domain and PRPD features + BPNN + DST (multi-information fusion)	82.94% for TRPD only 92% for PRPD only 97.25% for combined
[37]	Transformer	Detection	DGA	N/A	Non-code ratio + data augmentation + MLP	85.1%-95.5% overall
[39]	Elec. machine	Classification	PRPD	N/A	Statistical features + different classifiers	SVM/MLP performs best
[40]	Transformer	Classification	PRPD (impedance)	N/A	LBP&HOG + SVM	99.3% (HOG-SVM)
[42]	GIS	Localization	Time-series data (UHF)	N/A	STFT + edge detector + SVM	100%
[43]	Transformer	Classification	Time-series data (acoustic)	1 MS/s	DWT + kNN/DT/SVM	SVM hits the best
[44]	Transformer	Classification	Time-series data (UHF)	20 GS/s	Statistical features + SVM	99.14%
[45]	GIS	Classification	PRPS (UHF)	N/A	Polar coordinate patterns + k-means clustering + SVM	98.3%
[46]	GIS	Classification	Time-series data (UHF)	N/A	Chromatic parameters + SVM	86.67%
[47]	XLPE cable	Classification	PRPD (impedance)	N/A	Statistical features + ANN/SVM	89.05-95.49/98.6-100%
[48]	Transformer	Classification	Time-series data (HFCT)	200 MS/s	PSD based features + SVM	Almost perfect
[49]	Transformer	Classification (aging status)	PRPD	N/A	Statistical features + PCA+ SVM	82.06%
[50]	GIL	Classification	PRPD (UHF & optical sensor)	N/A	PCA + LDA/kNN/SVM	Up to 91.67/93.33/95%
[51]	GIS	Severity assessment	PRPD (UHF)	N/A	Statistical features + mRMR feature selection + SVM/BPNN	Higher than 85/84%
[52]	XLPE cable	Classification	PRPD (impedance)	N/A	Photoelectric fusion pattern + PCA + ANN/SVM	Superb in noisy conditions
[53]	HVDC cable	Classification	Time-series data (impedance)	N/A	Time domain features + SVM DWT based features + SVM	Up to 99.63% Up to 100%
[54]	Elec. machine	Classification	Time-series data (HFCT)	24 kS/s	GLCT+LBP+MCSVM	87% (mixed data) 100% for most sites (individual site data)
[55]	Substation	Localization	Time-series data (R-F)	2 GS/s	RSS + LSSVM	<2.5 m error (72%)
[56]	GIS	Classification	PRPD (impedance)	N/A	PCA + OCSVM + two-stage LR	59-100%
[57]	Elec. equip.	Detection	Time-series data (HFCT)	200 MS/s	Normalized PSD + OCSVM	99.67% with domain adaptation
[58]	Elec. equip.	Classification	PRPD (impedance)	N/A	Different feature extraction techniques + different ML classifiers	Fuzzy classifiers show high classification rate
[59]	Transformer	Classification	Time-series data (UHF)	60 MS/s	EEMD-sample entropy + RVM	100%
[60]	Elec. equip.	Classification	Time-series data (UHF)	N/A	VMD-MDE + PCA + HMSVM	100%
[61]	Transformer	Classification	PRPD	N/A	2D PCA + FkNNC/BPNN/SVM	96.4/94.4/97.5%
[62]	XLPE cable	Classification	PRPD (impedance)	N/A	Fractal features + GA-SVM	Higher than 95%
[63]	Substation	Localization	Time-series data (R-F)	2 GS/s	Time domain features + feature selection + kNN	1.65 m mean error
[64]	Transformer	Classification	Time-series data (acoustic)	10 MS/s	PCA + kNN	94% (same conditions) 90% (different conditions)
[65]	HV bushing	Classification	Time-series data (acoustic)	10 MS/s	Frequency spectrum + RF/kNN/ANN/SVM/NB/Others	90.33-99.65% (SNR:20dB) 38.41-95.98% (SNR-40dB)
[66]	Transformer	Classification	Time-series data (acoustic)	10 MS/s	PSD + DT/RF/GB/SVM/LDA	SVM performs worst when there is variation in dataset

[68]	Transformer	Detection	DGA	N/A	MLP-gas concentration/FIS	Not mentioned
[69]	Elec. equip.	Classification	PRPD (impedance)	N/A	Statistical features + FIS	FIS is feasible
[70]	Elec. equip.	Classification	PRPD (impedance)	N/A	Statistical features + ANN/FL	65-98/100%
[71]	GIS	Severity assessment	PRPD (UHF)	50 MS/s	Statistical features + FL	Average of 91.75%
[72]	Transformer	Detection	DGA	N/A	Combined method+ANFIS+DST	76.7% for PD detection
[73]	Elec. equip.	Detection	Optical signal (optical sensor) temperature, and humidity	N/A	SCM-FCM + improved ANFIS	18.1% maximum relative error
[74]	XLPE cable	Classification	PRPD (impedance)	N/A	PCA + ANN/SVM/ANFIS	ANN/SVM performs best under noisy conditions
[76]	Transformer	Classification	PRPD	N/A	Statistical features + ANN/SVM/kNN/DT/RF	RF hits the best: 97.78%
[77]	Elec. equip.	Denoising	Time-series data (HFCT)	100 MS/s	Energy spectral density + DT + WT	Superior for denoising
[78]	Transformer	Classification	PRPD (impedance)	100 MS/s	Statistical features + RF	94.44%
[79]	Substation	Localization	Time-series data (R-F)	N/A	WPD + RF	1.9152 m mean error
[80]	Transformer	Classification	Time-series data (impedance)	100 MS/s	Phase information with PD pulse waveform + RF /LSSVM	85.9-92.05/70.25-73.7%
[81]	Elec. equip.	Localization	Optical signal (optical sensor)	100 MS/s	S-transform + recursive features elimination + RF	95.6%
[82]	GIS and Transformer	Separation	Time-series data (UHF/acoustic/CT)	N/A	Linear prediction cepstrum coefficient + PCA + IF + FCM	IF is effective
[83]	HV cable	Detection Recognition	Time-series data (HFCT)	100 MS/s	Various features + RF based feature selection + SVM/BPNN	100/100% for detection 90/85% for recognition
[84]	Elec. machine	Classification	PRPD (impedance)	500 MS/s	PRPD data + tensor-based classifier	Greater than 98.7%
[85]	Transformer	Localization	Time-series data (HFCT)	1 GS/s	Statistical features + MLR/ANN	Up to 70/66.67%
[86]	Elec. equip.	Classification	Average PD magnitude/number of pulses/operating temperature/ humidity/loading (PD portable analyzer console)	N/A	Designed feature vector + GMM	At least 86.3%
[87]	Transformer	Separation	Time-series data (acoustic)	1 MS/s	BSS	Successful under experimental conditions
[88]	Elec. equip.	Separation	Time-series data (UHF)	10 GS/s	ICA	Successful under experimental conditions
[89]	Elec. equip.	Classification	PRPD (PD meter LDS-6)	N/A	Signal norms + SRC/ANN	81.6-99.7/82.6-95.9%
[90]	Elec. equip.	Classification	Time-series data (impedance)	40 MS/s	VMD entropy + SVDD +FCM clustering classifier	99.58% (known/unknown) 99.17% (types)
[92]	Transformer	Detection	DGA	N/A	Duval's gas values + Gaussian BN	96.3% for PD detection
[93]	Transformer	Classification	PRPD (impedance) Time-series data (HFCT)	N/A	Statistical features + KPLS-VPMCD VMD entropy + KPLS-VPMCD	99.62% 88.33%
[94]	EPR cable	Classification	Time-series data (HFCT)	100 MS/s	Statistical features + RS theory	93%
[95]	GIS	Localization	Time-series data (UHF)	5 GS/s	TDOA + k-means clustering	0.6 m maximum error
[96]	OH line	Detection	Time-series data (inductor)	20 MS/s	SVD + PSO	61.85%

4 DEEP LEARNING BASED METHODS

Conventional ML classifiers with shallow structures require a powerful feature extractor that solves the selectivity-invariance dilemma [97], and they cannot be continually improved by increasing the size of the training data. DL, as a subset of ML techniques, is getting more attention because it requires less need for feature engineering and can achieve higher performance with the help of big data, and revolutions in algorithms and hardware. The simplest DL structure is the deep neural network (DNN), which consists of multiple fully-connected layers for deep feature extraction and a classification layer (generally the softmax layer). The mathematical representations are similar to MLP (BPNN) introduced in Section 3.1, while the number of hidden layers is generally greater than three. In [98], Catterson *et al* develop a DNN based PD classification method using UHF PRPD, which is one of the earliest works to apply DL in PD diagnostics. Each input sample consists of 50 power cycles of PRPD with a phase window size of 5.625°, resulting in a one dimensional (1D) matrix with a size of 3200×1, where the value of the matrix represents the relative magnitude of the recorded PD in each phase window. The impacts of the number of hidden neurons, number of hidden layers, and the types of activation functions, including sigmoid function and rectified linear unit (ReLU), are investigated. Based on the experimental results, ReLU activation function demonstrates excellent performance with deeper structure compared to sigmoid function. This is normally the case in DL, since ReLU can effectively mitigate the problem of gradient vanishing during the training process [99]. Furthermore, DNN with five hidden layers and ReLUs can improve the classification accuracy from 72% to 86%, compared to the shallow ANN with sigmoid activation functions.

4.1 AUTOENCODER

A simple AE consists of two parts, an encoder and a decoder. Given the input dataset $\{\boldsymbol{x}_i, y_i\}_{i=1}^{n}$with *n* samples, the encoding process and decoding process can be represented as follows:

$$f_e(\boldsymbol{x}_i) = \boldsymbol{h_i} = \sigma_e(\boldsymbol{w_e}^T \boldsymbol{x}_i + \boldsymbol{b}_e) \quad (9)$$

$$f_d(\boldsymbol{h_i}) = \boldsymbol{x}_i{}' = \sigma_d(\boldsymbol{w_d}^T \boldsymbol{h_i} + \boldsymbol{b}_d) \quad (10)$$

where the subscript *e* and *d* represent the encoder and decoder; $\boldsymbol{h_i}$ is the features extracted by encoder; $\boldsymbol{x}_i{}'$ is the reconstructed sample by decoder; $\theta = \{\boldsymbol{w}, \boldsymbol{b}\}$ and σ are the network parameters and activation function, respectively. The optimization objective of AE is to minimize the reconstruction error of input samples. Therefore, it can be optimized using the following cost function:

$$\min_{\theta_e,\theta_d} Cost(\theta_e,\theta_d) = \frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{x}-\boldsymbol{x}_i{}'\|^2 \tag{11}$$

To achieve better performance, variants of AE are proposed such as sparse AE (SAE) [100, 101], denoising AE (DAE) [102], and variational AE (VAE) [103–105]. There are typically two ways to construct AE with deep structure. The first way is achieved by stacking multiple AEs to form the stacked AE. The output from the encoder part of the first AE is used as the input to the second AE. Then, greedy layer-wise pre-training is done to establish the stacked AE. The typical process to construct a stacked AE is visualized in Figure 6. After establishing the stacked AE with pre-training parameters, a decision layer (i.e. a softmax layer) is connected, and labels can be used to fine-tune the whole structure in a supervised way to achieve classification. In [102], Wang *et al* develop a stacked DAE (SDAE) based DL methodology for PD recognition in HV cables. 18 conventional PD features (e. g. pulse width and mean voltage) and 16 wavelet features (e. g. detailed energy) are combined to form the input vector to the SDAE. The proposed method is tested using a dataset with five types of PD signals, where three types of PD signals have a high mutual similarity. This study investigates the influence of DAE layers and the size of the training samples. The proposed SDAE method can achieve a classification rate of 92.19% and improves by 5.33% and 6.09% compared to SVM method and BPNN method, respectively. To fully free the human labor in feature engineering stage, researchers start to use raw data as input to construct end-to-end diagnosis method. In [100], Tang *et al* employ stacked SAE (SSAE) to assess the PD severity in GIS using UHF PRPD raw data as input (1D matrix with a size of 360×1). The impacts of number of layers of SAE, number of training samples, and number of neurons in the hidden layer of the last SAE on classification accuracy are investigated in detail. The proposed DL based method is compared to SVM classifier with statistical features of PRPD and demonstrates a better ability in feature learning and classification. The performance achieves the best when the number of SAEs is 3 or 4. Furthermore, the number of neurons in the hidden layer of the last SAE seems to have less impacts on the accuracy based on the experimental study. In [101], Duan *et al* develop an end-to-end stacked SAE based PD classification method, where PD current waveform captured by HFCT is used as the input to the proposed DL-based framework. The impacts of sparsity parameter, number of hidden nodes, depth of the framework, and activation functions, on PD classification accuracy are investigated to find the optimal sets of hyperparameters. Based on the experimental results under lab environment, single SAE combined with a softmax classifier is sufficient to achieve satisfactory performance. Conventional ML classifiers (e. g. ANN and SVM) with raw data and extracted features by PCA and t-SNE are investigated, respectively. As expected, conventional methods are less effective than the proposed DL method. The authors state the PRPD analysis (UHF and IEC 60270 method) acts as a feature extraction technique to construct distinguishable features, which is contradictive to the initial idea of DL: automatically extracting features without human expertise. However, more investigation and comparative studies should be carried out.

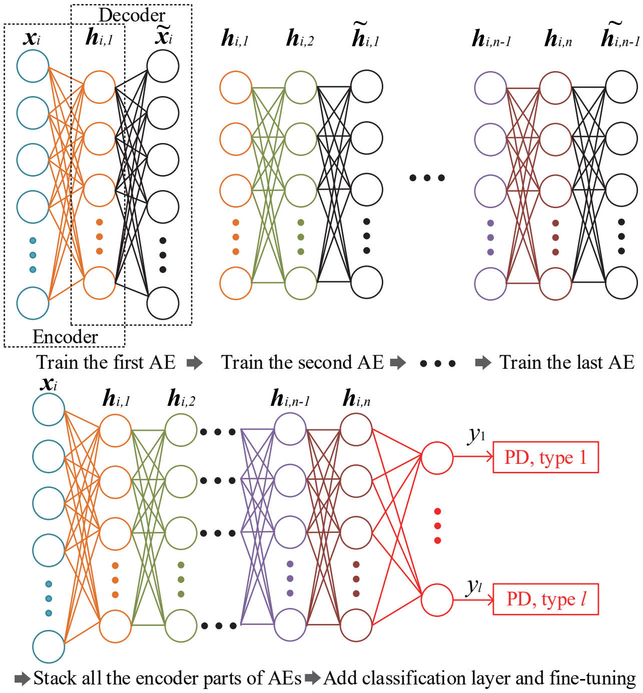


Figure 6. Diagram of a stacked AE for PD classification.

The other way to construct AE with deep structure is replacing a single layer with multiple layers in both encoder and decoder parts. In [105], Dai *et al* develop a data matching method for PD classification in GIS using VAE and raw UHF PRPS data. The VAE is firstly trained using both experimental data and substation field data (collected from more than 30 substations in China with 42 different cases of power equipment defects) in an unsupervised manner. To validate the effectiveness of data matching, each test sample, which is only from the substation field dataset, is mapped to the latent space through the encoder of the trained VAE. By calculating the cosine distance between two eigenvectors from two different PD data, the matching degree can be estimated. In this way, similar cases can be identified using data mining on historical PD databases, and PD diagnosis and equipment status evaluation can be performed. Based on the experimentally validated results on the complex field dataset, DL based methods such as CNN, DBN, and proposed VAE, show better identification rate of different PD types compared to conventional data matching method with statistical features. Among three DL based methods, the proposed VAE performs the best. In [103], Zemouri *et al* aim to solve the following problem: although the performance of the ML based PD diagnostics can be improved by using sufficient labeled data, one of the main problems that all industries face is the massive high dimensionality unlabeled data. Labeling all collected data is not an effective solution, because it is time-consuming and requires significant labor resources. Then, a PD classification framework using convolutional VAE based DL model with minimum labeled PD data is developed for hydrogenators. The encoder part of convolutional VAE is primarily used as a deep feature extractor and a DNN with few fully-connected layers and a softmax layer is used for classification. One major challenge in this research is how to select the most significant PD data for labeling and how to determine the minimum data size for training. Authors find DNNs cannot always find the best features from the randomly selected limited amount of labeled data. To address this issue, expert knowledges are integrated using handcrafted feature extraction. Besides acts as a deep feature extractor, the encoder part of convolutional VAE is also used as a visualization tool to identify the conflict zones in the feature space. Then, new samples around these conflict zones are selected by the expert for labeling, and a new training iteration starts using the previous training dataset with additional labeled data. Several iterations are performed until the conflict zone is minimized as acceptable by the expert, and the optimal set of training samples can be determined. However, the proposed method is somehow contradictive to the basic idea of DL: the performance of the framework heavily depends on the expert knowledge. Besides being used for classification, VAE can also be used for data augmentation [104].

4.2 CONVOLUTIONAL NEURAL NETWORK

As shown in Figure 7, a convolutional neural network (CNN) model typically consists of a convolutional operators-based feature extractor, fully-connected layers for higher level reasoning, and a classification layer. The computational complexity of convolutional layers is less compared to the fully-connected layers in terms of required matrix multiplication operations because of the configuration method, where each neuron in a convolutional layer is only connected to a small set of neurons in the following layer. Also, such configuration makes CNNs excellent in extracting regional characteristics of the input sample. For an input matrix $\boldsymbol{x}_i^{j-1}$ with P channels from the previous layer, K filters with the size of $H_f \times L_f$, and a step size s=1, the convolutional operation of k^{th} filter at j^{th} layer can be represented as follows:

$$\begin{aligned}(\boldsymbol{x}_i^j)_{h_o,l_o,k} &= \sigma(\boldsymbol{x}_i^{j-1} * \boldsymbol{w}_k^j + b_k^j) \\ &= \sigma(\sum_{p=0}^{P-1}\sum_{h_f=0}^{H_f-1}\sum_{l_f=0}^{L_f-1}(\boldsymbol{x}_i^{j-1})_{s\times h_o+h_f, s\times l_o+l_f,p} \\ &\times (\boldsymbol{w}_k^j)_{h_f,l_f,p} + b_k^j)\end{aligned} \tag{12}$$

where $\boldsymbol{x}_i^j$ denotes the output feature map at j^{th} layer with the size of $H_o \times L_o \times K$; $h_o = \{1, \ldots H_o\}$, $l_{ip} = \{1, \ldots L_o\}$, $k = \{1, \ldots K\}$ represent the row, column, depth index of the output feature map, respectively; $\boldsymbol{w}_k^j$ and b_k^j are the weight matrix and bias coefficient of k^{th} filter in j^{th} layer, respectively; σ denotes the activation function, which is typically ReLU for DNN since it can mitigate the gradient vanishing problem [99]. After each CNN layer, a pooling layer is usually used to achieve dimension reduction. Max pooling (MaxP) layer is the most common pooling layer, which can be represented as follows:

$$\begin{aligned}&(\boldsymbol{x}_i^j)_{h_o,l_o,k} \\ &= max((\boldsymbol{x}_i^{j-1})_{h_o:h_o+H_{\mathrm{MaxP}}-1, l_o:l_o+L_{\mathrm{MaxP}}-1,k}))\end{aligned} \tag{13}$$

where the size of the max operator is $H_{\mathrm{MaxP}} \times L_{\mathrm{MaxP}}$, the operation step size is 1, the size of the output feature is $H_o \times L_o \times K$. Then, the hierarchical features can be found by stacking several CNN layers and pooling layers. Next, the last pooling layer is flattened to 1D vector and connected to fully-connected layers for further reasoning. Finally, a classification layer (i.e. softmax layer) is connected at last to map the input sample into the target class.

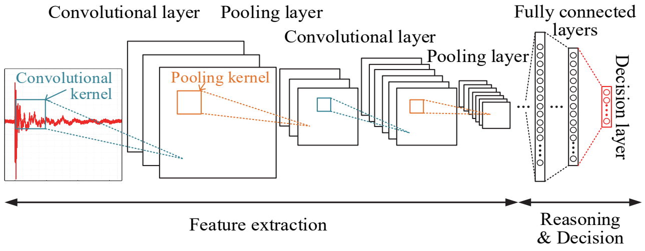


Figure 7. General structure for a convolutional neural network.

CNN is initially proposed to solve image recognition problems, where the input samples are images (2D matrices). In [106], Li *et al* make use of computer vision technique based on a deep capsule network for fault detection (corona, ageing, local hot spot) in power system equipment. Three images are captured by an ultraviolet (UV) camera, an infrared camera, and a visible light camera. A fused image is then constructed by the capsule network model, which includes key information of UV intensity, temperature, and physical information of the power equipment under monitoring. Then faults occur in the power equipment can be detected and displayed accurately in the constructed fused image. In [107], the authors state in some circumstances for PD on-site detection techniques, due to the limitation of the portable detectors (limited storage speed and space), detected samples are usually stored as unstructured datasets such as images. In [108], Puspitasari *et al* investigate the performance of CNN with 2D images (the segmented RBG color image of the PD pulse is converted to grayscale images for input) of the time series signals capture by different sensors in PD classification. An average rate of 92%, 84.5%, and 99.6% are obtained using waveforms from TEV sensor, SCS, and HFCT, respectively.

To directly apply 2D CNN, researchers also use PRPS as input: In [109], Yang *et al* propose a CNN based PD recognition method using PRPS. The PRPS raw data is converted to an RBG color image as the input of CNN. The best classification accuracy of the proposed method is 91%. To improve the property of generalization of CNN, a data augmentation method is adopted: the PRPS raw data is superimposed with environmental interferences; then, the superimposed data is converted to RBG image and superimposed with Gaussian noise to generated augmented data. With the help of data augmentation, the best classification rate improves to 97.58%. In [110], Song *et al* develop a CNN based method for PD recognition in GIS. The input is UHF PRPS raw data, where the *x*-axis, *y*-axis, and the value of the matrix represent the phase, period, and the magnitude of the PD pulse, respectively. The initial parameters of CNN are obtained by carrying out unsupervised learning to an AE model. After that, the encoder part of the AE is connected to two fully-connected layers and a softmax layer to construct the diagnosis model, and fine-tuned by training samples. According to the experimental results under the lab environment, the proposed method can improve classification accuracy compared to conventional ML based methods based on SVM and BPNN, and the improvement is more considerable when using more training samples. Authors also use a complex dataset to evaluate the proposed method as well as conventional ML methods, where the samples are collected from the live GIS in more than 30 substations using two types of portable PD detection devices. The classification accuracy of CNN decreases by 8.9%, while that of SVM and BPNN reduce by 17% and 20.4%. The results demonstrate the robustness of CNN in handling diverse data.

It is impossible to directly use 1D time series signal as input to 2D CNN since there is a mismatch in dimension between the input data and input layer. Therefore, researchers introduce several ways to convert the 1D data to 2D data. The first way is creating 2D binarization images using original 1D time series signals. In [104], Wang *et al* propose an end-to-end framework using CNN and raw data of the UHF signal for PD recognition in GIS. Instead of using large CNNs such as AlexNet and VGG16, a light-scale CNN (modified from LeNet5) is proposed to optimize the time and diagnosis performance of CNN model. The input is a 64×64 image shrunk from a TRPD single-channel banalization image with the size of 600×438 using the down-sampling technique. Those time-series data come from an experiment under laboratory environment and simulation using the finite-difference time-domain (FDTD) method. Furthermore, data augmentation is performed by using a conditional VAE. Because both simulation and data augmentation can enlarge the training dataset with improved diversity, the property of generalization of the CNN model can be improved. Compared to conventional ML methods using statistical features in [71] as input, including SVM, BPNN, and DT, the proposed method can achieve better classification accuracy when the number of training samples is greater than 500. When only limited samples are available for training, SVM and DT show better performance. AlexNet and VGG16, only achieves 90.63% and 86.41%, compared to 98.13% with the proposed method when the number of training sample is 2560. This is mainly because the performance of the complex CNN model is limited by the size of the dataset. Another comparison is conducted between the modified CNN and original LeNet5. The main difference between LeNet5 and modified CNN is the input size: 64×64 versus 28×28. LeNet5 is initially proposed for character recognition, and the input samples are relatively simple. However, UHF PD signals are much more complicated than character because of the stochastic nature of PD. Theoretically, shrinking the original image to a smaller size may result in the massive loss in useful information of PD characteristics. Although the model size increases with larger input size, the proposed method outperforms LeNet5 with a significant margin (98.13% compared to 75.04%) based on the experimental results. Therefore, one must take the input size into account to achieve a reasonable balance between the model size and accuracy when designing the model structure for different applications. In [111], The same author publishes another paper regarding GIS PD classification using the similar methodology in [104] but with MobileNets (the input image is shrunk from 600×438 to 224×224). The depth-wise separable convolutions and reverse residual structure can mitigate the gradient vanishing problem during the training of a model with a deeper structure. Similarly, a 96.5% recognition rate is achieved on the testing dataset, outperforming other CNN structures such as VGG16 and Lenet5. Overall, the two proposed frameworks can achieve improved classification accuracy and reduced computation time, which can facilitate the real-time deployment of the DL models in resource-constraint IoT edge devices. However, more investigations are needed: for example, the parameter number of the MobileNets in [111] is 2.24 million, the required storage space is 12.8MB, and the sampling rate of the input samples is 10 GS/s. Although the computational burden is greatly reduced, the feasibility of the proposed methodology for PD classification is subject to further online real-time validation experiments.

Secondly, inspired by music recommendation and speech recognition applications, time-frequency analysis is performed, which can easily convert original 1D data into the 2D presentation (i.e. spectrogram). In [112], Che *et al* propose a PD recognition method for XLPE cable based on CNN and acoustic signal captured by optical fiber distributed acoustic sensing (FDAS) system. The original 1D time domain signal is converted to 2D spectral frame representation using mel-frequency cepstrum coefficients (MFCC) analysis. The proposed MFCC-CNN achieves a 96.3% classification rate on a dataset consisting of internal, corona, surface PD, and noise. Although the results are promising, a comparison between MFCC based 2D features and other types of 2D features is not conducted. In [113], Lu *et al* develop a CNN based PD detection method for switchgear and GIS using a spectrogram of TEV signal extracted by joint time-frequency analysis. The proposed method demonstrates good capability of noise-rejection and achieves around 95.73% detection rate without applying any de-noising techniques. In [114], the bispectrum of the HFCT signal calculated by higher-order statistics is used as the input of ResNet-34 for EMI classification in HV equipment. Classification rates of 80.83%, 92.87%, and 80% are achieved for corona, PD, and minor PD, respectively. In [115], the same authors also conduct a preliminary research on an EMI classification method using complex bispectrum of HFCT signal and deep complex CNN. 88.33%, 96.67%, and 74.17% classification rates are obtained for corona, PD, and minor PD, respectively. In [116], Li *et al* propose a CNN based PD recognition method using the UHF signal with STFT as a pre-processing method for GIS. The proposed method outperforms some conventional ML methods such as SVM with HHT-entropy and SVM with DWT-entropy based on comparative studies. Authors state STFT offers a perceptual intuition expression that has clear physical meaning without much information loss. However, for these kinds of methods, the results can be affected by the types of time-frequency transformations and the selection of parameters such as decomposition level for DWT and window size for STFT. The impacts of the window size of STFT on CNN are unknow and not investigated in [116].

Thirdly, multiple 1D time-series signals (e.g. collected from multiple sensors or multiple channels) can be stacked to form the 2D data. In [117], Banno *et al* use TEV and CNN for PD classification in switchgear. A simple pre-processing method is adopted (downsampling, scaling, random shifting and inverting, and pulse extracting) to form a 2D input consisting of 2 channels (max and min). After that, CNN is applied for feature extraction and classification. In [118], Wang *et al* propose an SF6 gas on-line monitoring method based on CNN in GIS. The physical index measurements over time of the chamber are obtained by pressure, temperature and infrared photoacoustic gas sensors, and used as input to a CNN classifier. The proposed method outperforms the simple threshold and SVM classifiers. However, the implementation details of CNN are not given. Besides the three aforementioned conversion methods, reshaping is another method which is one of the most popular 1D-to-2D conversion methods in fault diagnosis in other fields, such as bearing fault diagnosis and photovoltaic fault diagnosis [119, 120].

Recently, 1D-CNN is proposed to directly bridge the relationship between 1D time-series signals and classification results. In [121], in order to directly extract information from 1D time-series signal even without any simple arrangements, Khan *et al* propose an end-to-end framework based on 1D-CNN with original HFCT PD pulse waveform for PD detection. Based on the experimental results, the proposed method is better than conventional ML methods using SVMs (linear kernel and RBF kernel) with handcrafted features and achieves 97.38% and 93.23% on simulated PD pulses and real PD pulses, respectively. In [122], Woon *et al* develop a 1D-CNN based transformer PD classification method using raw acoustic emission signals. The proposed method is compared to conventional ML classifiers with FFT based features presented in previous work in [66]. When there are no variations between training samples and testing samples, the proposed 1D-CNN based method performs slightly worse than SVM and RF based methods. On the contrast, when variations existed (i.e. different locations and oil temperature), the proposed method demonstrates the superior property of generalization, and outperforms conventional ML based methods with a significant margin in some cases. It implies 1D-CNN can extract more common features from the raw signal, and the conventional ML based methods with handcrafted features are easier to overfit on the training dataset.

There are other types of applications using CNN with feature extraction techniques. In [123], Zhang *et al* investigate DL based ultrasonic PD identification in the transformer. A feature vector consists of 193 handcrafted features is used as input to regular RNN, DNN, and CNN models, and achieves 89.4%, 91.2%, and 93.9%, respectively. However, more details of the models used in this study need to be presented in order to make a fair comparison among those DL methods. In [124], Peng *et al* develop a 1D-CNN based DL methodology for PD recognition in HV cables. Similar to [102], 17 conventional PD features (e. g. pulse width and mean voltage) and 16 wavelet features (e. g. detailed energy) are combined to form the input vector to the CNN. The proposed method is tested using a dataset with five types of PD signals, where three types of PD signals have high mutual similarity. The hyperparameters (number of layers of CNN, kernel size, activation functions, and pooling layers) are also carefully studied and selected, and the optimal structure consists of 3 convolutional layers, three pooling layers, and one fully-connected layer for classification. Although the optimal set of hyperparameters is chosen, the proposed CNN based method only improves the classification rate, which is 92.57%, by 6.47% and 4.76% compared with BPNN and SVM. In [107], Wan *et al* propose a 1D-CNN based PD recognition method for GIS. UHF time-series signal with 100ms duration is saved in terms of RBG image, and then the greyscale images of every PD pulse are extracted from the original image after several preprocessing steps. To further reduce the complexity of the model, 1D-CNN is employed, and the 1D input array is obtained by summing each column of the 2D matrix of the extracted image. The proposed method can achieve an average recognition rate of 88.9% on an imbalance dataset compared to 70% and 67.2% obtained from SVM and BPNN. However, the performance of those methods is highly affected by the PD feature construction.

4.3 RECURRENT NEURAL NETWORK

The links of a recurrent neural network (RNN) between nodes form a directed graph along a temporal sequence, which makes RNN capable of exploring the dynamic behavior of time-series data. Long short-term memory (LSTM) model, consisting of many LSTM blocks, is one of the best-performing and most popular RNNs. The most brilliant part of LSTM is that it introduces an internal recurrence besides the outer recurrence in traditional RNNs. With such modification, the LSTM model is easier to train since the gradient can flow for long durations [67]. An LSTM block contains several units to control the flow of information, including a state unit $\boldsymbol{s}_i^t$, a forget gate unit $\boldsymbol{f}_i^t$, a external input gate unit $\boldsymbol{g}_i^t$, and a output gate unit $\boldsymbol{q}_i^t$ for a time step t, layer index i, and the current input vector $\boldsymbol{x}_i^t$. Then, the mathematical representation of a LSTM block can be formulated as follows:

$$\boldsymbol{f}_i^t = \sigma_{sm}(\boldsymbol{u}_i^{f^T}\boldsymbol{x}_i^t + \boldsymbol{w}_i^{f^T}\boldsymbol{h}_i^{t-1} + \boldsymbol{b}_i^f) \tag{14}$$

$$\boldsymbol{s}_i^t = \boldsymbol{f}_i^t\boldsymbol{s}_i^{t-1} + \boldsymbol{g}_i^t\sigma_{tanh}\left(\boldsymbol{u}_i^{s^T}\boldsymbol{x}_i^t + \boldsymbol{w}_i^{s^T}\boldsymbol{h}_i^{t-1} + \boldsymbol{b}_i^s\right) \tag{15}$$

$$\boldsymbol{g}_i^t = \sigma_{sm}(\boldsymbol{u}_i^{g^T}\boldsymbol{x}_i^t + \boldsymbol{w}_i^{g^T}\boldsymbol{h}_i^{t-1} + \boldsymbol{b}_i^g) \tag{16}$$

$$\boldsymbol{h}_i^t = \sigma_{tanh}(\boldsymbol{s}_i^t)\boldsymbol{q}_i^t \tag{17}$$

$$\boldsymbol{q}_i^t = \sigma_{sm}\left(\boldsymbol{u}_i^{q^T}\boldsymbol{x}_i^t + \boldsymbol{w}_i^{q^T}\boldsymbol{h}_i^{t-1} + \boldsymbol{b}_i^q\right) \tag{18}$$

where $\boldsymbol{h_i}$, $\boldsymbol{w_i}$, $\boldsymbol{u_i}$, $\boldsymbol{b_i}$are the current hidden layer vector (it is also the output for the current LSTM unit), recurrent weights, input weights, and biases for i^{th} layer of LSTM block, respectively; the superscripts *s*, *f*, *g*, and, *q* indicate the correspondence of the parameters to different units; σ_{sm} and σ_{tanh} are the sigmoid and tanh activation function, respectively. An illustration of an LSTM unit is also provided in Figure 8. After the information over time is obtained by the LSTM model (consisting of several LSTM modules), fully-connected layers are used to reason the output of the LSTM model in many-to-one mode or in many-to-many mode. Finally, a softmax layer is connected at last to achieve classification.

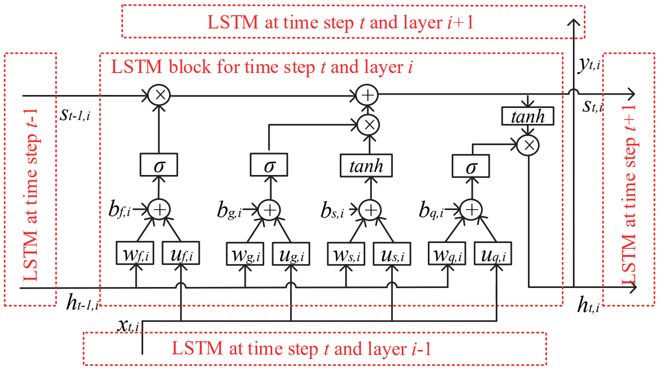


Figure 8. Diagram of a LSTM block in a LSTM model.

Similarly, researchers try to directly apply LSTM to automatically extract useful features from raw data sequences such as time-series data or time sequence of PRPDs. In [125], Adam *et al* conduct a preliminary investigation on transformer PD classification using raw data of PD pulse waveforms and the LSTM model. Even though the LSTM model is not well-designed, it is able to correctly classify PD with an accuracy of 97.04%, which is slightly lower than that of an RF classifier with handcrafted features. The capability of the LSTM model needs to be further investigated in more complex data sources. In [126], Nguyen *et al* develop an LSTM model based PD classification for GIS using series of one-power-cycle-PRPD as input. The input is a 2D PRPD matrix with a size of $M \times N$, where M is the number of power cycles, N is the number of data points in one-power-cycle-PRPD, and the value of the matrix represent the relevant amplitude of PD. Then, a LSTM model (many-to-one) is established with M time steps and L LSTM layers, where the input vector for i^{th} time step is the one-power-cycle-PRPD in i^{th} power cycle. Therefore, important temporal features are directly extracted from PRPD raw data by the LSTM model. Authors conduct multiple experiments with different set of hyperparameters and find that $M = 60$ and $L = 2$ can achieve satisfactory classification accuracy (96.72%). Proposed LSTM model also outperforms conventional ML methods such as shallow ANN with the same 2D PRPD matrix and SVM (linear kernel and RBF kernel) with handcrafted features. To enhance the generalization performance of the LSTM model in the scenarios of a small-size imbalanced dataset, simple data augmentation is achieved by slicing the experimental data with overlap. However, bias effects are still quite obvious in the classification results: the recognition accuracy for corona, floating, particle, and void are 97.04%, 79.54%, 93.18%, and 99.94%, respectively; the original experimental number of experiments for them are 94, 35, 66, 242, respectively. Therefore, more investigations can be done using other advanced data augmentation techniques.

Inspired by the applications of human pose recognition in videos, the combined model of CNN and LSTM is proposed for PD classification, where CNN can extract important regional features from the 2D matrix, and LSTM can further explore the temporal information of those extracted key features. In [127], the same authors from [116] further develop a hybrid DL structure based on CNN and LSTM for PD recognition in GIS using multi-spectrogram of UHF signals. As well known, STFT can provide both time and frequency information of signals, and the window length should be carefully selected to achieve a good balance between time and frequency resolution [128]. The smaller window size highlights the fast-changing signals (e.g. pulses), while the larger window size emphasizes the frequency information. To maximize the usage of the time-frequency information, three spectrograms with higher time resolution, higher frequency resolution, and medium resolution are calculated with different time windows, respectively. For each spectrogram, there is a CNN with specified, designed filters to extract the more in-depth features: row filters are used to extract temporal information from high time resolution spectrogram; column filters are used to extract frequency information from high frequency resolution spectrogram; 2-D filters are used to explore the time-frequency relationship in medium resolution spectrogram. The outputs from those CNNs are flattened and concatenated, and then attached to an MLP to form a multi-column CNN. For the GIS system, the characteristics of UHF signals can be affected by the propagation and reflections. Therefore, sensors at different locations can provide more information to reveal the important patterns of different types of PD over time. To further maximize the diagnosis accuracy, an LSTM network is used to analyze the output of the multi-column CNN from four sensors and make the classification. Based on the validation experiment on a dataset consisting of 13 defect modes with different types of PD and relative angles between the defect position and sensors, the proposed methodology achieves 97.51% in recognizing both PD types and locations and 98.2% in recognizing the PD types only. In addition, authors extensively compare their method with conventional ML methods, and improvement can be obtained by the proposed method with a large margin. The main concern about this method is the computational burden for real-time implementation, since the sampling rate is high (10 GHz) and the time duration (100 ns) is small. Therefore, optimization of the model structure and efficient real-time implementation can be considered. In [129], Zhou *et al* combine CNN and LSTM model to explore the useful spatial information of UHF PRPDs over time for PD classification in transformers. For each time, the input of a CNN is a 2D PRPD matrix at that particular time step, where the *x*-axis represents the phase, the *y*-axis represents the relative amplitude of PD, and the value of the matrix stands for the number of discharges. After extracting the spatial information, the output of the CNN is connected to an LSTM module at that particular time step. The outputs from all LSTM modules are connected to a fully-connected layer (many-to-many), which is followed by a softmax layer for classification. The number of LSTM nodes is 100 in this study. Based on the experimental results on a dataset with four typical insulation defect models collected in an oil-immersed power transformer, better performance can be obtained by the proposed CNN-LSTM structure compared to CNN and LSTM.

Other researchers also explore the combination of handcrafted features and the LSTM model. In [130], Dong *et al* propose a PD detection based on time-series decomposition and LSTM model for aerial covered conductors. Seasonal and trend decomposition using Loess (STL) is applied to extract the trend, seasonal, and residual features of the voltage signals in one power cycle with several windows, where residual component consists of important information about the noise. Then, three handcrafted features of residual component (number of peaks, the sum of absolute peak heights, and standard deviation of absolute peak heights) are extracted from each window to form the input vector of the LSTM model. Both oversampling and noise reduction techniques are applied. Based on the experimental study, the optimal number of windows is 4 (each window corresponds to data of 1/4 power cycle), and the detection accuracy for a 4-time step LSTM classifier is 78.76%. The proposed method also outperforms the conventional ML classifiers, such as BPNN and SVM, with the same input vectors. The proposed method is validated using a public dataset (voltage signals of the stray electric field along the aerial covered conductors measured by a simple meter) shared by ENET Centre in the Czech Republic on Kaggle. In [131], Balouji *et al* propose a PD classification method for power electronics applications using the LSTM model (many-to-one) and features extracted from PD pulses in PWM waveform within a predefined PD cycle (5 ms in this study). Similar to the structure in [126], the input to each time step is replaced by the feature vector (5×1) extracted from time domain signal of a single PD pulse, consisting of useful information of the PD shape such as maximum amplitude and time duration. The time step and number of LSTM layers for the proposed many-to-one LSTM model are 20 and 3, respectively. Therefore, the size of the input matrix for the LSTM model is $5 \times n$ for each PD cycle. The number of PDs might vary from cycle to cycle, to ensure the size of the feature vectors is fixed, zero-padding is applied. A large PD dataset (21284 PD pulses and five different types) is used to evaluate 23 ML algorithms, including the proposed DL method and 22 conventional ML classifiers. 98.3% classification accuracy can be obtained by the LSTM model, and the highest accuracy among all conventional ML methods with data vector (for non-recursive ML method, the input is reshaped to $5 * n \times 1$) is 95.5% achieved by ensemble bagged DT.

4.4 DEEP BELIEF NETWORK

The restricted Boltzmann machine (RBM), as illustrated in Figure 9, is the fundamental unit to construct the deep belief network (DBN). RBM is an energy-based model with the joint probability distribution specified by its energy function:

$$P(\boldsymbol{v},\boldsymbol{h}) = \frac{1}{Z} exp(-E(\boldsymbol{v},\boldsymbol{h})) \tag{19}$$

Z and $E(\boldsymbol{v},\boldsymbol{h})$ are the partition function for normalization purpose and the energy function for an RBM shown as follows:

$$E(\boldsymbol{v},\boldsymbol{h}) = -\boldsymbol{b}^T\boldsymbol{v} - \boldsymbol{c}^T\boldsymbol{h} - \boldsymbol{v}^T\boldsymbol{w}\boldsymbol{h} \tag{20}$$

$$Z = \sum_{\boldsymbol{v}}\sum_{\boldsymbol{h}} exp\{-E(\boldsymbol{v},\boldsymbol{h})\} \tag{21}$$

where $\boldsymbol{v} = \{v_1, \dots v_j\}$ and $\boldsymbol{h} = \{h_1, \dots h_k\}$ are visible binary units and hidden binary units; $\boldsymbol{b}$ is the bias vector for visible layer; $\boldsymbol{c}$ is the bias vector for hidden layer; $\boldsymbol{w}$ is the weight matrix. Z is intractable, requiring significant computation efforts to compute. However, the conditional distribution, which is straightforward to calculate because of the bipartite graph structure of the RBM, can be used to evaluate Z with Gibbs sampling [132]. The conditional distribution of the k^{th} hidden unit and j^{th} visible unit can be calculated as follows, respectively:

$$P(h_k = \mathbf{1}|\boldsymbol{v}) = \sigma_{sm}(c_k + \boldsymbol{v}^T\boldsymbol{w}_{:,k}) \tag{22}$$

$$P\big(v_j = \mathbf{1}|\boldsymbol{h}\big) = \sigma_{sm}(b_j + \boldsymbol{w}_{:,j}\boldsymbol{h}) \tag{23}$$

where σ_{sm} is the sigmoid activation function. With the input dataset $\{\boldsymbol{x}_i, y_i\}_{i=1}^{n}$ with n samples and corresponding labels, the RBM is optimized with the following maximum likelihood estimation with contradictive diverse learning algorithm [132]:

$$\max_{\boldsymbol{w},\boldsymbol{b},\boldsymbol{c}} LL(\boldsymbol{w},\boldsymbol{b},\boldsymbol{c}) = \frac{1}{n}\sum_{i=1}^{n} \log\big(P(\boldsymbol{x}_i;\boldsymbol{w},\boldsymbol{b},\boldsymbol{c})\big) \tag{24}$$

The trained RBM can define the parameters of the first layer of the DBN. Then, the second RBM can be trained using the hidden layer of the first RBM as the visible layer. By stacking the RBMs, a DBN can be established. Finally, the parameters are taken from the trained DBN and used to initialize an MLP. After that, a classification layer is added (typically a softmax layer), and discriminative fine-tuning is performed so that MLP can be used for a classification task.

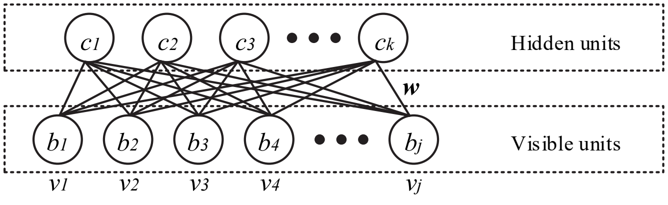


Figure 9. Basic structure of the restricted Boltzmann machine.

In [133], Karimi *et al* develop a new PD classification method based on DBN and PRPD raw variables. The PRPD raw variables, including the maximum charge, average charge, and the total numbers of PDs, are calculated along the 360°AC power cycles with different size of phase window lengths ($1^{\circ}, 3^{\circ}, 5^{\circ}, 6^{\circ}, 8^{\circ}, 10^{\circ}$). Given W phase windows per power cycle and C power cycles, a PRPD matrix with $3 \times W$ columns and C rows is created. In order to compare with the methods using feature extraction before feeding into DBN, three features including statistical features, signal norm-based features [89], and combined of them are prepared. Based on comprehensive numerical results, DBN with PRPD raw variables (8°) performs the best in different scenarios, and DBN with conventional features achieve worse classification accuracy compared to DBN with raw data. Two popular conventional ML classifiers, ANFIS and SVM, are also used for comparison. It is found that conventional ML classifiers cannot achieve satisfactory accuracy using PRPD raw variables. This is normally the case, since conventional structures are too shallow so that the classifiers are not able to extract hierarchical features and handle complex input samples. When the feature extraction is performed before feeding into ANFIS and SVM, their accuracy dramatically increases. However, DBN with extracted features is still better than ANFIS and SVM in average performance. PD classification studies under noisy conditions are also performed, and the proposed DBN method (without feature extraction) can achieve reasonable classification accuracies (around 90%) even though $SNR = -24.8$ dB, while SVM fails to work properly. It is suggested to apply de-noising techniques when PD signals are contaminated with intense noise. Similar results can be found in [134] by the same authors: DBN based method outperforms the conventional AI based methods such as DT. In [135], Dai *et al* propose a DGA based method to classify the transformer defects using non-code ratio and DBN. The results show that the proposed method can achieve 91.2% and 96% PD detection accuracy for single-fault and multiple-fault datasets, respectively. It also outperforms the conventional ML methods such as SVM and BPNN with the same input.

4.5 OTHERS

In [136], Guan *et al* propose a deep-RF based transformer PD recognition method using time-series signals collected by HFCT. Based on experimental results, deep-RF have better classification accuracy compared to ensemble RF classifier, ENN classifier, and SVM classifier. It also investigates the impact of the ratio of training sample to test sample: the proposed deep-RF classifier experiences a 3.95% decrease from 98.39% to 94.44% when the ratio changes from 0.7:0.3 to 0.3:0.7.

Generative adversarial network (GAN), as an exciting recent innovation in DL, was initially proposed by Goodfellow *et al* in 2014 [137]. A GAN typically consists of two deep neural networks, a generator (G) and a discriminator (D). The main objective of G is to generate fake samples to fool D, while that of D is to distinguish fake samples from the real samples as many as possible. A two-player minmax problem can be formulated, and the optimization objective can be written as:

$$\min_{G}\max_{D} L_{GAN}(D,G) = \mathbb{E}_{\boldsymbol{x}\sim P_{real}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim P_{\boldsymbol{Z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))] \tag{25}$$

where $\boldsymbol{x}$ and $\boldsymbol{z}$ are samples from the real data distribution $P_{real}(\boldsymbol{x})$ and a noise distribution $P_{\boldsymbol{Z}}(\boldsymbol{z})$, respectively; $D(\cdot)$ is the output from D, which is the probability of a sample being "real"; the output from G, $G(\cdot)$, is a fake sample generated from the latent space; $\mathbb{E}(\cdot)$ denotes the expectation operator.

In [138], Wu. *et al* employ GAN to synthesize UHF PD pulses for data augmentation. By double the training samples (the ratio of the original number of PD samples to the number of augmented PD samples is 1), the classification accuracy of SVM classifier and LR classifier increase by 0.52% and 1.72%, respectively. In [139], similarly, Wang *et al* employ GAN to generate more synthesized UHF PD pulses for data augmentation. The PD classification accuracy improves from 94.33% to 95.67% by doubling the training dataset with generated samples. Besides generating 1D time-series data, some researchers also use GANs to create 2D samples such as spectrogram and reshaped time-series data. J. Ardila-Rey *et al* employ deep convolutional GAN (DCGAN) to generate high-quality artificial PD samples (every input PD signal is reshaped to the 2D matrix), and the temporal and spectral behavior of generated PD samples are proved to be similar to that of real samples from experiments [140]. In [141], DCGAN is used by Petri *et al* for data augmentation in PD classification in GIS. DCGAN is widely used for directly synthesize 2D images in applications such as image recognition, and demonstrates better properties compared to original GAN based on DNN. Then, spectrograms (after removing the lower frequencies content) of original UHF signals (three power cycles) are used as the input of DCGAN, and 4 DCGANs are trained for four different types of defects. A training dataset with 400 generated spectrograms (100 samples for each class) is used to train a CNN classifier, and evaluation on both spectrograms of original data for training DCGANs and unseen data gives correct classification results. However, training GANs is not easy and suffer from many problems such as mode collapse, gradient vanishment, and non-convergence. Although DCGAN with proper structures can mitigate the abovementioned problems, other types of GAN, such as Wasserstein GAN (WGAN) [142], are still worthwhile to be investigated because of the excellent training stability. A summary of DL based PD diagnostics methods is presented in Table 2.

5 DISCUSSION AND RECOMMENDATIONS

Although ML methods have been demonstrating promising results and excellent performance in PD diagnostics, there are still several barriers preventing them from being applied for practical applications. Several key aspects, along with the potential solutions and future directions are discussed below:

Table 2. Summary of DL based PD diagnostics.

Ref.	Application	Objective	Input (sensor type)	Sampling rate	Methodology	Accuracy
[98]	Elec. equip.	Classification	PRPD (UHF)	N/A	PRPD data + CNN	86%
[100]	GIS	Severity assessment	PRPD (UHF)	5 GS/s	PRPD data + SSAE	92.2%
[101]	Elec. equip.	Classification	Time-series data (CT)	1 GS/s	Raw data + SSAE	99.7%
[102]	HV cable	Classification	Time-series data (HFCT)	N/A	Handcrafted features + SDAE	92.19%
[103]	Elec. machine	Classification	Discharge rate vs. Amplitude (PDA)	N/A	Handcrafted features + convolutional VAE	65%
[104]	GIS	Classification	Time-series data (UHF)	10 GS/s	Raw data + conditional VAE for data augmentation + light-scale CNN	98.13%
[105]	GIS	Classification	PRPS (UHF)	10GS/s	PRPS data + VAE	Outperform CNN/DBN
[106]	Elec. equip.	Detection	Images (cameras)	N/A	Images + capsule network	Detected accurately
[107]	GIS	Classification	Time-series data (UHF)	N/A	Images exported by CRO + preprocessing + 1D-CNN	88.9%
[108]	Elec. equip.	Classification	Time-series data (HFCT/SCS/TEV)	100 MS/s	Images exported by CRO + grayscale conversion + CNN	99.6/84.5/92%
[109]	Elec. equip.	Classification	PRPS	N/A	PRPS RBG image + data augmentation (superimposed with different noises) + CNN	Best model: 97.58%
[110]	GIS	Classification	PRPS (UHF)	100 MS/s – 10GS/s	PRPS data + CNN	95.6% (experiment data)/86.7% (mixed data)
[111]	GIS	Classification	Time-series data (UHF)	10 GS/s	Raw data + CNN (MobileNet)	96.5%
[112]	XLPE cable	Classification	Time-series data (acoustic)	50 kS/s	MFCC + CNN	96.3%
[113]	Switchgear	Detection	Time-series data (TEV)	N/A	Spectrogram + CNN	95.73%
[114]	Elec. equip.	Classification	Time-series data (HFCT)	24 kS/s	Bispectrum + CNN (ResNet-34)	80.83% (corona), 82.87% (PD), 80% (minor PD)
[115]	Elec. equip.	Classification	Time-series data (HFCT)	24 kS/s	Complex bispectrum + CNN	88.33% (corona), 96.67% (PD), 74.17% (minor PD)
[116]	GIS	Classification	Time-series data (UHF)	N/A	STFT + CNN	100%
[117]	Switchgear	Classification	Time-series data (TEV)	100 MS/s	Raw data + preprocessing + CNN	97.37-100%
[118]	GIS	Detection	Time-series data (pressure/temperature/infrared photoacoustic gas sensors)	N/A	Raw data + CNN	95.7%
[121]	Cable	Detection	Time-series data (HFCT)	500 MS/s	Raw data + CNN	97.38% (simulation) 93.23% (experiment)
[122]	Transformer	Classification	Time-series data (acoustic)	10 MS/s	Raw data + CNN	Above 88.8%
[123]	Transformer	Classification	Time-series data (ultrasonic)	1 MS/s	Vector of various features + CNN/DNN/RNN	93.9/91.2/89.4%
[124]	HV cable	Classification	Time-series data (HFCT)	100 MS/s	Handcrafted features + CNN	92.57%
[125]	Transformer	Classification	Time-series data (impedance)	100 MS/s	Raw data +LSTM	97.04%
[126]	GIS	Classification	PRPD (UHF)	N/A	Series of PRPDs + RNN (LSTM)	96.74%
[127]	GIS	Classification Localization	Time-series data (UHF)	10 GS/s	Multi-resolution spectrogram + CNN + LSTM	97.51% (both) 98.2% (recognition)
[129]	Transformer	Classification	PRPD (UHF)	10 GS/s	Series of PRPDs + CNN-LSTM	89-100%
[130]	OH lines	Detection	Time-series data (voltage sensor)	40 MS/s	STL + LSTM	79%
[131]	Electro. device	Classification	Time-series data	200 MS/s	Series of statistical features of one PD cycle (5ms) + LSTM	98.3%
[133]	Elec. equip.	Classification	PRPD (LDS'6 PD meter)	N/A	Raw PRPD variables + DBN	98-99.8%
[134]	Elec. equip.	Classification	PRPD (LDS'6 PD meter)	N/A	Vector of signal norms + DBN	84.3-93.5%
[135]	Transformer	Detection	DGA	N/A	Non-code ratios + DBN	96% for PD (PD dataset) 91.2 for PD (mixed dataset)
[136]	Transformer	Classification	Time-series data (HFCT)	20 MS/s	Raw data + Deep RF	96.53-99.07%
[138]	HV cable	Classification	Time-series data (impedance)	100 MS/s	Raw data + GAN + LR/SVM	Improves SVM by 0.52% LR by 1.72%
[139]	Elec. equip.	Classification	Time-series data (UHF)	5 GS/s	Raw data + GAN for data augmentation + ANN	Improves ANN by 1.34%
[141]	OH insulator	Classification	Time-series data (UHF)	N/A	STFT + GAN for data augmentation + CNN	Evaluation on real data gives correct results
[143]	DC cable	Classification	Time-series data (HFCT)	N/A	2D feature map + CNN (based on Alexnet)	91%

(1) ***Imbalanced dataset***: training ML classifiers with the imbalanced dataset can introduce bias into classes: they tend to focus on classifying classes with sufficient samples while ignore or misclassify the minority classes. Several researchers report this phenomenon, which can be found in [37, 107, 126, 130]. To overcome this issue, some traditional data augmentation methods are adopted, such as sliding window [126], ASMOTE [37], and noise superposition [109, 138]. However, these techniques show certain disadvantages. For example, ASMOTE is not very practical for high dimensional data, and can cause increases in the overlapping of classes. Furthermore, the real noise has different characteristics compared to some common noise for superposition, such as Gaussian noise. There is no sufficient direct evidence that whether the artificial PD samples superimposed with the noise still keep their key features or not. As a result, some advanced data augmentation methods using DL algorithms such as GAN and VAE are introduced, and demonstrate excellent capabilities in keeping the key features of the original PD samples [104, 138–141]. However, a comparative study between traditional and DL based data augmentation methods is yet to be conducted. Besides overall accuracy, other metrics such as *precision*, *recall*, *specificity*, and *F-score* should be used to provide more information for evaluating the reliability of the intelligent PD diagnostics under imbalanced conditions [96].

Another method to solve the imbalanced classification problem is the one-class classification [56, 57, 90]. However, there are still some challenges presented in the existing literatures. For example, under the scenario of learning with unlabeled data and without negative data, which is the case in the PD study in [57], OCSVM requires a more substantial amount of training data to set an accurate positive class boundary [144]. For example, one might need to consider all noise sources for the applied installation to get the complete noise distribution to avoid misclassification. Therefore, OCSVM based PD detection scheme using the method mentioned in [57] is more suitable for the applications where the environmental noise density has fewer variations. Furthermore, kernel-SVM can hardly deal with large datasets (i.e. a complete noise distribution). Therefore, one-class classification using modern DL techniques can be investigated as an alternative solution for conventional methods such as OCSVM. It is worthwhile to mention that one-classification is similar to binary classification, which is not quite suitable for PD recognition but PD detection. However, it shows the potential to identify unknown or multi-source PD [56, 90], which can be an effective pre-processing step for PD recognition.

(2) ***Small dataset***: many researchers investigate the relationship between diagnostics accuracy and the size of the training dataset [31, 100, 104, 110, 136]. The diagnostics accuracy reduces with fewer training samples for all types of ML methods investigated in the abovementioned referenced paper. It is challenging to train a DL based classifier from scratch with good generalization and satisfactory accuracy using insufficient training samples [67]. As expected, it is reported in [100, 104, 110] that conventional ML methods outperform DL methods with small datasets. The most common way to tackle this problem is designing the common features and selecting the most relevant features. However, this is a time-consuming process involving significant labor efforts. Recently, several advanced algorithms are introduced to assist the training process of ML models, especially for DL models. As mentioned above, DL based data augmentation techniques can generate synthetic samples to extend the dataset [104, 138–141].

Another possible way is applying transfer learning. The developments of transfer learning for fault diagnostics are motivated by the following two main facts: firstly, massive labeled fault data can be obtained under the laboratory environment, and few labeled fault data can be obtained from in-service equipment because they may not be allowed to operate under the fault conditions continuously; Secondly, variations exist between the source domain data (i.e. PD data collected under laboratory environment) used during the development and the target domain data (i.e. PD data collected from in-service equipment) encountered in the operation of the field. Transfer learning aims to leverage the knowledge of source domains to enhance the performance of the ML models on the target domain dataset with less required training samples. The procedures of transfer learning and general ML are illustrated and compared in Figure 10. Transfer learning has been showing promising results in other fields of fault diagnosis [145], while it is not well investigated for intelligent PD diagnostics.

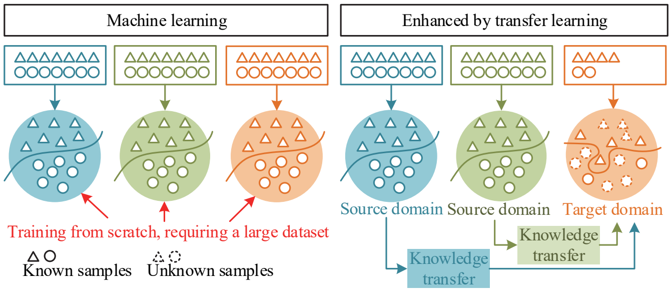


Figure 10. Machine learning vs. transfer learning.

(3) ***Inconsistency between training and testing dataset***: for most of the paper, the whole dataset is divided into training dataset and testing dataset with certain ratios (for most cases, the ratio is 80% to 20%). There is one important assumption: the testing dataset can represent the data encountered in the operation of the field. In other words, it assumes the physical properties and conditions of the applied installation remain the same. However, this is usually not the case, especially for practical applications as indicated in the previous paragraph. As reviewed above, performance degradation of ML methods is observed caused by different reasons, such as background environmental noise [57], changes in resolution (amplitude and phase) [34], different equipment [57], sensor locations [66, 122], operating temperature of the insulation [64, 66, 122], the presence of barrier [64, 66], and the dimension parameters of PD [101]. The conventional ML methods employed handcrafted features designed by experts, and there is no guarantee that the extracted features can fully represent the unique characteristics of PD under different conditions. As a result, DL is proposed to automate and optimize the feature learning, and end-to-end DL framework demonstrates better property of generalization compared to conventional ML methods when variations existed between training and testing dataset [101, 122]. Although better diagnosis performance can be obtained by DL methods, accuracy reduction of up to 23.3% can still be observed in some cases.

It is suggested the ML based classifiers should be trained with the PD signals collected from complex data sources covering as many conditions as possible [64]. In this case, the testing data is drawn from the same distribution with the training data. Due to the excellent learning capability on complex datasets, using DL methods can achieve better PD recognition accuracy with a complex dataset compared to conventional ML methods [54, 105, 110]. However, the performance of ML algorithms will degrade with a more comprehensive dataset. Moreover, collecting sufficient data from different operating conditions and different equipment is time-consuming and not realistic. A more feasible and cost-effective solution is to apply transfer learning to achieve better diagnosis accuracy with fewer data [145].

(4) ***Unlabeled dataset*****:** the majority of researchers are focusing on supervised learning, which indicates that the datasets need to be fully annotated. However, one of the main challenges for many industries is how to handle a large amount of unlabeled data (i.e. historical data) [103]. Unlabeled datasets are usually easier to obtain and require less labor to create. Unsupervised learning techniques, such as AE and DBN, provide feasible solutions for PD diagnostics using massive unlabeled data. Additionally, transfer learning also demonstrates promising results with massive unlabeled data in other fault diagnostics fields [146]. Therefore, the development of effective DL methods with unlabeled data is a valuable research direction in the future.

(5) ***Model complexity and real-time capability*****:** one of the main challenges to apply ML, especially for DL, to real industry environment is the real-time implementation of ML models. The majority of the researchers validate their ML based PD diagnostics algorithms using pre-recorded data in an off-line manner. Computational complexity and the accuracy of the ML models are dependent on the internal structure of the model, required sampling rate for data acquisition, size of the input sample, etc. Although DL methods demonstrate superior PD diagnostics accuracy over conventional methods (i.e. in [114, 115]), the expenses of these accuracy improvements are the exponentially increasing computational resources as well as energy consumption. In [104, 111], two light-scale CNN models are adopted to achieve reduced complexity and higher accuracy. However, there are fewer discussions on the selection of critical parameters, such as sampling frequency and data period. The effectiveness of those algorithms is subject to on-line real-time validation experiment. Furthermore, the choices and designs of the models should be contingent on the complexity of the problem and resource constraints of the specific application in order to achieve a reasonable balance. As a result, it is recommended to report essential information such as training/testing time and sensitivity to hyperparameters of the models, which can enable direct comparison across different models [147]. On the other hand, researches on efficient hardware implementation in low-cost devices such as FPGA [32] and application specific integrated circuit (ASIC) can also pave the way to realize the cost-effective solution based on DL methods for on-line intelligent PD diagnostics. For applications where the computation resource is the most important factor, conventional ML methods can be also considered.

(6) ***Interpretability*****:** although DL algorithms can achieve the best-in-class accuracy, the interpretability of DL is not well developed. The lack of core understanding restrains the development of DL at a fundamental level. The DL models are selected based on trails and errors instead of rigorous theories. Some researchers try to visualize the learnt CNN kernels in order to explain their physical meaning in terms of PD characteristics [104, 111, 116, 127]. Furthermore, the weights of the DL model are visualized to indicate the specific PD pattern that has been learnt by each neuron [97]. Even though those studies give some intuitions about the interpretability of DL algorithms, more thorough investigations should be carried out. Some possible methods for interpreting and understanding DL models are investigated in [148, 149] with solid theoretical analysis, which demonstrates the potential of facilitating the process of choosing the hyperparameters such as the filter size and number of layers to determine the optimal model structure for PD diagnostics appropriately.

As for designing and implementing DL algorithms, there are many frameworks and libraries available. Tensorflow is the most popular DL framework, which is based on a static computational graph. It also provides support for the ASIC customization, and recently Tensorflow-Lite is designed for DL mobile solutions. Additionally, the TensorBoard in Tensorflow provides effective visualization of parameters as well as data during the training progress. Keras is more user-friendly because it is a high-level application programming interface (API) on top of other popular lower level libraries such as Tensorflow. Keras is suitable for fast prototyping, while it is less configurable and less flexible. DL toolbox of Matlab also provides a good starting point for beginners with all commonly used DL algorithms. PyTorch, unlike other DL frameworks, operates with a dynamic graph, which allows users to update the computational graph during the runtime. It gives more flexibility during the development of DL algorithms compared to other DL frameworks. There is also an experimental release of PyTorch-Mobile for mobile solutions recently. Another popular DL framework for a mobile solution, Caffe2, is merging with PyTorch. Chainer also allows to dynamically update the computational graph and is faster than the majority of the frameworks based on Python. However, it has low popularity because of less documentation and community support. A summary of popular DL frameworks as well as libraries are listed in Table 3, which gives a preliminary guide for researchers who are interested in building and implementing PD diagnostics using DL algorithms. A more comprehensive survey among different DL frameworks can be found in [150]. Tensorflow and PyTorch are recommended.

6 CONCLUSIONS

This paper presents a systematic review of the state-of-the-art literatures published in the last five years using conventional ML and DL algorithms in the field of condition monitoring for PD diagnostics. Useful information, including application, objective, input signal, sensor type, sampling rate, key methodology, and accuracy, are summarized and compared to show how ML algorithms applied to different PD diagnostic applications. The advantages and disadvantages of different ML methods are also discussed in each sub-section. The following conclusions can be drawn:

Table 3. Summary of popular DL frameworks.

DL framework/ library	Core language	interface	GPU?	Mobile solution ?	Popularity
Tensorflow	C++, Python	Python, Java, Go, C++	Yes	No	Very high
Keras	Python	Python	Yes	No	High
Torch	C, Lua	C, C++, Lua, OpenCL	Yes	No	Low
PyTorch	Python, C	Python, ONNX	Yes	No	High
Caffe	C++	C++, Python, Matlab	Yes	No	High
Caffe2	C++	C++, Python, ONNX	Yes	Yes	Low
Theano	Python	Python	Yes	No	Low
DL toolbox	C, C++	Matlab	Yes	No	Very low
DL4j	Java	Java, Scala, Python	Yes	No	Medium
MXNet	C++	C++, Python, R, *et al*	Yes	No	Medium
Chainer	Python	Python	Yes	No	Very low
CNTK	C++	Python, C++, *et al*	Yes	Yes	Medium

(1) DL methods, which have attracted wide attentions recently, can achieve best-in-class accuracy and require fewer efforts for feature engineering in PD diagnostics. However, there is no solid mathematical foundation as compared to some conventional ML methods such as SVM. As black-box models, DL methods are also less interpretable compared to some conventional ML methods such as DT.

(2) Since the experimental conditions in different literatures are different, to facilitate further developments of robust ML algorithms and enable the direct comparison among different intelligent PD diagnostics, it calls for the establishment of a comprehensive and open-access dataset.

(3) The majority of existing literatures are focusing on ML based PD diagnostics under sinusoidal AC excitation, while there are limited studies that apply ML to PD diagnostics under DC and PWM excitation.

(4) Most studies using DL in this field are focusing on PD detection and PD classification, while other aspects such as PD localization are not well investigated as compared to conventional ML based methods.

(5) There are several barriers preventing ML, especially DL, from being applied for practical applications, including imbalanced dataset, small dataset, inconsistency between training and testing dataset, unlabeled dataset, model complexity and real-time capability, and interpretability. Those aspects are still not well explored and need further development. Potential solutions to the existing challenges have been suggested to facilitate the applications of intelligent PD diagnostic systems in scenarios of the real industry.

Therefore, there is a significant scope for improvement of the ML especially DL based PD diagnostics.

APPENDIX

Acronyms

ANFIS	Adaptive Neuro Fuzzy Inference System
ANN	Artificial Neural Network
API	Application Programming Interface
ASIC	Application Specific Integrated Circuit
ASMOTE	Adaptive Synthetic Minority Oversampling Technique
BN	Bayesian Network
BPNN	Backpropagation Neural Network
BSS	Blind Signal Separation
CNN	Convolutional Neural Network
CPLDs	Complex Programmable Logic Devices
DAE	Denoising Autoencoder
DBN	Deep Belief Network
DCGAN	Deep Convolutional Generative Adversarial Network
DGA	Dissolved Gas Analysis
DL	Deep Learning
DNN	Deep Neural Network
DST	Dempster-Shafer Theory
DT	Decision Tree
DWT	Discrete Wavelet Transform
EEMD	Ensemble Empirical Mode Decomposition
ELM	Extreme Learning Machine
EMI	Electromagnetic Interference
ENN	Ensemble Neural Network
EPR	Ethylene-Propylene Rubber
FCM	Fuzzy C-Means
FDAS	Fiber Distributed Acoustic Sensing
FDTD	Finite-Difference Time-Domain
FFT	Fast Fourier Transform
FIS	Fuzzy Inference System
FL	Fuzzy Logic
FPGAs	Field-Programmable Gate Arrays
GA	Genetic Algorithm
GAN	Generative Adversarial Network
GB	Gradient Boosting
GIL	Gas-Insulated Transmission Lines
GIS	Gas-Insulated Switchgear
GLCT	General Linear Chirplet Transform
GMM	Gaussian Mixture Model
GRNN	Generalized Regression Neural Network
HF	High Frequency
HFCT	High-Frequency Current Transformer
HMSVM	Hypersphere Multiclass Support Vector Machine
HOG	Histogram of Oriented Gradient
HV	High Voltage
HVDC	High-Voltage Direct Current
ICA	Independent Component Analysis
IF	Isolation Forest
IoT	Internet-of-Things
KL	Kullback-Leibler
kNN	k-Nearest Neighbors
KPLS	Kernel Partial Least Square
LBP	Local Binary Pattern
LDA	Linear Discriminant Analysis
LR	Logistic Regression
LSSVM	Least-Square Support Vector Machine
LSTM	Long Short-Term Memory
MaxP	Max Pooling
MDE	Multi-scale Dispersion Entropy
MFCC	Mel-Frequency Cepstrum Coefficients
ML	Machine Learning
MLP	Multilayer Perception

MLR	Multiple Linear Regression
mRMR	minimum-Redundancy and Maximum Relevance
OAA	One-Against-All
OAO	One-Against-One
OCSVM	One-Class Support Vector Machine
OH	Overhead
OS-ELM	Online-Sequential Extreme Learning Machine
PCA	Principal Component Analysis
PD	Partial Discharge
PNN	Probabilistic Neural Network
PRPD	Phase-Resolved Partial Discharge
PRPS	Phase-Resolved Pulse Sequence
PSD	Power Spectrum Density
PSO	Particle Swarm Optimization
RBF	Radial Basis Function
RBFN	Radial Basis Function Network
ReLU	Rectified Linear Unit
RF	Random Forest
R-F	Radio-Frequency
RBM	Restricted Boltzmann Machine
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
RS	Rough Set
RSS	Received Signal Strength
RSSI	Received Signal Strength Indicator
RVM	Relevance Vector Machine
SAE	Sparse Autoencoder
SCM	Subtractive Clustering Method
SCS	Surface Current Sensor
SDAE	Stacked Denoising Autoencoder
SNR	Signal-to-Noise Ratio
SRC	Sparse Representation Classifier
SSAE	Stacked Sparse Autoencoder
SSR	Signal Strength Ratio
STFT	Short-Time Fourier Transform
STL	Seasonal and Trend decomposition using Loess
SVD	Singular Value Decomposition
SVDD	Support Vector Data Description
SVM	Support Vector Machine
TEV	Transient Earth Voltage
TRPD	Time-Resolved Partial Discharge
UHF	Ultra-High Frequency
UV	Ultraviolet
VAE	Variational Autoencoder
VHF	Very-High Frequency
VMD	Variational Mode Decomposition
VPMCD	Variable Predictive Model-based Class Discrimination
WGAN	Wasserstein Generative Adversarial Network
WPD	Wavelet Packet Decomposition
XLPE	Cross-Linked Polyethylene

REFERENCES

[1] O. Kessler, "The importance of partial discharge testing: PD testing has proven to be a very reliable method for detecting defects in the insulation system of electrical equipment and for assessing the risk of failure," IEEE Power Energy Mag., vol. 18, no. 2, pp. 62–65, 2020.

[2] High-voltage test techniques – Partial discharge measurements, IEC 60270:2000, 2000.

[3] W. J. K. Raymond *et al,* "Partial discharge classifications: review of recent progress," Measurement, vol. 68, pp. 164–181, 2015.

[4] M. G. Danikas, N. Gao, and M. Aro, "Partial discharge recognition using neural networks: A review," Electr. Eng., vol. 85, no. 2, pp. 87–93, 2003.

[5] A. Mas'ud *et al,* "Artificial Neural Network Application for Partial Discharge Recognition: Survey and Future Directions," Energies, vol. 9, no. 8, p. 574, Jul. 2016.

[6] Barrios *et al,* "Partial discharge classification using deep learning methods—survey of recent progress," Energies, vol. 12, no. 13, p. 2485, Jun. 2019.

[7] B. Fruth and J. Fuhr, "Partial discharge pattern recognition. A tool for diagnosis and monitoring of ageing," CIGRE paper 15/33-12, 1990.

[8] C. Gao *et al,* "Partial discharge localization inside transformer windings via fiber-optic acoustic sensor array," IEEE Trans. Power Del., vol. 34, no. 4, pp. 1251–1260, 2019.

[9] H. Chai, B. T. Phung, and S. Mitchell, "Application of UHF sensors in power system equipment for partial discharge detection: A review," Sensors, vol. 19, no. 5, p.1029, Feb. 2019.

[10] H. Chai, S. Lu, B. T. Phung, and S. Mitchell, "Comparative Study of Partial Discharge Localization based on UHF Detection Methods," *Int. Conf. Exhib. Electr. Distrib. (CIRED),* 2019.

[11] C. Gao *et al,* "Localization of partial discharge in transformer oil using Fabry-Pérot optical fiber sensor array," IEEE Trans. Dielectr. Electr. Insul., vol. 25, no. 6, pp. 2279–2286, 2018.

[12] L. Cheng and T. Yu, "Dissolved Gas Analysis Principle-Based Intelligent Approaches to Fault Diagnosis and Decision Making for Large Oil-Immersed Power Transformers: A Survey," Energies, vol. 11, no. 4, p. 913, Apr. 2018.

[13] L. S. Lumba, U. Khayam, and R. Maulana, "Design of pattern recognition application of partial discharge signals using artificial neural networks," *Int. Conf. Electr. Eng. Informatics (ICEEI),* 2019, pp. 239–243.

[14] T. R. Sukma *et al,* "Classification of partial discharge sources using waveform parameters and phase-resolved partial discharge pattern as input for the artificial neural network," *Int. Conf. Cond. Monit. Diagnosis (CMD),* 2018.

[15] A. A. Soltani and A. El-Hag, "Denoising of radio frequency partial discharge signals using artificial neural network," Energies, vol. 12, no. 18, p. 3485, Sep. 2019.

[16] S. Kainaga, A. Pirker, and U. Schichler, "Identification of partial discharges at DC voltage using machine learning methods," *Int. Symp. High Volt. Eng. (ISH),* 2017.

[17] G. Luo, D. Zhang, K. J. Tseng, and J. He, "Impulsive noise reduction for transient earth voltage-based partial discharge using wavelet-entropy," IET Sci. Meas. Technol., vol. 10, no. 1, pp. 69–76, Jan. 2016.

[18] Y. Wang *et al*, "Multi-scale analysis and pattern recognition of ultrasonic signals of PD in a liquid/solid composite of an oil-filled terminal," Energies, vol. 13, no. 2, p. 366, Jan. 2020.

[19] Z. Li *et al,* "UHF partial discharge localization algorithm based on compressed sensing," IEEE Trans. Dielectr. Electr. Insul., vol. 25, no. 1, pp. 21–29, Feb. 2018.

[20] E. T. Iorkyase, C. Tachtatzis, R. C. Atkinson, and I. A. Glover, "Localisation of partial discharge sources using radio fingerprinting technique," *Loughbrgh. Antennas Propag. Conf. (LAPC),* 2015.

[21] A. A. Zahed *et al*, "Comparison of different fourth order Hilbert fractal antennas for partial discharge measurement," IEEE Trans. Dielectr. Electr. Insul., vol. 24, no. 1, pp. 175–182, Feb. 2017.

[22] S. Polisetty, A. El-Hag, and S. Jayram, "Classification of common discharges in outdoor insulation using acoustic signals and artificial neural network," High Volt., vol. 4, no. 4, pp. 333–338, Dec. 2019.

[23] Y. Khan, "Partial discharge pattern analysis using PCA and back-propagation artificial neural network for the estimation of size and position of metallic particle adhering to spacer in GIS," Electr. Eng., vol. 98, no. 1, pp. 29–42, Mar. 2016.

[24] A. Dobrzycki, S. Mikulski, and W. Opydo, "Using ANN and SVM for the detection of acoustic emission signals accompanying epoxy resin electrical treeing," Appl. Sci., vol. 9, no. 8, p. 1523, Apr. 2019.

[25] S. Anjum, S. Jayaram, A. El-Hag, and A. N. Jahromi, "Detection and classification of defects in ceramic insulators using RF antenna," IEEE Trans. Dielectr. Electr. Insul., vol. 24, no. 1, pp. 183–190, Feb. 2017.

[26] M. H. Wang, S. Der Lu, and M. J. Hsieh, "Application of extension neural network algorithm and chaos synchronization detection method to

partial discharge diagnosis of power capacitor," Measurement, vol. 129, pp. 227–235, Dec. 2018.

[27] E. T. Iorkyase *et al,* "Low-complexity wireless sensor system for partial discharge localisation," IET Wirel. Sens. Syst., vol. 9, no. 3, pp. 158–165, 2019.

[28] H. Hou, G. Sheng, S. Li, and X. Jiang, "A novel algorithm for separating multiple PD sources in a substation based on spectrum reconstruction of UHF signals," IEEE Trans. Power Deliv., vol. 30, no. 2, pp. 809–817, Apr. 2015.

[29] N. Zhou *et al,* "Error correction method based on multiple neural networks for UHF partial discharge localization," IEEE Trans. Dielectr. Electr. Insul., vol. 24, no. 6, pp. 3730–3738, Dec. 2017.

[30] J. He *et al,* "Partial discharge pattern recognition algorithm based on sparse self - coding and extreme learning machine," *IEEE Conf. Energy Internet Energy Syst. Integr. (EI2),* 2018.

[31] Q. Q. Zhang, H. Song, and G. H. Sheng, "Online sequential extreme learning machine for partial discharge pattern recognition of transformer," *IEEE Power Eng. Soc. Transm. Distrib. Conf. (T&D),* 2018.

[32] C. Gianoglio *et al,* "Hardware friendly neural network for the PD classification," *Annu. Rep. Conf. Electr. Insul. Dielectr. Phenom. (CEIDP),* 2018, pp. 538–541.

[33] S. Lin, X. Liu, J. Fang, and Z. Xu, "Is extreme learning machine feasible? A theoretical assessment (Part II)," IEEE Trans. Neural Networks Learn. Syst., vol. 26, no. 1, pp. 21–34, Jan. 2015.

[34] A. A. Mas'ud, B. G. Stewart, and S. G. McMeekin, "An investigative study into the sensitivity of different partial discharge φ-q-n pattern resolution sizes on statistical neural network pattern classification," Measurement, vol. 92, pp. 497–507, Oct. 2016.

[35] A. Mas'ud, J. Ardila-Rey, R. Albarracín, and F. Muhammad-Sukki, "An ensemble-boosting algorithm for classifying partial discharge defects in electrical assets," Machines, vol. 5, no. 3, p. 18, Aug. 2017.

[36] L. Li, J. Tang, and Y. Liu, "Partial discharge recognition in gas insulated switchgear based on multi-information fusion," IEEE Trans. Dielectr. Electr. Insul., vol. 22, no. 2, pp. 1080–1087, Apr. 2015.

[37] V. Tra, B. P. Duong, and J. M. Kim, "Improving diagnostic performance of a power transformer using an adaptive over-sampling method for imbalanced data," IEEE Trans. Dielectr. Electr. Insul., vol. 26, no. 4, pp. 1325–1333, Aug. 2019.

[38] V. Kecman, *Support Vector Machines: Theory and Applications*. Springer Science & Business Media, 2005.

[39] H. M. M. G. T. Herath *et al,* "Comparison of supervised machine learning techniques for PD classification in generator insulation," *IEEE Int. Conf. Ind. Inf. Syst. (ICIIS),* 2017, pp. 1–6.

[40] K. Firuzi, M. Vakilian, B. T. Phung, and T. R. Blackburn, "Partial Discharges Pattern Recognition of Transformer Defect Model by LBP & HOG Features," IEEE Trans. Power Deliv., vol. 34, no. 2, pp. 542–550, Apr. 2019.

[41] K. Firuzi, M. Vakilian, B. T. Phung, and T. R. Blackburn, "A Hybrid transformer pd monitoring method using simultaneous iec60270 and rf data," IEEE Trans. Power Deliv., vol. 34, no. 4, pp. 1374–1382, Aug. 2019.

[42] X. Li, X. Wang, A. Yang, and M. Rong, "Partial Discharge Source Localization in GIS Based on Image Edge Detection and Support Vector Machine," IEEE Trans. Power Deliv., vol. 34, no. 4, pp. 1795–1802, Aug. 2019.

[43] M. Kunicki and D. Wotzka, "A Classification Method for Select Defects in Power Transformers Based on the Acoustic Signals," Sensors, vol. 19, no. 23, p. 5212, Nov. 2019.

[44] B. M. A. Desai, R. Sarathi, J. Xavier, and A. Senugupta, "Partial Discharge Source Classification using Time-Frequency Transformation," *IEEE Int. Conf. Ind. Inf. Syst. (ICIIS),* 2018, pp. 362–366.

[45] R. Yao *et al*, "A New Discharge Pattern for the Characterization and Identification of Insulation Defects in GIS," Energies, vol. 11, no. 4, p. 971, Apr. 2018.

[46] X. Wang *et al*, "UHF Signal Processing and Pattern Recognition of Partial Discharge in Gas-Insulated Switchgear Using Chromatic Methodology," Sensors, vol. 17, no. 12, p. 177, Jan. 2017.

[47] J. Jineeth, R. Mallepally, and T. K. Sindhu, "Classification of Partial Discharge Sources in XLPE Cables by Artificial Neural Networks and Support Vector Machine," *IEEE Electr. Insul. Conf. (EIC),* 2018, pp. 407–411.

[48] G. Robles, E. Parrado-Hernández, J. Ardila-Rey, and J. M. Martínez-Tarifa, "Multiple partial discharge source discrimination with multiclass support vector machines," Expert Syst. Appl., vol. 55, pp. 417–428, Aug. 2016.

[49] P. Xie, "Analysis of fault of insulation aging of oiled paper of a large-scale power transformer and the prediction of its service life," IEEJ Trans. Electr. Electron. Eng., vol. 14, no. 8, pp. 1139–1144, Aug. 2019.

[50] Y. Zang *et al,* "A Novel Partial Discharge Detection Method Based on the Photoelectric Fusion Pattern in GIL," Energies, vol. 12, no. 21, p. 4120, Oct. 2019.

[51] J. Tang *et al, "*Feature selection for partial discharge severity assessment in gas-insulated switchgear based on minimum redundancy and maximum relevance," Energies, vol. 10, no. 10, p. 1516, Oct. 2017.

[52] W. J. K. Raymond, H. A. Illias, and A. H. A. Bakar, "High noise tolerance feature extraction for partial discharge classification in XLPE cable joints," IEEE Trans. Dielectr. Electr. Insul., vol. 24, no. 1, pp. 66–74, Feb. 2017.

[53] N. Morette, T. Ditchi, and Y. Oussar, "Feature extraction and ageing state recognition using partial discharges in cables under HVDC," Electr. Power Syst. Res., vol. 178, Jan. 2020.

[54] I. Mitiche *et al*, "Classification of EMI discharge sources using time–frequency features and multi-class support vector machine," Electr. Power Syst. Res., vol. 163, pp. 261–269, Oct. 2018.

[55] E. T. Iorkyase *et al,* "Radio location of partial discharge sources: A support vector regression approach," IET Sci. Meas. Technol., vol. 12, no. 2, pp. 230–236, Mar. 2018.

[56] H. Janani, B. Kordi, and M. J. Jozani, "Classification of simultaneous multiple partial discharge sources based on probabilistic interpretation using a two-step logistic regression algorithm," IEEE Trans. Dielectr. Electr. Insul., vol. 24, no. 1, pp. 54–65, Feb. 2017.

[57] E. Parrado-Hernández, G. Robles, J. Ardila-Rey, and J. Martínez-Tarifa, "Robust condition assessment of electrical equipment with one class support vector machines based on the measurement of partial discharges," Energies, vol. 11, no. 3, p. 486, Feb. 2018.

[58] H. Janani and B. Kordi, "Towards automated statistical partial discharge source classification using pattern recognition techniques," High Volt., vol. 3, no. 3, pp. 162–169, Sep. 2018.

[59] H. Shang, K. Lo, and F. Li, "Partial Discharge Feature Extraction Based on Ensemble Empirical Mode Decomposition and Sample Entropy," Entropy, vol. 19, no. 9, p. 439, Aug. 2017.

[60] H. Shang, F. Li, and Y. Wu, "Partial Discharge Fault Diagnosis Based on Multi-Scale Dispersion Entropy and a Hypersphere Multiclass Support Vector Machine," Entropy, vol. 21, no. 1, p. 81, Jan. 2019.

[61] S. Zhang *et al*, "Improving recognition accuracy of partial discharge patterns by image-oriented feature extraction and selection technique," IEEE Trans. Dielectr. Electr. Insul., vol. 23, no. 2, pp. 1076–1087, Apr. 2016.

[62] Y. Duan *et al,* "PD pattern recognition of XLPE cable based on parameter optimal support vector machine Algorithm," *IEEE Conf. Ind. Electron. Appl. (ICIEA),* 2019, pp. 355–359.

[63] E. T. Iorkyase, C. Tachtatzis, I. A. Glover, and R. C. Atkinson, "RF-based location of partial discharge sources using received signal features," High Volt., vol. 4, no. 1, pp. 28–32, Mar. 2019.

[64] M. Harbaji, K. Shaban, and A. El-Hag, "Classification of common partial discharge types in oil-paper insulation system using acoustic signals," IEEE Trans. Dielectr. Electr. Insul., vol. 22, no. 3, pp. 1674–1683, Jun. 2015.

[65] R. Hussein, K. B. Shaban, and A. H. El-Hag, "Robust feature extraction and classification of acoustic partial discharge signals corrupted with noise," IEEE Trans. Instrum. Meas., vol. 66, no. 3, pp. 405–413, Mar. 2017.

[66] W. L. Woon, A. El-Hag, and M. Harbaji, "Machine learning techniques for robust classification of partial discharges in oil-paper insulation systems," IET Sci. Meas. Technol., vol. 10, no. 3, pp. 221–227, May 2016.

[67] I. Goodfellow, B. Yoshua, and C. Aaron, *Deep Learning*, MIT Press, 2016.

[68] J. Faiz and M. Soleimani, "Assessment of computational intelligence and conventional dissolved gas analysis methods for transformer fault diagnosis," IEEE Trans. Dielectr. Electr. Insul., vol. 25, no. 5, pp. 1798–1806, Oct. 2018.

[69] L. A. Lumba, U. Khayam, and L. S. Lumba, "Application of fuzzy logic for partial discharge pattern recognition," *Int. Conf. Electr. Eng. Informatics (ICEEI),* 2019, pp. 210–215.

[70] A. A. Mas'ud, J. A. Ardila-Rey, R. Albarracín, F. Muhammad-Sukki, and N. A. Bani, "Comparison of the performance of artificial neural

networks and fuzzy logic for recognizing different partial discharge sources," Energies, vol. 10, no. 7, p. 1060, Jul. 2017.

[71] F. Zeng, Y. Dong, and J. Tang, "Feature extraction and severity assessment of partial discharge under protrusion defect based on fuzzy comprehensive evaluation," IET Gener. Transm. Distrib., vol. 9, no. 16, pp. 2493–2500, Dec. 2015.

[72] T. Kari *et al,* "An integrated method of ANFIS and Dempster-Shafer theory for fault diagnosis of power transformer," IEEE Trans. Dielectr. Electr. Insul., vol. 25, no. 1, pp. 360–371, Feb. 2018.

[73] J. Wang *et al,* "Evaluation on partial discharge intensity of electrical equipment based on improved ANFIS and ultraviolet pulse detection technology," IEEE Access, vol. 7, pp. 126561–126570, 2019.

[74] W. J. K. Raymond, H. A. Illias, and A. H. A. Bakar, "Classification of partial discharge measured under different levels of noise contamination," PLoS One, vol. 12, no. 1, p. e0170111, Jan. 2017.

[75] K. X. Lai, B. T. Phung, and T. R. Blackburn, "Application of data mining on partial discharge part I: Predictive modelling classification," IEEE Trans. Dielectr. Electr. Insul., vol. 17, no. 3, pp. 846–854, Jun. 2010.

[76] S. Wang, C. Ping, and G. Xue, "Transformer partial discharge pattern recognition based on random forest," J. Phys. Conf. Ser., vol. 1176, no. 6, p. 062025, Mar. 2019.

[77] A. A. Soltani and S. M. Shahrtash, "A pattern recognition based method for optimum decomposition level determination in wavelet transform for noise reduction of partial discharge signals," IET Sci. Meas. Technol., Sep. 2019.

[78] I. H. Kartojo, Y. B. Wang, G. J. Zhang, and Suwarno, "Partial discharge defect recognition in power transformer using random forest," *IEEE Int. Conf. Dielectr. Liq. (ICDL)*, 2019.

[79] E. T. Iorkyase *et al,* "Improving RF-based partial discharge localization via machine learning ensemble method," IEEE Trans. Power Deliv., vol. 34, no. 4, pp. 1478–1489, Aug. 2019.

[80] W. Si *et al*, "Defect pattern recognition based on partial discharge characteristics of oil-pressboard insulation for UHVDC converter transformer," Energies, vol. 11, no. 3, p. 592, Mar. 2018.

[81] S. Bag *et al,* "S-transform aided random forest based PD location detection employing signature of optical sensor," IEEE Trans. Power Del., vol. 34, no. 4, pp. 1261–1268, Aug. 2019.

[82] Y. Wang *et al,* "Separating multi-source partial discharge signals using linear prediction analysis and isolation forest algorithm," IEEE Trans. Instrum. Meas., pp. 1–1, Jul. 2019.

[83] X. Peng *et al*, "Random forest based optimal feature selection for partial discharge pattern recognition in HV cables," IEEE Trans. Power Deliv., vol. 34, no. 4, pp. 1715–1724, Aug. 2019.

[84] C. Gianoglio *et al,* "Tensor based algorithm for automatic partial discharges pattern classification," *Eur. Conf. Power Electron. Appl. (EPE),* 2019.

[85] A. M. Gonçalves Júnior, H. de Paula, and W. do Couto Boaventura, "Practical partial discharge pulse generation and location within transformer windings using regression models adjusted with simulated signals," Electr. Power Syst. Res., vol. 157, pp. 118–125, Apr. 2018.

[86] H. C. Yan, J. H. Zhou, and C. K. Pang, "Gaussian mixture model using semisupervised learning for probabilistic fault diagnosis under new data categories," IEEE Trans. Instrum. Meas., vol. 66, no. 4, pp. 723–733, Apr. 2017.

[87] C. Boya, M. Ruiz-Llata, J. Posada, and J. A. Garcia-Souto, "Identification of multiple partial discharge sources using acoustic emission technique and blind source separation," IEEE Trans. Dielectr. Electr. Insul., vol. 22, no. 3, pp. 1663–1673, Jun. 2015.

[88] C. Boya, G. Robles, E. Parrado-Hernández, and M. Ruiz-Llata, "Detection of partial discharge sources using UHF sensors and blind signal separation," Sensors, vol. 17, no. 11, p. 2625, Nov. 2017.

[89] M. Majidi, M. S. Fadali, M. Etezadi-Amoli, and M. Oskuoee, "Partial discharge pattern recognition via sparse representation and ANN," IEEE Trans. Dielectr. Electr. Insul., vol. 22, no. 2, pp. 1061–1070, Apr. 2015.

[90] J. Gao, Y. Zhu, and Y. Jia, "Pattern recognition of unknown partial discharge based on improved SVDD," IET Sci. Meas. Technol., vol. 12, no. 7, pp. 907–916, Oct. 2018.

[91] T. Le, D. Tran, W. Ma, and D. Sharma, "A unified model for support vector machine and support vector data description," *Int. Jt. Conf. Neural Networks (IJCNN),* 2012.

[92] J. I. Aizpurua *et al,* "Power transformer dissolved gas analysis through Bayesian networks and hypothesis testing," IEEE Trans. Dielectr. Electr. Insul., vol. 25, no. 2, pp. 494–506, Apr. 2018.

[93] Y. Jia and Y. Zhu, "Partial discharge pattern recognition using variable predictive model-based class discrimination with kernel partial least squares regression," IET Sci. Meas. Technol., vol. 12, no. 3, pp. 360–367, May 2018.

[94] X. Peng *et al,* "Rough set theory applied to pattern recognition of partial discharge in noise affected cable data," IEEE Trans. Dielectr. Electr. Insul., vol. 24, no. 1, pp. 147–156, Feb. 2017.

[95] M. X. Zhu *et al*, "Localization of multiple partial discharge sources in air-insulated substation using probability-based algorithm," IEEE Trans. Dielectr. Electr. Insul., vol. 24, no. 1, pp. 157–166, Feb. 2017.

[96] S. Misak *et al*, "A novel approach of partial discharges detection in a real environment," *Int. Conf. Environ. Electr. Eng. (EEEIC),* 2016.

[97] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, May 2015.

[98] V. M. Catterson and B. Sheng, "Deep neural networks for understanding and diagnosing partial discharge data," *IEEE Electr. Insul. Conf. (EIC),* 2015, pp. 218–221.

[99] V. Nair and G. E. Hinton, "Rectified linear units improve Restricted Boltzmann machines," *Int. Conf. Mach. Learn. (ICML),* 2010, pp. 807–814.

[100] J. Tang *et al*, "Assessment of PD severity in gas-insulated switchgear with an SSAE," IET Sci. Meas. Technol., vol. 11, no. 4, pp. 423–430, Jul. 2017.

[101] L. Duan *et al*, "Identification of partial discharge defects based on deep learning method," IEEE Trans. Power Del., vol. 34, no. 4, pp. 1557–1568, Aug. 2019.

[102] G. Wang *et al*, "Partial discharge pattern recognition of high voltage cables based on the stacked denoising autoencoder method," *Int. Conf. Power Syst. Technol. (POWERCON),* 2019, pp. 3778–3792.

[103] R. Zemouri *et al,* "Deep convolutional variational autoencoder as a 2D-visualization tool for partial discharge source classification in hydrogenerators," IEEE Access, vol. 8, pp. 5438–5454, 2020.

[104] Y. Wang *et al,* "Partial discharge pattern recognition of gas-insulated switchgear via a light-scale convolutional neural network," Energies, vol. 12, no. 24, p. 4674, Dec. 2019.

[105] J. Dai *et al,* "Partial discharge data matching method for GIS case-based reasoning," Energies, vol. 12, no. 19, p. 3677, Sep. 2019.

[106] Y. Li *et al*, "Image fusion of fault detection in power system based on deep learning," Cluster Comput., pp. 1–9, Jul. 2018.

[107] X. Wan *et al*, "Pattern recognition of partial discharge image based on one-dimensional convolutional neural network," *Int. Conf. Cond. Monit. Diagnosis (CMD),* 2018.

[108] N. Puspitasari *et al*, "Partial discharge waveform identification using image with convolutional neural network," *Int. Univ. Power Eng. Conf. (UPEC),* 2019, pp. 1–4.

[109] Y. Yang *et al,* "Research on partial discharge diagnosis based on data augmentation and convolutional neural," IOP Conf. Ser. Mater. Sci. Eng., vol. 677, no. 5, p. 052101, Dec. 2019.

[110] H. Song, J. Dai, G. Sheng, and X. Jiang, "GIS partial discharge pattern recognition via deep convolutional neural network under complex data source," IEEE Trans. Dielectr. Electr. Insul., vol. 25, no. 2, pp. 678–685, Apr. 2018.

[111] Y. Wang *et al*, "A MobileNets convolutional neural network for GIS partial discharge pattern recognition in the ubiquitous power internet of things context: optimization, comparison, and application," IEEE Access, vol. 7, pp. 150226–150236, Oct. 2019.

[112] Q. Che *et al,* "Partial discharge recognition based on optical fiber distributed acoustic sensing and a convolutional neural network," IEEE Access, vol. 7, pp. 101758–101764, Jul. 2019.

[113] Y. Lu, R. Wei, J. Chen, and J. Yuan, "Convolutional neural network based transient earth voltage detection," *Int. Symp. Parallel Distrib. Comput. (ISPDC),* 2017, pp. 386–389.

[114] I. Mitiche *et al,* "Deep residual neural network for EMI event classification using Bispectrum representations," *Eur. Signal Process. Conf. (EUSIPCO),* 2018, pp. 186–190.

[115] I. Mitiche *et al,* "Deep complex neural network learning for high-voltage insulation fault classification from complex bispectrum representation," *Eur. Signal Process. Conf. (EUSIPCO),* 2019.

[116] G. Li *et al*, "Partial discharge patterns recognition with deep Convolutional Neural Networks," *Int. Conf. Cond. Monit. Diagnosis (CMD),* 2016, pp. 324–327.

[117] K. Banno, Y. Nakamura, Y. Fujii, and T. Takano, "Partial discharge source classification for switchgears with transient earth voltage sensor using convolutional neural network," *Int. Conf. Cond. Monit. Diagnosis (CMD),* 2018.

[118] S. Wang, Y. Xia, C. Ping, and G. Xue, "Study on SF6 gas on-line monitoring method based on machine learning," *IEEE Inf. Technol. Mechatronics Eng. Conf. (ITOEC),* 2018, pp. 240–244.

[119] R. Liu *et al*, "Dislocated Time Series Convolutional Neural Architecture: An Intelligent Fault Diagnosis Approach for Electric Machine," IEEE Trans. Ind. Informatics, vol. 13, no. 3, pp. 1310–1320, 2017.

[120] S. Lu *et al*, "DA-DCGAN: An Effective Methodology for DC Series Arc Fault Diagnosis in Photovoltaic Systems," IEEE Access, vol. 7, pp. 45831–45840, 2019.

[121] M. A. Khan, J. Choo, and Y. H. Kim, "End-to-end partial discharge detection in power cables via time-domain convolutional neural networks," J. Electr. Eng. Technol., vol. 14, no. 3, pp. 1299–1309, May 2019.

[122] W. L. Woon, Z. Aung, and A. El-Hag, "Intelligent monitoring of transformer insulation using convolutional neural networks," *Int. Work. Data Anal. Renew. Energy Integr. (DARE),* 2018, pp. 127–136.

[123] Q. Zhang, J. Lin, H. Song, and G. Sheng, "Fault identification based on PD ultrasonic signal using RNN, DNN and CNN," *Int. Conf. Cond. Monit. Diagnosis (CMD),* 2018.

[124] X. Peng *et al,* "A Convolutional neural network-based deep learning methodology for recognition of partial discharge patterns from high-voltage cables," IEEE Trans. Power Del., vol. 34, no. 4, pp. 1460–1469, Aug. 2019.

[125] B. Adam and S. Tenbohlen, "Classification of multiple PD sources by signal features and LSTM networks," *IEEE Int. Conf. High Volt. Eng. Appl. (ICHVE),* 2019.

[126] M. T. Nguyen, V. H. Nguyen, S. J. Yun, and Y. H. Kim, "Recurrent neural network for partial discharge diagnosis in gas-insulated switchgear," Energies, vol. 11, no. 5, p. 1202, May 2018.

[127] G. Li *et al*, "Partial discharge recognition with a multi-resolution convolutional neural network," Sensors, vol. 18, no. 10, Oct. 2018.

[128] S. Lu, B. T. Phung, and D. Zhang, "A comprehensive review on DC arc faults and their diagnosis methods in photovoltaic systems," Renew. Sustain. Energy Rev., vol. 89, pp. 88–98, 2018.

[129] X. Zhou *et al,* "Research on transformer partial discharge UHF pattern recognition based on CNN-lSTM," Energies, vol. 13, no. 61, pp. 1–13, Dec. 2019.

[130] M. Dong and J. Sun, "Partial discharge detection on aerial covered conductors using time - series decomposition and long short - term memory network," ArXiv, pp. 1–21, Jul. 2019.

[131] E. Balouji, T. Hammarstrom, and T. McKelvey, "Partial discharge classification in power electronics applications using machine learning," *IEEE Glob. Conf. Signal Inf. Process. (GlobalSIP),* 2019, pp. 1–5.

[132] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," Neural Comput., vol. 14, no. 8, pp. 1771–1800, Aug. 2002.

[133] M. Karimi *et al*, "A novel application of deep belief networks in learning partial discharge patterns for classifying corona, surface and internal discharges," IEEE Trans. Ind. Electron., vol. 67, no. 4, pp. 3277–3287, Apr. 2020.

[134] M. Karimi, M. Majidi, M. Etezadi-Amoli, and M. Oskuoee, "Partial discharge classification using deep belief networks," *IEEE Power Eng. Soc. Transm. Distrib. Conf. (T&D),* 2018.

[135] J. Dai, H. Song, G. Sheng, and X. Jiang, "Dissolved gas analysis of insulating oil for power transformer fault diagnosis with deep belief network," IEEE Trans. Dielectr. Electr. Insul., vol. 24, no. 5, pp. 2828–2835, Oct. 2017.

[136] J. Guan, M. Guo, and S. Fang, "Partial discharge pattern recognition of transformer based on deep forest algorithm," J. Phys. Conf. Ser., vol. 1437, no. 1, p. 012083, Jan. 2020.

[137] I. J. Goodfellow *et al,* "Generative adversarial nets," *Conf. Neural Inf. Process. Syst. (NIPS),* 2014, pp. 2672–2680.

[138] W. Yijiang *et al*, "Partial discharge data augmentation of high voltage cables based on the variable noise superposition and generative adversarial network," *Int. Conf. Power Syst. Technol. (POWERCON),* 2018, pp. 3855–3859.

[139] X. Wang, H. Huang, Y. Hu, and Y. Yang, "Partial discharge pattern recognition with data augmentation based on generative adversarial networks," *Int. Conf. Cond. Monit. Diag. (CMD),* 2018.

[140] J. A. Ardila-Rey *et al,* "Artificial Generation of Partial Discharge Sources Through an Algorithm Based on Deep Convolutional Generative Adversarial Networks," IEEE Access, vol. 8, pp. 24561–24575, 2020.

[141] L. de Paula Santos Petri *et al*, "Partial discharge spectrogram data augmentation based on generative adversarial networks," *Int. Conf. Electr. Comput. Technol. Appl. (ICECTA),* 2019, pp. 1–5.

[142] I. Gulrajani *et al*, "Improved Training of Wasserstein GANs," *Conf. Neural Inf. Process. Syst. (NIPS),* 2017, pp. 5768–5778.

[143] Y. Zhu *et al,* "Partial discharge pattern recognition of DC XLPE cables based on convolutional neural network," *Int. Conf. Cond. Monit. Diagnosis (CMD),* 2018.

[144] S. S. Khan and M. G. Madden, "One-class classification: Taxonomy of study and review of techniques," Knowl. Eng. Rev., vol. 29, no. 3, pp. 345–374, 2014.

[145] H. Zheng *et al*, "Cross-Domain Fault Diagnosis Using Knowledge Transfer Strategy: A Review," IEEE Access, vol. 7, pp. 129260–129290, 2019.

[146] L. Guo *et al*, "Deep Convolutional Transfer Learning Network: A New Method for Intelligent Fault Diagnosis of Machines with Unlabeled Data," IEEE Trans. Ind. Electron., vol. 66, no. 9, pp. 7316–7325, 2019.

[147] E. Strubell, A. Ganesh, and A. McCallum, "Energy and Policy Considerations for Deep Learning in NLP," *Annu. Meet. Assoc. Comput. Linguist. (ACL),* 2019, pp. 3645–3650.

[148] G. Montavon, W. Samek, and K. R. Müller, "Methods for interpreting and understanding deep neural networks," Digit. Signal Process., vol. 73, pp. 1–15, Feb. 2018.

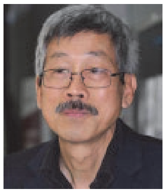
[149] Q. shi Zhang and S. chun Zhu, "Visual interpretability for deep learning: a survey," Front. Inf. Technol. Electron. Eng., vol. 19, no. 1, pp. 27–39, Jan. 2018.

[150] G. Nguyen *et al,* "Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey," Artif. Intell. Rev., vol. 52, no. 1, pp. 77–124, Jun. 2019.

Shibo Lu (S'17) received his Bachelor's degree in Electrical Engineering and its Automation from North China Electric Power University, China, and in Electrical Engineering from University of Wisconsin-Milwaukee, USA, in 2016. He is currently a Ph.D. research student in the School of Electrical Engineering and Telecommunications, the University of New South Wales, Sydney, Australia. His research interests are DC arc fault detection in photovoltaic systems, artificial intelligence applications in electrical engineering (e.g. fault diagnosis), and real-time condition monitoring.

Hua Chai (S'19) received Double bachelor's degrees in both Electrical Engineering and Project Management from North China Electric Power University, China, in 2014. He received an MPhil degree in Electrical Engineering from the University of New South Wales (UNSW), Australia, in 2019. He is currently a Ph.D. research student in the School of Electrical Engineering and Telecommunications, the University of New South Wales, Sydney, Australia. His research interests are partial discharge monitoring in high voltage equipment, renewable energy and microgrid modelling and power engineering education.

Animesh Sahoo (S'17) received the M.S. degree in Electrical Engineering from the Indian Institute of Technology, Madras, India, in 2016. He is currently working towards the Ph.D. degree with the School of Electrical Engineering and Telecommunications, University of New South Wales (UNSW), Sydney, Australia. Before joining UNSW, he was working as a Research Engineer for one year with the Electrical Machines and Drives Laboratory, National University of Singapore. His research interests include decentralized control and reliability aspects of grid-connected inverters, grid synchronization; fault ride-through operation. Additionally, he has worked on condition-based monitoring with predictive maintenance of high- and medium-voltage electrical power equipment.



B.T. Phung (SM'12) gained his Ph.D. in electrical engineering in 1998 and is currently an Associate Professor in the School of Electrical Engineering & Telecommunications at the University of New South Wales, Sydney, Australia. He has long involved in research/development on partial discharge measurement and practical on-line condition monitoring of high-voltage equipment. His research interests include dielectric and electrical insulation materials, high-voltage engineering, electromagnetic transients in power systems, power system equipment design and diagnostic testing.