IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 26 January 2024, accepted 15 February 2024, date of publication 22 February 2024, date of current version 29 February 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3368866*

TOPICAL REVIEW

# A Comprehensive Review of Signal Processing and Machine Learning Technologies for UHF PD Detection and Diagnosis (II): Pattern Recognition Approaches

**JIACHUAN LONG[1], (Member, IEEE), LIJUAN XIE[1], XIANPEI WANG[2], JUN ZHANG[3], (Senior Member, IEEE), BING LU[3], (Member, IEEE), CHUN WEI[1], DANGDANG DAI[4], GUOWEI ZHU[4], AND MENG TIAN[5], (Member, IEEE)**

[1]School of Electronics and Information Engineering, Wuhan Donghu University, Wuhan 430212, China
[2]Electronic Information School, Wuhan University, Wuhan 430072, China
[3]Metrology Department, China Electric Power Research Institute, Wuhan 430074, China
[4]State Grid Information & Communication Branch of Hubei Electric Power Company Ltd., Wuhan 430077, China
[5]School of Automation, Wuhan University of Technology, Wuhan 430070, China

Corresponding author: Lijuan Xie (xielijuan_lucky@163.com)

This work was supported in part by the Key Research and Development Program of Hubei Province under Grant 2020BAB109, in part by the Youth Foundation of Wuhan Donghu University under Grant 2023dhzk003, and in part by the China National University Student Innovation & Entrepreneurship Development Program under Grant 202311798002.

**ABSTRACT** Partial discharge (PD) pattern recognition approaches are designed to identify the types or severities of the insulation defects within the high voltage equipment, which is vital for evaluating potential harmfulness and making follow-up maintenance plan. In recent years, many advanced machine learning (ML) algorithms have been introduced to this field and achieved remarkable results. As the second one of the two-part papers, we aim to give a comprehensive review regarding the pattern recognition approaches for ultra-high frequency (UHF) PD data in this paper. These methods are grouped into three categories, which are the traditional ML-based PD type recognition, the deep learning-based (DL) PD type recognition, and PD severity assessment. Specifically, for the first topic, feature extraction methods, dimensionality reduction methods and classification methods are reviewed separately. For the second topic, many state-of-the-art DL methods are discussed, including the deep belief network (DBN), deep autoencoder network (DAN), convolutional neural network (CNN), recurrent neural network (RNN), generative adversarial network (GAN), graph convolutional network (GCN), deep ensemble learning (DEL), etc. For the third topic, the relevant algorithms are also divided into the conventional ML-based ones and the DL-based ones, which are studied in detail respectively. Finally, a brief discussion about the application effects of the above technologies is given, and some future directions are suggested. This paper covers almost every aspect of the PD pattern recognition and highlights the latest progress, which can provide valuable references for scholars in this field.

**INDEX TERMS** Partial discharge, ultra-high frequency, pattern recognition, machine learning, deep learning, type recognition, severity assessment.

## ABBREVIATIONS

| | |
|---|---|
| AutoML | Automated machine learning. |
| AC-BEGAN | Boundary equilibrium generative adversarial network with auxiliary classifier. |
| ADDA | Adversarial discriminative domain adaptation. |
| AOKA | Adaptive optimal kernel amplitude. |
| AE | Autoencoder. |
| AL | Adversarial learning. |
| ACGAN | Auxiliary classifier generative adversarial network. |

The associate editor coordinating the review of this manuscript and approving it for publication was Zahid Akhtar.

| | |
|---|---|
| ANN | Artificial neural network. |
| ANOVA | Analysis of variance. |
| ANFIS | Adaptive neuro-fuzzy inference system. |
| BIC | Bayesian information criterion. |
| BLIMF | Band-limited intrinsic mode functions. |
| BPNN | Back-propagation neural network. |
| BPA | Basic probability assignment. |
| BFO | Bacterial foraging optimization. |
| BSS | Blind source separation. |
| BERT | Bidirectional encoder representations from transformers. |
| CD | Contrastive divergence. |
| CE | Cumulative energy. |
| CM | Chromatic methodology. |
| CNN | Convolutional neural network. |
| CWT | Complex wavelet transform. |
| CAE | Convolutional autoencoder. |
| CVAE | Convolutional variational autoencoder. |
| CGAN | Conditional generative adversarial network. |
| CS | Compressed sensing. |
| DAE | Denoising autoencoder. |
| DANN | Domain adversarial neural network. |
| DCGAN | Deep convolutional generative adversarial network. |
| DL | Deep learning. |
| DEL | Deep ensemble learning. |
| DT-CWT | Dual-tree complex wavelet transform. |
| DAN | Deep autoencoder network. |
| DAG | Directed acyclic graph. |
| DBN | Deep belief network. |
| DCG | Directed cyclic graph. |
| DBSCAN | Density-based spatial clustering of applications with noise. |
| DS | Dempster-Shafer. |
| DenseNet | Densely connected convolutional network. |
| DSC | Depthwise separable convolution. |
| DNAS | Differentiable neural architecture search. |
| DWT | Discrete wavelet transform. |
| EL | Ensemble learning. |
| ERT | Extremely randomized tree. |
| ENN | Ensemble neural network. |
| EXNN | Extension neural network. |
| FBNN | Feed-backward neuron network. |
| FCM | Fuzzy c-means. |
| FC | Fully connected. |
| FD | Fractal dimension. |
| FFNN | Feed-forward neural network. |
| FrFT | Fractional Fourier transform. |
| Fuzzy-ART | Fuzzy adaptive resonance theory. |
| FLI | Fuzzy logic inference. |
| FML | Fuzzy maximum-likelihood. |
| FNN | Fuzzy neural network. |
| FDD | Fast developing discharge. |
| GAN | Generative adversarial network. |
| GBDT | Gradient boosting decision tree. |
| GCN | Graph convolutional network. |
| GIL | Gas insulated line. |
| GILBS | Gas insulated load break switch. |
| GIS | Gas insulated switchgear. |
| GSP | Graph signal processing. |
| GLCM | Gray-level co-occurrence matrix. |
| GLCMOP | Gray-level co-occurrence matrix of optimal parameters. |
| GS | Grid search. |
| Grad-CAM | Gradient class activation map. |
| GM | Generative modeling. |
| HHT | Hilbert-Huang transformation. |
| HVGSA | Horizontal visibility graph spectral analysis. |
| HOG | Histogram of oriented gradient. |
| HVDC | High voltage direct current. |
| HFCT | High-frequency current transformer. |
| HNN | Hopfield neural network. |
| ICA | Independent component analysis. |
| IMF | Intrinsic mode function. |
| ISSA | Improved sparrow search algorithm. |
| ISOMAP | Isometric feature mapping. |
| K-CV | K-fold cross validation. |
| KG | Knowledge graph. |
| KL | Kullback-Leibler. |
| KPCA | Kernel principal component analysis. |
| LBP | Local binary pattern. |
| LDA | Linear discriminant analysis. |
| LE | Laplacian eigenmap. |
| LLE | Locally linear embedding. |
| LSTM | Long short-term memory. |
| LCNN | Lightweight convolutional neural network. |
| LMD | Local mean decomposition. |
| LS-SVM | Least square support vector machine. |
| MAE | Mean absolute error. |
| MCNN | MobileNets convolutional neural network. |
| ML | Machine learning. |
| MMD | Maximum mean discrepancy. |
| MMG | Mathematical morphology gradient. |
| MRMR | Minimum redundancy maximum relevance. |
| MFDFA | Multi-fractal detrended fluctuation analysis. |
| NMF | Non-negative matrix factorization. |
| NSGA-II | Nondominated sorting genetic algorithm II. |
| OOB | Out-of-bag. |
| OCWE | Optimal complex wavelet energy. |
| PAM | Pulse activation map. |
| PD | Partial discharge. |
| PDIV | Partial discharge inception voltage. |
| PCA | Principal component analysis. |
| PNN | Probabilistic neural network. |
| PDE | Probability density estimation. |
| PCPRA | Polar coordinate phase resolved analysis. |
| PRPD | Phase resolved partial discharge. |
| PRPS | Phase resolved pulse sequence. |

| | |
|---|---|
| PRPSA | Phase resolved pulse sequence analysis. |
| PSO | Particle swarm optimization. |
| ResNet | Residual network. |
| RNN | Recurrent neural network. |
| RF | Random forest. |
| RFE | Recursive feature elimination. |
| RBM | Restricted Boltzmann machine. |
| RBFNN | Radial basis function neural network. |
| ROC | Receiver operating characteristic. |
| SAE | Sparse autoencoder. |
| SAGPool | Self-attention graph pooling. |
| SHAP | Shapley additive explanation. |
| SWT | Synchrosqueezed wavelet transform. |
| SSAE | Stacked sparse autoencoder. |
| SSC | Spatial separable convolution. |
| STFT | Short-time Fourier transform. |
| SVM | Support vector machine. |
| SLT | Statistical learning theory. |
| SLLE | Supervised locally linear embedding. |
| SOFM | Self-organizing feature mapping. |
| SURF | Speeded up robust features. |
| SVD | Singular value decomposition. |
| SI | Swarm intelligence. |
| STM | Statistical moment matching. |
| TL | Transfer learning. |
| TRPD | Time resolved partial discharge. |
| t-SNE | t-distributed stochastic neighbor embedding. |
| UHF | Ultra-high frequency. |
| VAE | Variational autoencoder. |
| VMD | Variational mode decomposition. |
| VSWR | Voltage standing wave ratio. |
| WGAN | Wasserstein generative adversarial network. |
| WGAN-GP | Wasserstein generative adversarial network with gradient penalty. |
| 1D-CNN | One-dimensional convolutional neural network. |
| 2D-PCA | Two-dimensional principal component analysis. |

## I. INTRODUCTION

It's well known that a correlation exists between the pattern of PD signals and the insulation defects. On one hand, for different types of insulation defects, the excited PD signals are inherently diverse due to different discharge mechanism. Thus, accurate recognition of the defect types is helpful to analyze the possible failure causes. On the other hand, with the gradual deterioration of insulation, the characteristics of PD signal also change. Therefore, severity assessment of the insulation defects based on PD signals is also necessary to evaluate the operating status and predict remaining life of power equipment [1], [2], [3]. Based on the above reasons, pattern recognition is essential for the PD fault diagnosis.

As one of the most popular methods in practice, the UHF PD fault diagnosis has attracted wide attention. Pattern recognition approaches based on UHF PD data have been broadly studied and remarkable progress has been achieved. In general, there are two research paradigms regarding this issue, which are the traditional ML-based methods and the DL-based methods. The former mainly adopts the route of "artificial feature extraction + shallow classifier", while the latter follows an end-to-end route of "automatic feature extraction + deep classifier". Fig. 1 shows the typical workflow of the above two paradigms. As can be seen, the major differences between them are reflected in two aspects: a) the DL methods can extract features from raw data without human intervention, while the traditional ML methods need a feature engineering process conducted by experts; b) the DL methods usually adopt deeper classifier network than that in the traditional ML methods [4]. In practice, the two roadmaps have their own advantages and disadvantages. When the data volume is sufficient, and the computing/storage resources are abundant, the DL methods can usually get higher diagnostic accuracy. Otherwise, traditional ML methods might be more appropriate, especially when resource consumption and result interpretability are the priorities. Hence, in this paper, we will make comprehensive reviews about the application of the two schemes in PD type recognition and severity assessment.

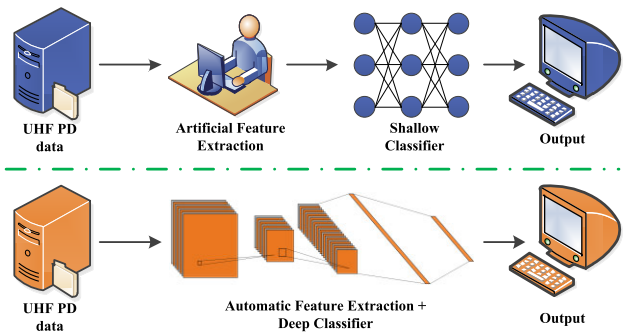



**FIGURE 1.** Typical workflow of the traditional ML-based methods and the DL-based methods.

Before our paper, some researchers have published review papers on PD pattern recognition. As early as in 2005, Sahoo et al. had summarized the application of pattern recognition technologies in PD diagnosis, and they discussed the feature extraction and traditional classification methods in detail [5]. Later, due to the rapid developments of ML algorithms, many relevant review papers had been published [6], [7], [8], [9]. However, these works only focus on the traditional ML methods. In recent years, the DL algorithms have also been introduced due to their automatic feature extraction abilities and excellent nonlinear fitting capabilities. Application of some primitive DL algorithms in PD diagnosis was firstly summarized in [10], such as DBN, RNN and CNN. Whereas, other novel DL methods like GCN or improved DL methods like TL were not mentioned in this paper. Most recently, Lu et al. carried out a survey on the ML methods in PD diagnosis, and both classical ML and modern DL approaches were included [11]. However,

only the classification methods were highlighted in this paper when discussing the classical ML approaches. In fact, well-designed features are often more conducive to accurate PD diagnosis than the classifiers. Thus, attention should also be paid to the feature engineering step. In addition, the severity assessment problem was not studied in depth in [11]. Last but not least, the ML field (especially the DL area) is experiencing unprecedented period of explosive development, with many new algorithms being proposed and applied in PD diagnosis every year. Therefore, it's essential to keep up with those cutting-edge developments.

In this paper, our motivation is to provide a review that covers almost every aspect of the PD pattern recognition, highlight some most powerful technologies, and supplement the latest algorithmic progress in this field. Keyword-based string including partial discharge and pattern recognition were mainly used to search the relevant papers in popular database such as the Google Scholar, IEEE, Science Direct, CNKI, etc. Considering that many pattern recognition methods applied for UHF PD data were also applicable for pulse current, ultrasonic and optical PD data, representative journal papers adopting the above kinds of PD data were also reviewed. In addition, only those papers discussing the PD type recognition and severity assessment topics were selected. All other PD-related papers regarding chemical methods, signal denoising, and source localization were excluded from this review paper.

The major contributions of this paper can be summarized as follows:

- We separately discuss the two major applications of pattern recognition methods in PD diagnosis, which is the type recognition and severity assessment. For each application, related ML-based and DL-based methods are described in detail.
- For each topic, the latest progress will be included. For example, the EL methods which show higher accuracy and better robustness than single ML method, will be highlighted. Moreover, some latest DL algorithms such as TL, GAN and GCN will also be emphasized.
- Before the conclusion, we summarize the application effects of existing methods, and then propose some potential directions for future improvement, including: a) PD mechanisms under new application scenarios; b) new-type UHF sensors; c) establishment of common database; d) pattern recognition methods under small /unbalanced samples or different working conditions; e) combined data-driven and knowledge-driven ML methods; f) multi-source fusion methods.

A diagrammatic view of the organization of this article is shown in Fig. 2. We begin by discussing the traditional ML-based and DL-based type recognition methods in Section II and Section III, respectively. In Section IV, various severity assessment methods are summarized. Afterwards, we make a discussion about the application effects of existing pattern recognition methods and prospect some future directions in Section V. Finally, this paper is ended with the conclusions in Section VI.

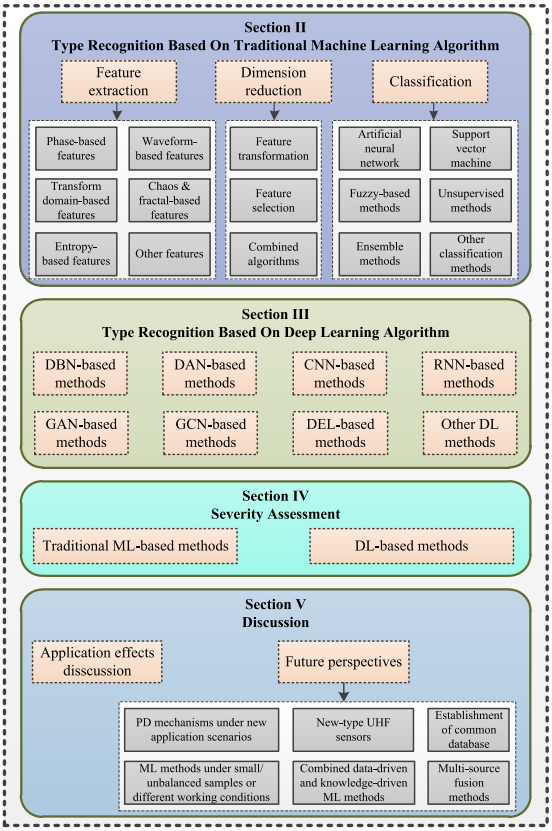



**FIGURE 2.** Diagrammatic view of the organization of this paper.

## II. TYPE RECOGNITION BASED ON TRADITIONAL MACHINE LEARNING ALGORITHM

In the early stage of intelligent PD diagnosis, traditional ML-based research paradigm was often adopted. That is, firstly, discriminative features were designed by the experts. Then, the features were selected or mapped into a lower transform space to reduce the data dimensionality. Finally, those well-chosen features were used as the input of shallow classifier, and output the diagnosis results. Following this processing flow, we will discuss the methods of feature extraction, dimensionality reduction, and classification successively in this section.

### A. FEATURE EXTRACTION METHODS

For traditional ML-based methods, extracting distinguishable features that can reflect different fault types is the key to the success of pattern recognition. In the field of PD recognition, the commonly extracted features can be divided into the following categories: a) the phase-based features; b) the waveform-based features; c) the transform domain-based features; d) the Chaos/fractal-based features; e) the entropy-based features; f) other features. In what follows, we will detail some representative literatures related to the above topics, as listed in Table 1.

**TABLE 1. Summary of the feature extraction methods for PD signals.**

| Category | Refer-ence | Sensor type | Extracted features | Reported accuracy |
|---|---|---|---|---|
| Phase-based features (statistic features) | [12] | IEC 60270 | Q-quantiles and other statistical operators of the PRPD patterns. | 99.1% |
| | [13] | IEC 60270 | Peak discharge, average discharge, discharge rate from the PRPD patterns. | 97.2% |
| | [14] | IEC 60270 | PCA features of six different PRPD patterns. | Not less than 60% under noisy conditions |
| | [15] | IEC 60270 | 15 statistical parameters of 2-D PRPD plots. | At least 65% |
| | [16] | IEC 60270 | Nine features in each degree of each cycle from the 3-D PRPS pattern. | Average of 92.05% |
| Phase-based features (image features) | [17] | UHF | HOG features of the 3-D PRPS graph. | 99.44% |
| | [18] | IEC 60270 | LBP and HOG features of the 2-D PRPD grayscale images. | 96.1% for LBP and 99.3% for HOG |
| | [19] | UHF | GLCMOP features of the 2-D PRPD grayscale image. | Average of 94.6% |
| | [20] | UHF | Speeded up robust features of PRPS grayscale images. | 89.71% |
| | [21] | IEC 60270 | TD-FW-2DPCA features of the 2-D PRPD gray images. | Up to 97.88% |
| Waveform-based features | [22] | IEC 60270 | Rise time, decaying time, and pulse width of the PD pulses. | Not found |
| | [23] | HFCT | Skewness and kurtosis features of the PD pulses. | At least 85.3% |
| | [24] | HFCT | 17 features of individual PD pulse, such as peak voltage, polarity, pulse width, etc. | Average of 93% |
| | [25] | IEC 60270 | PD pulse waveform features like rise time, fall time, slew rate, and pulse width. | Higher than 90% |
| | [26] | HFCT | Maximum amplitude, time deviation, rising and falling slope of the PD envelope. | Not found |
| | [27] | UHF | 16 features of the envelope of UHF PD waveform, such as pulse width, skewness, kurtosis, etc. | Up to 96% |
| | [28] | UHF | Width, area, and sharpness of the CE and MMG waveforms of PD signals. | Nearly 98% |
| | [29] | HFCT | Laguerre expansion coefficients of PD pulses. | Up to 96.9% |
| Transform domain-based features | [30] | HFCT | Mean, standard deviation, skewness, and kurtosis of the wavelet coefficients at various scales. | Average of 97.3% |
| | [31] | UHF | Skewness, kurtosis, and energy of the wavelet coefficients. | Up to 95.3% after denoising |
| | [32] | HFCT | Statistical moments of the distribution of the wavelet detail coefficients. | Not less than 80% |
| | [33] | UHF | Peak, peak-to-peak durations, sub-band scale energy/entropy, bandwidth of the CWT coefficients. | 94.9% for laboratory testing, 91.6% for field testing |
| | [34] | UHF | The OCWE features of the real and imaginary parts of | Above 92.5% |

**TABLE 1.** ***(Continued.)*** **Summary of the feature extraction methods for PD signals.**

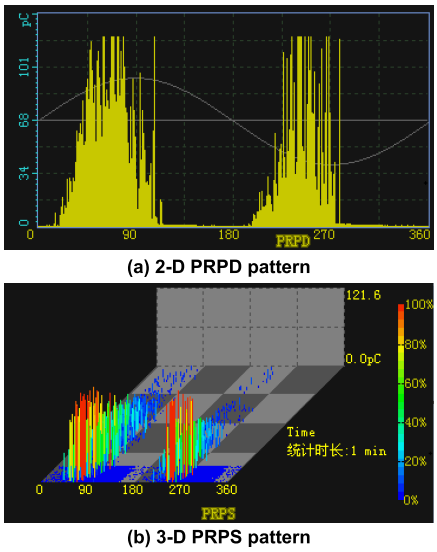
| | | | | |
|---|---|---|---|---|
| | | | UHF PD signals after decomposed by DTCWT. | |
| | [35] | UHF | Skewness and kurtosis of every detail coefficients obtained by DTCWT. | 98.35% for simulated defects, 92.1% for field tests |
| | [36] | HFCT | Texture features extracted from the GLCM calculated by CWT. | Average of 94.2% |
| | [37] | UHF | Six statistical features based on the S-transform modular matrix. | 97.33% without denoising |
| | [38] | UHF | 24-D feature vector from the time-frequency matrix of the UHF PD signal. | Up to 98.33% |
| | [39] | UHF | Geometric moments and Hu's invariant moments of the time-frequency image. | Up to 95% |
| | [40] | Acoustic | Spectrum values located in the low-frequency band (0-0.2 MHz). | At least 98.16% |
| | [41] | UHF | Extract the principal components of the AOKA matrices of PD signals as the features. | Highest accuracy of 94.33% |
| | [42] | UHF | Energy ratios obtained from two antennas optimized at different frequency bands. | Not found |
| | [43] | UHF | Nine features in frequency domain and four features in wavelet domain. | Highest accuracy of 96.8% |
| | [44] | UHF | Dominant wavelength, signal strength and saturation in chromatic space. | Average of 86.67% |
| | [45] | HFCT | Hyper statistical parameters and hyper edges of the Hypergraph of PD data. | Up to 98.8% |
| Chaos/fractal-based features | [46] | IEC 60270 | Average and standard deviation of the fractal dimension, and lacunarity. | Maximum of 95% |
| | [47] | UHF | Fractal dimension features of the time-frequency energy distribution map. | Up to 96.5% |
| | [48] | HFCT | Fractal dimension and lacunarity of the 3-D Hilbert energy spectrum of PD signals. | 78.3% in the worst case |
| | [49] | UHF | Fractal dimension and lacunarity of the 3-D $\beta-n-\eta$ spectrum of PD signals. | 77.5% in the worst case |
| | [50] | HFCT | The chaos eye coordinates after post-processing. | Average of 90% |
| | [51] | UHF | Largest Lyapunov exponent of each column of the 3-D PRPD pattern matrix. | Average of 89.2% |
| | [52] | UHF | Normalized and differenced version of the discharge amplitudes and time differences of PD pulses. | Not found |
| | [53] | HFCT | Lacunarity and fractal dimension of characteristic matrix based on the Chaos Synchronization technique. | 63.8% in the worst case |
| Entropy-based features | [54] | UHF | Multi-scale dispersion entropy of the effective IMFs obtained by VMD. | Almost 100% |
| | [55] | IEC 60270 | Multi-scale dispersion entropy of the BLIMFs by | Higher than 92% |

**TABLE 1.** *(Continued.)* **Summary of the feature extraction methods for PD signals.**

| | | | | |
|---|---|---|---|---|
| | | | synchrosqueezed windowed Fourier transform. | |
| | [56] | HFCT | Tsallis entropy of each subband coefficient of the PD contourlet transform image. | Up to 93.13% |
| | [57] | IEC 60270 | Distribution parameters of the sample entropy values of BLIMFs. | Not less than 99.17% |
| | [58] | UHF | Dominant CWT energy entropy and their scales. | Average of 89.81% |
| | [59] | UHF | Wavelet energy spectrum entropy and coefficient entropy of PD signal. | Up to 93.3% |
| | [60] | UHF | Multidimensional energy and multidimensional sample entropy parameters of PD signals. | Average of 96.4% |
| Other features | [61] | HFCT | Seven spectral graph features based on the HVGSA. | 98.85% ±0.74% |
| | [62] | HFCT | Four network features of the weighted adjacency matrices based on the complex network theory. | Up to 99.12% ±0.17% |
| | [63] | UHF+ Optical | Count the photon numbers in the three sub-bands in each integration window, then calculate their ratios. | Average of 87.9% |
| | [64] | Optical | Spectral distribution and intensity characteristics of the light emissions from PD sources. | Not found |
| | [65] | IEC 60270 | The codeword frequency matrix after the vector quantization of the PD apparent charge sequence at different test voltages. | Reached 91% |
| | [66] | UHF | Seven multifractal spectrum parameters based on the MFDFA technology. | Exceeded 90% |
| | [67] | IEC 60270 | Voltage change and phase angle change between three successive PD pulses. | Overall accuracy of 97.1% |
| | [68] | UHF | 13 TRPD features and 16 PRPD features. | Up to 97.25% |
| | [69] | HFCT | Features from the PRPD spectrum, time-domain waveforms, and frequency-domain waveforms. | Average of 89.375% |
| | [70, 71] | UHF | 15 features from PRPSA pattern+33 features from PRPD pattern+34 features from PCPRA pattern. | Reached 91% |
| | [72] | HFCT | 119 two-dimensional and 1082 three-dimensional features based on the 34 typical features of the raw data. | Exceeded 90% |
| | [73] | IEC 60270 | Original waveforms of PD pulses combined with phase information. | Highest accuracy of 92.05% |
| | [74] | IEC 60270 | The empirical spectrum distribution of sample covariance matrix S and the eigenvalue distribution of matrix product Z. | Up to 93.5% |
| | [75] | Acoustic | Singular features extracted by the Hankel SVD and hyper-features extracted from the hypergraph. | At least 90% for all testing cases |

### 1) PHASE-BASED FEATURES

In PD fault diagnosis, the phase-based spectra are probably the most popular signal patterns in practical applications. The basic principle of this kind of patterns is to visualize the relationships between the power frequency phase and other features of PD signal, then extract distinguishable parameters for different PD types.




(a) 2-D PRPD pattern

(b) 3-D PRPS pattern

**FIGURE 3.** Typical phase-based spectra of PD signals.

Two widely adopted phase-based patterns are the PRPD pattern and PRPS pattern, which are demonstrated in Fig. 3.

a) PRPD pattern ($\varphi - q - n$): obtained by counting the number $n$, amplitude $q$ and corresponding phase $\varphi$ of PDs within multiple power frequency cycles.

b) PRPS pattern ($\varphi - q - t$): obtained by counting the amplitude $q$ and corresponding phase $\varphi$ of PDs along the time-axis.

For either PRPD or PRPS pattern, there are two common feature extraction schemes. The first one is to calculate the statistical parameters directly from the PRPD or PRPS pattern, while the second one is to compute image-related features from PRPD or PRPS spectra. In next few paragraphs, we will discuss them separately.

PRPD or PRPS statistical features are those parameters that can represent the univariate or multivariate distribution of the elements from set $\{\varphi, q, n, t\}$. In general, there are two types of statistical operators for PRPD or PRPS pattern: a) the statistical moments describing the univariate distribution, including the mean $\mu$, standard deviation $\sigma$, skewness $S_k$, kurtosis $K_u$ and so on; b) the parameters representing the differences of PD distribution between positive and negative part of a power frequency cycle, such as phase asymmetry $\Phi$, discharge asymmetry $Q$, cross-correlation coefficient $cc$, modified cross-correlation coefficient $mcc$, etc. More detailed explanation of these parameters can be found in [5].

The above statistical values and other derived parameters have been widely used in the identification of PD types. Some representative literatures are listed in Table 1. For example, Janani et al. constructed a feature dataset from the bivariate distribution $H_n(\varphi, q)$ of PRPD pattern in [12]. Authors in [15] adopted the 2-D PRPD plot as the PD pattern, and 15 statistical parameters were selected as the inputs of classifier.

Besides, many scholars have verified that the amplitude histogram of PD fits well with the two-parameter Weibull distribution [76], [77]:

$$H(q) = 1 - \exp[-(q/\alpha)^{\beta}] \tag{1}$$

where $\alpha$ and $\beta$ are the scale parameter and waveform parameter of Weibull distribution respectively.

As we discussed before, another widely employed feature extraction scheme is the image-related features from PRPD or PRPS spectra. In this kind of scheme, the PRPD or PRPS spectra are treated as images, and some popular characteristic features in the computer vision field can then be extracted from those graphic PD patterns. In practice, the HOG features, LBP features, SURF, GLCM features have been employed in PD recognition. In [17], Song et al. firstly established a basic PD database of GIS based on the 3-D PRPS graphs, then the HOG feature vector of each sample was computed. Firuzi et al. extracted both the LBP features and HOG features from PRPD grayscale images, and they evaluated the effectiveness of them in PD recognition [18]. The results showed that compared with existing methods using statistical features, the LBP or HOG features extracted in that paper can increase the average accuracy rate by more than 10%. In [19], Sun et al. proposed an improved texture feature extraction method namely GLCMOP and applied it to the 2-D PRPD grayscale images. To reduce the influence of noise on the PRPS graphs, Li et al. introduced the SURF algorithm to extract the feature points and feature descriptors of the PRPS grayscale images [20].

Despite the advantages of phase-based statistical features mentioned above, the following limitations are still non-negligible: a) it requires a large amount of PD data to calculate the statistical operators; b) it is difficult to obtain the power-frequency synchronization signal on the high-voltage side at real substation; c) these features cannot be applied to the electrical equipment that the phase information is unavailable, such as the HVDC GIS or HVDC transformer, etc.

#### 2) WAVEFORM-BASED FEATURES

The TRPD is another widely adopted PD pattern, which use the waveform features to reveal the inherent attributes of the PD time-domain pulses. Common TRPD parameters include the pulse rise time, pulse fall time, pulse width, peak discharge magnitude, average discharge magnitude, etc.

Generally, there are two key advantages of the waveform-based features over the phased-based features. The first one is that the PD waveforms are directly related to the insulation defects, thus the waveform-based features have more explicit physical interpretability. The other one is that since there is no need for phase information, the waveform-based features can be applied to HVDC equipment. Due to the above benefits, the waveform features based on TRPD patterns are also extensively applied in practice.

In [22], the authors extracted the waveform characteristics such as rise time, decay time, pulse width, amplitude from PD pulses to distinguish AC and DC PDs. They found that the most obvious differences between AC and DC PDs were reflected in the amplitude. Su et al. illustrated that not only the waveform features, but also the statistical parameters (e.g. skewness and kurtosis) of the individual PD pulse can be employed to achieve PD classification [23]. Considering the practical UHF PD signal show characteristics of oscillation and fluctuation, the envelop-based TRPD features are also effective. Therefore, authors of [26] and [27] extracted this kind of features. In [28], the CE curve was chosen to characterize UHF PD signals, and features from both time-domain and frequency-domain were adopted. Different from all the above schemes, Janani et al. proposed a novel feature extraction method based on the orthogonal series expansion in [29]. By setting a proper scaling factor and order of Laguerre functions, the corresponding coefficients were selected as the fingerprints for PD identification.

#### 3) TRANSFORM DOMAIN-BASED FEATURES

Sometimes, the discrepancies between different types of PD signals in PRPD or TRPD patterns are not obvious. In this case, the original signal can be transformed to another space through linear or nonlinear transformation, and features in the transformed domain may be more distinguishable.

One of the most frequently used transforms for PD feature extraction is the wavelet transform. For example, Evagorou et al. proposed a representative feature extraction method by using the wavelet packet transform in [30]. The PD signals coupled by HFCT sensor were firstly decomposed into wavelet coefficients at various scales, then the first four cumulants (i.e. mean, standard deviation, skewness and kurtosis) of wavelet coefficients at each scale were formed as the fingerprints. Feature extraction methods reported in [31], [32], [33] basically follow a similar scheme as in [30]. Compared with traditional wavelet transform, the DT-CWT is less prone to frequency aliasing and information loss. In [34], Tian et al. introduced DT-CWT to decompose the UHF PD signals, and extracted the OCWE features from the real and imaginary parts of higher decomposed layers. Similarly, Han et al. employed the DT-CWT to obtain the detail coefficients at different scales of raw PD signals, and skewness and kurtosis values of detail coefficients were computed and formed as the feature vector [35].

Another important type of feature extraction methods is based on the time-frequency transformation, such as the CWT, STFT and S-transform, etc. In [36], the CWT was applied to the time-domain PD signals to construct GLCM. Then, 18 texture features were derived from the GLCM.

In [37], Long et al. proposed an improved S-transform to map the time-domain UHF PD signal into the time-frequency domain, and established a frequency-maximum value matrix accordingly. Based on this matrix, six statistical parameters were computed and composed to a feature vector. A similar work was reported in [38]. Besides, other feature extraction methods based on time-frequency transformation can also be found in [39], [40], [41], [42], and [43], interested readers can refer to the original articles for more details.

Apart from the above-mentioned methods, other promising transform domain-based feature extraction approaches should also be paid attention to. For instance, Wang et al. developed a novel feature extraction scheme for UHF PD signals based on the CM in [44]. First, the UHF PD signals were transformed into the frequency domain, and each frequency spectrum was divided into 10 sections. Then, three chromatic parameters namely hue, lightness, saturation were calculated. Testing results showed that the chromatic parameters can not only be used to classify PD sources, but also revealed the propagation properties of the UHF signals in GIS. Another interesting work was reported in [45], in which the authors employed the hypergraph to extract features from PD signals. Specifically, the topological and geometrical features were obtained from the hyperedges of a PD hypergraph, and the extracted features performed well in PD classification.

#### 4) CHAOS/FRACTAL-BASED FEATURES

Considering that UHF PD signals inherently exhibit non-stationary and non-linear characteristics, nonlinearity features such as chaos and fractal can be employed to characterize the complex dynamic behavior of the system.

Most objects in nature are very complex and irregular. When there is self-similarity existed between the part and the whole of an object, it can be regarded as fractal. The FD and lacunarity are two important indexes in fractal theory, which can well reflect the basic characteristics of a certain system. In [46], Basharan et al. extracted the average and standard deviation of FD and lacunarity from the 3-D PRPD plots, and high classification accuracy was achieved by using these fractal features. Authors in [47], [48], and [49] basically adopted similar scheme as in [46], the difference was that they used different kinds of PD spectrograms as the source images for fractal features extraction. To sum up, the fractal features can well reflect the characteristics of different PD spectrograms, thus can achieve satisfactory recognition results.

On the other hand, some scholars have pointed out that PD phenomenon is not a rigorous random process, but a chaotic process. That is, there is a certain regularity behind this randomness. Thus, the Chaos theory can be employed to extract the distinguishable features. For example, Wang et al. introduced the Chaos synchronization detection method to PD fault diagnosis in [50]. They derived the mathematical expressions of the chaotic dynamic function, and defined the so-called chaos eyes as the features for PD diagnosis. Zhang et al. developed another chaotic feature extraction method for $\phi - v - n$ 3-D PD spectrogram in [51]. They firstly selected appropriate parameters to reconstruct the phase space of each phase sequence of PD spectrogram, and the largest Lyapunov exponent of each phase sequence was calculated. As a result, a 36-D chaotic feature vector was constructed. In [53], the authors combined the advantages of Chaos synchronization and fractal theory, and experimental results showed that the combined scheme can effectively extract the key information about different PD defects.

#### 5) ENTROPY-BASED FEATURES

Entropy is an important statistical measure to quantify the complexity or uncertainty of time series, which is particularly useful in representing characteristics of the non-linear, non-stationary signals. For this reason, it has been widely adopted in the feature extraction of industrial signals [78]. By far, many different forms of entropies have been created, some of which are listed and demonstrated in Table 2.

Many entropy-based feature extraction methods adopting the following roadmap: a) transform or decompose the raw PD signal into a series of sequences; b) calculate the selected kind of entropies of all sequences and combine them as the feature vector. For example, authors of [56] firstly performed the contourlet transform on the $q - \Delta t - n$ PD images, and Tsallis entropy for all sub-band coefficients were computed and formed as the input of the classifier. In [57], Gao et al. proposed a novel feature extraction procedure based on the VMD method and sample entropy. First, the VMD algorithm was used to decompose the PD signal into several BLIMFs. Second, the sample entropy of each BLIMF was computed and combined to get the final feature vector.

Considering the single-scale-based entropies are usually sensitive to small disturbances and may cause information loss, the multi-scale entropies are becoming more favorable in recent years. The key of this kind of method is to add a coarse graining step to the original time series signals, as illustrated in Fig. 4.

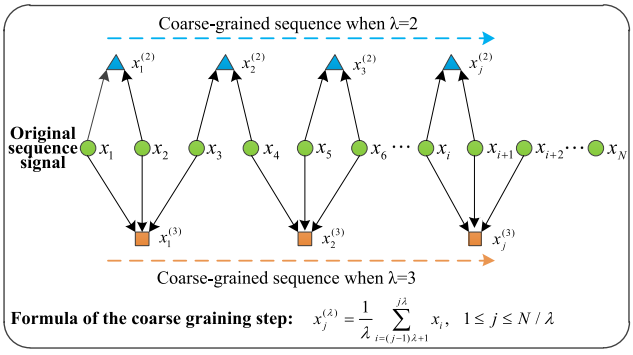



**FIGURE 4.** Illustration of the coarse graining step for time series signal.

In [54], Shang et al. calculated the multi-scale dispersion entropy of each BLIMF decomposed by VMD to form the

**TABLE 2. Explanations of some typical entropies.**

| Category | Calculation formulas/steps | Description |
|---|---|---|
| Shannon entropy | $H(X)=-\sum_{i=1}^{N} p(x_i)\log_2(p(x_i))$ | $\{x_1,x_2,...,x_N\}$ is the time series, $p(x_i)$ is the probability of $x_i$. |
| Rényi entropy | $H_\alpha(X)=\frac{1}{1-\alpha}\ln\sum_{i=1}^{N} p^\alpha(x_i)$, $\alpha\geq 0, \alpha\neq 1$ | $\alpha$ specifies the relative importance of small or large values of $p(x_i)$. |
| Approximate entropy | $ApEn=\phi^m(r)-\phi^{m+1}(r)$, $\phi^m(r)=\frac{1}{N-m+1}\times\sum_{i=1}^{N-m+1}\ln\left[\frac{num\{d[x_i,x_j]<r\}}{N-m+1}\right]$ | $\phi^m(r)$ is the mean value of pattern mean count, $m$ and $r$ is the embedding dimension and similarity coefficient respectively, $num\{d[x_i,x_j]<r\}$ means the count of the distance between $x_i$ and $x_i$ that lower than $r$. |
| Sample entropy | $SampEn=-\ln\left(\frac{B^{m+1}(r)}{B^m(r)}\right)$, $B^m(r)=\frac{1}{N-m+1}\times\sum_{i=1}^{N-m+1}\left[\frac{num\{d[x_i,x_j]<r\}}{N-m}\right]$ | $B^m(r)$ is the mean value of pattern mean count. Other parameters are the same as in approximate entropy. |
| Permutation entropy | **Step1**: obtain the reconstruction vector $\boldsymbol{X}_i=\{x_i,x_{i+\tau},...,x_{i+(m-1)\tau}\}$; **Step2**: obtain $\boldsymbol{S}$ by arranging $\boldsymbol{X}_i$ in ascending order, and denote $p_g$ as the probability of the $g$th permutation; **Step3**: $PE=\frac{-\sum_{g=1}^{k} p_g\ln(p_g)}{\ln(m!)}$ | $m$ and $\tau$ is embedding dimension and time delay, respectively. $\boldsymbol{S}$ is the permutated vector, which has $m!$ different types of permutations. |
| Dispersion entropy | **Step1**: obtain new time series $\boldsymbol{y}$: $y_i=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x_i} e^{\frac{-(t-\mu)^2}{2\sigma^2}}dt$, $i=1,...,N$ **Step2**: map $\boldsymbol{y}$ to $c$ classes: $z_i^c=round(c\cdot y_i+0.5)$ **Step3**: make time series $\boldsymbol{z}_j^{m,c}$: $\boldsymbol{z}_j^{m,c}=\{z_j^c,z_{j+d}^c,...,z_{j+(m-1)d}^c\}$ $j=1,2,....,N-(m-1)d$ **Step4**: map each $\boldsymbol{z}_j^{m,c}$ into a dispersion pattern $\pi_{v_0...v_{m-1}}$: $z_i^c=v_0,...,z_{i+(m-1)d}^c=v_{m-1}$ **Step5**: for each of $c^m$ possible dispersion patterns, obtain the relative frequency: $p(\pi_{v_0...v_{m-1}})=\frac{Number\{j\mid \boldsymbol{z}_j^{m,c}\ has\ type\ \pi_{v_0...v_{m-1}}\}}{N-(m-1)d}$ note that $j\leq N-(m-1)d$ **Step6**: calculate the dispersion entropy as follows: $DisEn(\boldsymbol{x},m,c,d)=-\sum_{\pi=1}^{c^m} p(\pi_{v_0...v_{m-1}})\ln[p(\pi_{v_0...v_{m-1}})]$ | $\sigma$ and $\mu$ are respectively the standard deviation and mean of the original time series $\{x_i\}$. $z_i^c$ is the $i$th member of $\boldsymbol{z}^c$. $m$ and $d$ are respectively embedding dimension and time delay. Finally, $p(\pi_{v_0...v_{m-1}})$ shows the number of dispersion patterns $\pi_{v_0...v_{m-1}}$ assigned to $\boldsymbol{z}_j^{m,c}$. |

feature vector for PD fault diagnosis. Testing results showed that the classification accuracy can almost reach 100%. Another similar scheme was proposed in [55], the difference is that the synchrosqueezed windowed Fourier transform was used to get the BLIMFs.

Apart from the above entropy-only feature extraction methods, integrated approaches that combining entropy with other features are also popular. For instance, Luo et al. developed a novel feature pair of energy entropies and their scales for UHF PD signals [58]. In [59], the wavelet energy spectrum entropy and coefficient entropy were combined. Authors of [60] established another joint feature region for UHF PD signals based on the multi-dimensional energy and sample entropy parameters. All these combined features can obtain satisfactory results.

#### 6) OTHER FEATURES

In addition to the feature types summarized above, scholars have also designed many new features for PD fault diagnosis, some of which are shown in Table 1.

For example, Roy et al. introduced the visibility graph to PD features design, enabling the analysis of non-stationary characteristics of PD signals in the graph domain [61], [62]. Testing results showed that superior performance can be achieved by using the GSP technology.

Ren et al. pointed out that the light intensity and integral spectrum of PD have close correlations with the PD activities. They developed a new feature extraction method utilizing multi-spectral information from the ultraviolet, visible and near-infrared regions [63], [64]. Experimental results verified the effectiveness of this optical PD diagnostic technique.

Other typical feature extraction approaches are also briefly described in Table 1, readers can refer to the corresponding papers for more details.

### B. DIMENSIONALITY REDUCTION METHODS

In practical ML applications, the high-dimensional data may not only contain redundant or irrelevant information, but also require much more cost for processing and storage. Therefore, performance of the ML methods may greatly deteriorate with the increase of the feature dimensionality, which is known as the ''curse of dimensionality'' [79].

A common way to alleviate this problem is to introduce the dimensionality reduction methods [80]. According to the working principles, the dimensionality reduction methods can be divided into two categories, which are the feature transformation algorithms and feature selection algorithms. The key of the first category is to map the original features into a low-dimensional subspace through linear or nonlinear transformation, in which the most important information of the raw data can be maintained. While the second category aims at choosing some of the best features from the original feature set based on specific indexes, thus an optimal feature subset can be formed. We summarize the taxonomy of the dimensionality reduction methods in Fig. 5. In addition,

typical application cases of these approaches in PD field are listed in Table 3.

**TABLE 3. Summary of the dimensionality reduction methods for PD signals.**

| Category | Reference | Dimensionality reduction algorithm |
|---|---|---|
| Feature transformation algorithms | [36] | PCA |
| | [54] | PCA |
| | [70] | PCA |
| | [47] | LDA |
| | [34] | Fisher LDA |
| | [35] | LLE |
| | [81] | SLLE |
| Feature selection algorithms | [27] | ERT |
| | [45] | RFE |
| | [72] | RF |
| | [61] | ANOVA |
| | [82] | RF + ANOVA |
| | [32] | Unsupervised feature selection based on the Dunn index |
| | [55] | MRMR |
| | [43] | MRMR |
| Combined algorithms | [83] | Linear transformation + Greedy search |
| | [84] | 2D-PCA + NSGA-II |

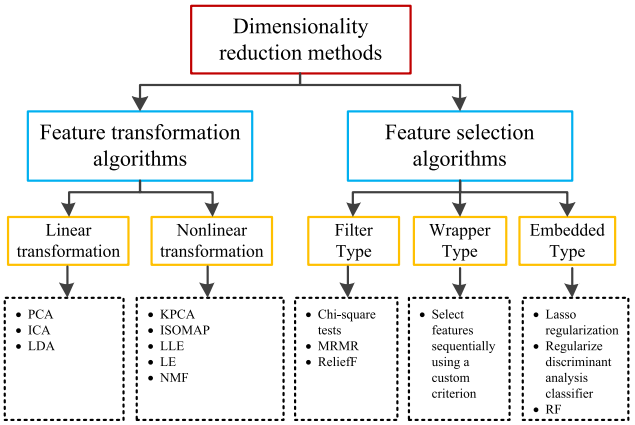



**FIGURE 5. Taxonomy of the dimensionality reduction methods.**

### 1) FEATURE TRANSFORMATION ALGORITHMS

According to whether the calculation involved is linear, the feature transformation methods can be divided into the linear ones and the nonlinear ones. Some representative algorithms are shown in Fig. 5, and their applications in dimensionality reduction of PD data will be detailed in this subsection.

Among all linear feature transformation algorithms, PCA is probably the most widely used one. The major principle of PCA is to map the original features into another feature space, in which the new features are the linear combination of the original ones. Through this transformation, most valuable information is retained in the first few principal components. In [36], Rostaminia et al. applied PCA on 18 texture features, and the first 6 principal components were used as the final features. In [54], PCA was adopted to extract representative components from the original multi-scale dispersion entropy features, resulting in a lower computational complexity. Yao et al. employed the PCA method to reduce the dimension of the combined PD feature set [70].

LDA is another popular linear transformation algorithm. The basic idea behind this method is to make the features of the same type as close as possible in the new subspace, while the features of different type are separated as far as possible. In [47], the authors used LDA to reduce the dimension of the extracted fractal dimensions features. Results showed that higher accuracy was achieved (improved by 9% comparing with no LDA method).

Nonlinear transformation methods are also desirable in feature dimensionality reduction for PD signals. One typical method of this kind is the LLE algorithm. It projects the original features into a low-dimensional space, and keeps the linear structure between the nearest neighbors unchanged. Detailed process of the LLE algorithm is illustrated in Fig. 6.

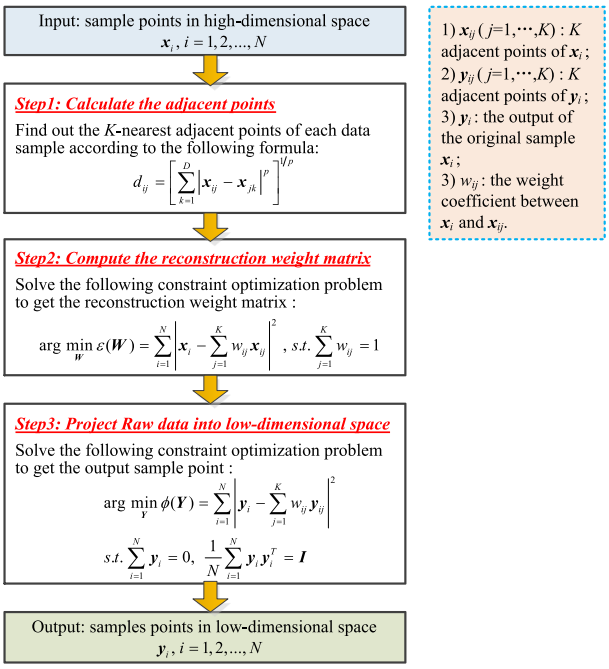



**FIGURE 6. Flowchart of the LLE algorithm.**

In [35], LLE algorithm was employed to reduce the feature parameters extracted from the UHF PD signals. According to the comparative results, the LLE algorithm showed higher accuracy and faster calculation speed over PCA. Considering that LLE is an unsupervised method, the class information of the training samples is ignored. To solve this problem, the SLLE method was developed by adding a distance correction term. Sun et al. applied SLLE to reduce feature dimension of the PD signals, and results demonstrated its superiority over the PCA and LLE methods [81].

### 2) FEATURE SELECTION ALGORITHMS

Feature selection methods reduce the dimension by selecting a small and more informative feature subset based on some criterions. According to the working mechanism, the feature

selection methods can be further categorized into three types: filter type, wrapper type, and embedded type. For the filter type algorithms, the feature selection process is regarded as a part of the data preprocessing step. However, the wrapper type and embedded type algorithms involve training process. Detailed description of these algorithms can be found in [85].

In [27], the ERT algorithm was applied to select the optimal feature subset. Results showed that ERT algorithm can greatly reduce the data volume. In [45], feature selection was implemented by two steps. First, the Helly property of Hypergraph was used to remove the redundant features from the raw PD data. Second, elimination of ill-structured and uninformative features was finished by the RFE algorithm. In two case studies, accuracy rate of 98.6% and 99.8% were respectively achieved by using only two and four features obtained by the proposed feature selection method.

The RF algorithm was also applied to the feature selection of PD signals [72]. The authors evaluated feature importance by examining the changes of classification error rate between the intact OOB samples and samples with particular feature permutation. Effectiveness of this approach was validated by the classification results using the obtained feature subset. In [61], the ANOVA test was employed to select the optimal HVG features of PD signals. Based on the ANOVA test, features showed high statistical significance were chosen as the finally inputs for classifier. Yao et al. proposed another feature selection scheme for UHF PD signals in [82], which combined the RF and ANOVA algorithms. In [32], the Dunn index was used as the quantifier of clustering quality, feature subset that provided the largest Dunn index was selected as the inputs for clustering algorithm. The results showed that this method can reduce the feature dimension to one while keeping the clustering accuracy above 80%.

In [55], a sequential feature selection algorithm namely MRMR was used to select a 18-D feature subset from the original 84-D feature vector for subsequent PD classification step. Similar feature selection scheme was also adopted in [43], and the authors demonstrated that using only the first feature indicated by the MRMR algorithm was sufficient for accurate recognition of UHF PD signals.

In essence, the feature selection issue can be considered as a multi-objective optimization problem regarding its two objectives: maximizing the accuracy and minimizing the number of selected features. Thus, powerful global searching techniques like the SI algorithms can be employed. Detailed introduction about the application of SI algorithms in feature selection can be found in [86], and we can expect that the SI-based feature selection will become a research hotspot in PD diagnosis field.

#### 3) COMBINED ALGORITHMS

To maximize the effectiveness of dimensionality reduction, some researchers combined the feature transformation algorithms and the feature selection algorithms. For example, Song et al. assembled the linear transformation method and the greedy search method for the feature reduction of HOG-based PD recognition [83]. Results showed that the overall recognition rate by using this method was 21.87% higher than that without it, and moreover, faster recognition speed was achieved.

Another combined feature dimension reduction method for PD signals was reported in [84]. At first, the PD gray images were transformed into vectors by using 2D-PCA algorithm. Then the NSGA-II algorithm was employed to select the most valuable features. The effectiveness of this method was validated by comprehensive case studies.

From above descriptions, we can observe that both feature transformation and feature selection are effective means for dimension reduction. However, in practice, users should note that the feature transformation algorithms may change the structure or attributes of the original PD dataset, resulting in poor physical interpretability and uncertain information loss. Thus, we should carefully choose the dimension reduction scheme according to the specific application requirements.

### C. CLASSIFICATION METHODS

After the well-designed features are obtained, the remaining work is to design an effective and reliable classifier. During the past decade, plenty of advanced classification algorithms have been developed to diagnose the PD faults, and it is almost impossible to cover all these studies in one single paper. In this subsection, only some latest and representative works will be discussed, as summarized in Table 4.

#### 1) ARTIFICIAL NEURAL NETWORK

ANN classifier is essentially a kind of bio-inspired nonlinear transformation technology that imitates the biological neural connections to fit the complex mapping relationship between feature vectors and classification categories. The basic unit of the ANN classifier is the artificial neuron model, which is demonstrated in Fig. 7. Despite its simplicity, combination of these units can form a complex and powerful network, which shows strong capability in non-linear fitting.

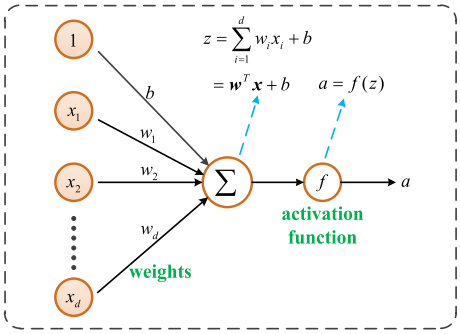



**FIGURE 7.** Graphic illustration of the artificial neuron model.

According to the topology structure, ANN can be divided into two categories: the FFNN and FBNN. For FFNN, each neuron only accepts the output of the previous layer and sends it to the next layer. For FBNN, however, the neurons can not

**TABLE 4.** Summary of the classification methods for PD signals.

| Category | Sensor type | Refer-ence | Major technology | Reported accuracy |
|---|---|---|---|---|
| Artificial neural network (ANN) | IEC 60270 | [87] | BPNN | No noise: 100%; 11dB: 83.33%; 7dB: 80%; 3dB: 72.22%. |
| | UHF | [31] | BPNN | Up to 95.3% after denoising |
| | Acoustic | [88] | BPNN | Recognition rate of 92.06% |
| | UHF | [68] | BPNN + DS fusion | Up to 97.25% |
| | HFCT | [50] | EXNN | Average of 90% |
| | HFCT | [30] | PNN | Average of 97.3% |
| | IEC 60270 | [89] | PNN | Average of 95.83% |
| | IEC 60270 | [90] | Fuzzy-ART neural network | 95.825% for known PD patterns, 94.2% for unknown PD patterns. |
| Support vector machine (SVM) | UHF | [27] | SVM | Up to 96% |
| | HFCT | [36] | SVM | Average of 94.2% |
| | UHF | [19] | SVM | Average of 94.6% |
| | UHF | [47] | SVM | Up to 96.5% |
| | UHF | [35] | SVM | 98.35% for simulated data, 92.1% for field data. |
| | IEC 60270 | [46] | SVM with user defined sigmoid kernel function | Maximum of 95% |
| | UHF | [39] | SVM optimized by cross-validation | Up to 95% |
| | UHF | [37] | SVM optimized by grid search | 97.33% without denoising |
| | UHF | [54] | SVM optimized by PSO | Almost 100% |
| | UHF | [20] | SVM optimized by BFO | Overall accuracy of 89.71% |
| | IEC 60270 | [91] | SVM optimized by ISSA | Overall accuracy of 96.25% |
| | Unknown | [92] | SVM optimized by FoSSA | Prediction accuracy of 97.5% |
| | Acoustic | [93] | SVM optimized by BAS and PSO | Overall accuracy of 93% |
| Fuzzy-based methods | IEC 60270 | [94] | FLI | Not found |
| | IEC 60270 | [95] | FLI | Almost 100% for the same size or type of PD sources |
| | HFCT | [23] | FCM | At least 85.3% |
| | HFCT | [96] | FCM | Not found |
| | IEC 60270 | [97] | Fuzzy decision tree | Average of 82.2% |
| | UHF | [41] | Fuzzy k-nearest neighbor classifier | Highest accuracy of 94.33% |
| | UHF | [28] | FML algorithm | Nearly 98% |
| Unsupervised methods | HFCT | [32] | K-means clustering | Not less than 80% |
| | IEC 60270 | [98] | X-means clustering | Average of 85% |
| | UHF + Optical | [63] | DBSCAN | Average of 87.9% |
| | IEC 60270 | [99] | Density-based clustering | Not found |
| | IEC | [100] | Stream clustering | Not found |
| | 60270 | | | |
| | HFCT | [53] | Extension theory-based clustering | 63.8% in the worst case |
| | UHF | [33] | SOFM | 94.9% for laboratory testing, 91.6% for field testing. |
| | HFCT | [101] | Automatic clustering + fuzzy logic | Not found |
| Ensemble methods | IEC 60270 + HFCT | [45] | RF | Up to 98.8% |
| | IEC 60270 | [73] | RF | Reaches 92.05% |
| | HFCT | [61] | RF | 98.85% ± 0.74% |
| | Acoustic | [40] | RF | At least 98.16% for actual PD signals |
| | IEC 60270 | [67] | AdaBoost | Overall accuracy of 97.1% |
| | HFCT | [102] | XGBoost | Overall accuracy of 96.93% |
| | IEC 60270 | [103] | ENN based on bagging | Average recognition rate more than 95% |
| | IEC 60270 | [15] | Bootstrap resampling + dynamically weighted aggregation | Reaches 90% |
| | UHF | [104] | Ensemble fuzzy-RBFNN | Reaches 99.37% ± 0.009% |
| | UHF | [71] | Ensemble method based on combined feature sets and adaptive boost technique | Reaches 91% |
| Other methods | UHF | [17] | Attribute selective naive Bayes classifier | 99.44% |
| | IEC 60270 | [65] | Fast matching algorithm | Reaches 91% |
| | HFCT | [69] | Fusion extreme learning machine | Average of 89.375% |
| | UHF | [49] | Extension recognition method | 77.5% in the worst case |
| | UHF | [59] | Hidden Markov model | Overall accuracy of 93.3% |
| | IEC 60270 | [105] | Variable predictive model and tanimoto similarity | 100% for known PD types, 93.33% for unknown types |
| | UHF | [43] | Six supervised ML algorithms | Average of 96.8% |

only accept the information from the previous layer, but also accept their own feedbacks. That is, neurons in FBNN have memory function. Taxonomy of ANN classifiers are shown in Fig. 8.

The ANN-based classifier has long been accepted as a reliable tool to distinguish different PD sources. For example, Illias et al. developed a FFNN classifier with two hidden layers in [87]. Authors of [31] and [88] adopted the similar BP-based ANN classifier to identify the different PD types in transformer and outdoor insulation, respectively. In [68], a multi-information fusion strategy for the recognition results of two BPNN sub-networks was proposed based on the DS evidence theory.

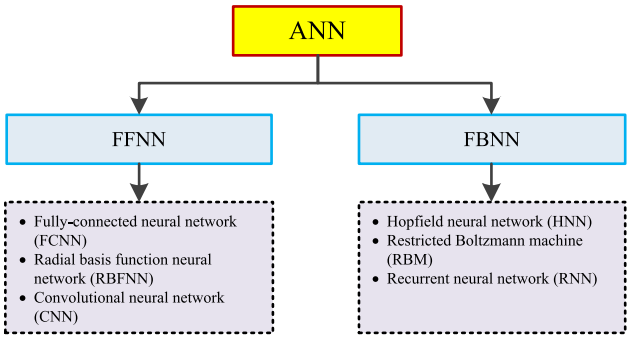



**FIGURE 8.** Taxonomy of ANN classifiers.

Another kind of ANN namely EXNN was also introduced to the PD type recognition [50]. Structure of the EXNN is illustrated in Fig. 9. As can be seen, it adopts double-layer structure and double-weight connections. By updating and adjusting these weight pairs, the network can be trained to identify different PD types. According to the results reported in [50], the EXNN can achieve accuracy rate as high as 90%.

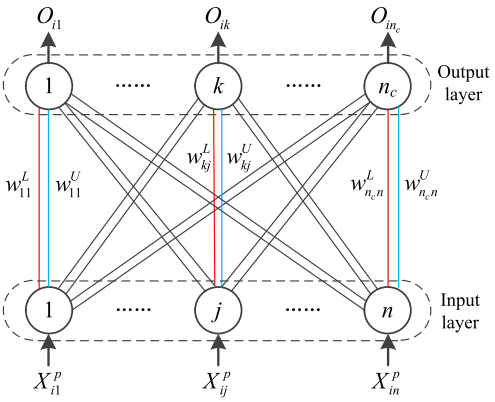



**FIGURE 9.** Structure of the extension neural network.

PNN is a popular FFNN which is developed based on the PDE technique and the minimum Bayesian risk criterion. Typical structure of PNN includes input layer, hidden layer, summation layer and output layer, as shown in Fig. 10.

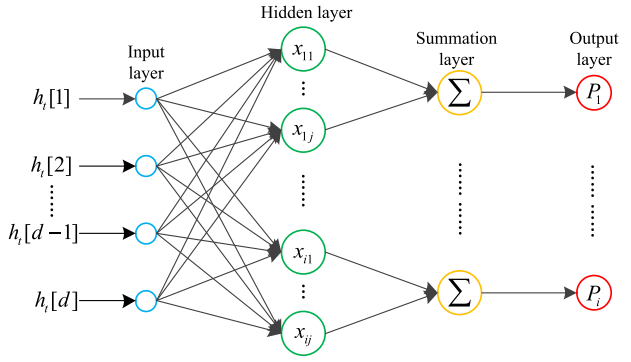



**FIGURE 10.** Structure of the probabilistic neural network.

Due to its simplicity and validity, PNN are broadly used in many industrial applications, including the PD classification field. In [30], the features extracted from wavelet coefficients were employed to train the PNN classifier, and an average recognition rate of 97.3% can be achieved. In [89], the PNN was adopted to classify the statistical characteristics of three typical PD defects in power transformer, and experimental results demonstrated the superiority of PNN over BPNN in both recognition accuracy and time consumption.

Recently, a novel ANN classifier based on the Fuzzy-ART was developed for PD recognition of power cables [90]. Combination of the fuzzy theory and ART system can well improve the learning ability of neural network. In addition, the Fuzzy-ART network shows good scalability, that is, the performance will be gradually improved with the continuous input of fault data. The above advantages make it very suitable for the online PD diagnosis application.

### 2) SUPPORT VECTOR MACHINE

SVM is a famous and effective ML algorithm proposed by Vapnik in the 1990s [106]. It is established based on the SLT, which aims to solving the pattern recognition problem under small samples. In SLT, error risk is redefined as the expected risk, rather than the commonly used empirical risk. The expected risk (denoted as $R(w)$) and empirical risk (denoted as $R_{emp}(w)$) satisfy the following relation with at least probability of $1-\eta$:

$$R(w) \le R_{emp}(w) + \sqrt{\left[h(\ln\frac{2n}{h}+1) - \ln\frac{\eta}{4}\right]\Big/ n} \quad (2)$$

where $h$ is the VC dimension of the model, $n$ is the number of training samples.

According to the SLT, the optimal hyperplane of a linear classification problem is defined as a hyperplane that can correctly separate the training samples, and meanwhile, the samples nearest to the hyperplane (so-called support vectors) has the largest distance from the hyperplane [106]. Purpose of the SVM algorithm is to find the optimal hyperplane, as illustrated in Fig. 11.

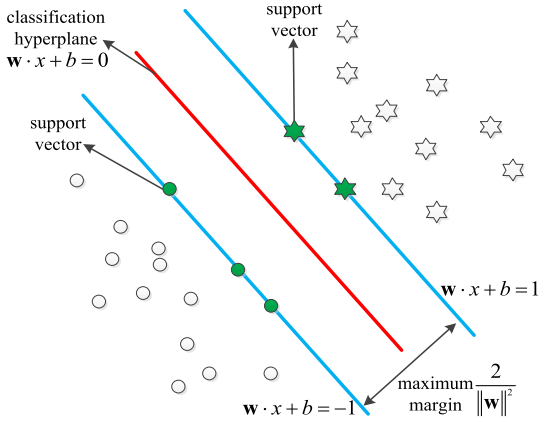



**FIGURE 11.** Illustration of SVM for a linear classification problem.

Given a training dataset $(x_i, y_i)_{i=1,\ldots,N}$, where $x_i$ and $y_i$ is the $i$th sample and corresponding class label respectively: $x_i \in R^d$, $y_i$ is set to 1 or $-1$. Then, the linear binary classification problem of SVM can be expressed as follows:

$$\min_{(\mathbf{w},b)} \frac{1}{2}\mathbf{w}^T\mathbf{w}, \quad s.t.\ y_i\left[(\mathbf{w}^T \cdot x_i) + b\right] - 1 \geq 0 \tag{3}$$

The above linear SVM can also be extended to nonlinear cases by introducing the kernel function, which can be formulated as follows:

$$\max_{\alpha} \quad Q_D(\alpha) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K(x_i, x_j)$$

$$s.t.\ C \geq \alpha_i \geq 0, \quad i = 1, \ldots, N, \quad \sum_{i=1}^{N}\alpha_i y_i = 0 \tag{4}$$

where $K(x_i, x_j)$ is the kernel function, $\alpha$ is a non-negative Lagrange multiplier, $Q_D(\alpha)$ is the Lagrange dual operator.

SVM has long been used for type recognition of PD faults. For example, authors of [19], [27], [35], [36], and [47] introduced SVM to classify PD data for different power equipment, and results of these studies are very competitive (all higher than 94%). However, the performance of SVM is highly depend on the setting of hyperparameters like the kernel function and the penalty coefficient. Therefore, optimized SVM algorithms were developed to improve the accuracy and reliability.

In [46], a user defined sigmoidal kernel function was designed to improve the flexibility of SVM algorithm, which is rewritten as:

$$g = \tanh(\gamma * U * V' + C') \tag{5}$$

where $U$ and $V$ are the axis of variables, $\gamma$ and $C$ are two coefficients. With this new kernel function, the improved SVM obtained higher recognition rate than that using other kernel functions such as polynomial and radial basis function.

Another commonly adopted optimization scheme is called the GS. A typical method of this kind is the K-CV. Li et al. developed an improved SVM algorithm to diagnose the UHF PD signals of GIS, in which the K-CV was employed to optimize the controlling parameter $\gamma$ and penalty parameter $C$ [39]. Similar GS-based optimization approach for SVM was also adopted in [37], results showed that the recognition rate reached up to 97.33% without denoising.

Unlike the GS-based optimization methods which require traversing the whole searching space, the heuristic algorithms attract more attention because they can converge to the optimal solution in less steps. The improved SVM algorithms developed in [20], [54], [91], [92], and [93] are all belong to this category. For example, in [20], Li et al. proposed an optimized SVM based on the BFO algorithm. The recognition error rate was used as the fitness function of BFO, and optimal parameters pair $(C, \gamma)$ of SVM were then obtained after a few iteration steps. In summary, these heuristic optimization methods seek to find out the most suitable hyperparameters of SVM in a highly automatic and efficient way. Interested readers can refer to [107] for a more comprehensive understanding of the application of heuristic optimization algorithms in SVM.

### 3) FUZZY-BASED METHODS

In essence, PD is a complex and stochastic phenomenon, and uncertainty often exists during the formation and propagation of PD signals, making absolutely accurate judgement of PD types very difficult. The fuzzy-based methods can well cope with this kind of problems. In fact, the fuzzy-based methods have been applied to PD type recognition for a long time. Here, we categorize them into three groups: a) FLI; b) fuzzy clustering; and c) FNN.

There are usually four processing steps in a FLI system, which are the fuzzification, fuzzy rules design, inference, and defuzzification. One major advantage of FLI is that the fuzzy rule sets are built based on the expert knowledge, thus they have clear physical meaning, making the results interpretable and easy to understand. In addition, the FLI system has high tolerances to incorrect data. Lumba et al. detailed a FLI system for PD type recognition in [94]. Specifically, feature vector $(n, \Delta n, q, \Delta q)$ from the PRPD pattern was used as the input, and a total of 81 fuzzy rules were designed. Finally, the judgment was finished by using the Mamdani method. In [95], the authors made a comparison of FLI and ANN in PD recognition, they found that both methods can identify the PD faults. However, the FLI classifier didn't require iterative training by large amount of data, which meant it may had better flexibility in practical application.

Fuzzy clustering is another important category of fuzzy-based method. Among many fuzzy clustering algorithms, the FCM might be the most widely used one. Major principle of FCM is to convert the clustering problem into a constrained nonlinear programming problem, whose objective function can be expressed by the following general formula:

$$\min J_m = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^m \left\|x_j - v_i\right\|^2 \tag{6}$$

where $x_j$ is the $j$th feature vector, $v_i$ is the $i$th clustering center, $u_{ij}$ is the membership function of $x_j$ belonging to $v_i$, $\|\cdot\|$ is the Euclidean distance operator. $m \in [1, \infty]$ is called the smoothing factor. The flowchart of FCM is given in Fig. 12. In [96], the FCM was employed to separate single PD signals from the overlapped time-frequency spectrum. Besides, Zhu et al. introduced another fuzzy clustering algorithm namely FML to classify UHF PD signals [28]. They demonstrated that FML can also successfully separate single PD signals from the mixed one.

FNN is also an important branch of fuzzy-based methods, which combines the characteristics of fuzzy system and neural network. In FNN, the input and output nodes are used to represent the input and output signals of the fuzzy system, and the hidden nodes are adopted to represent membership

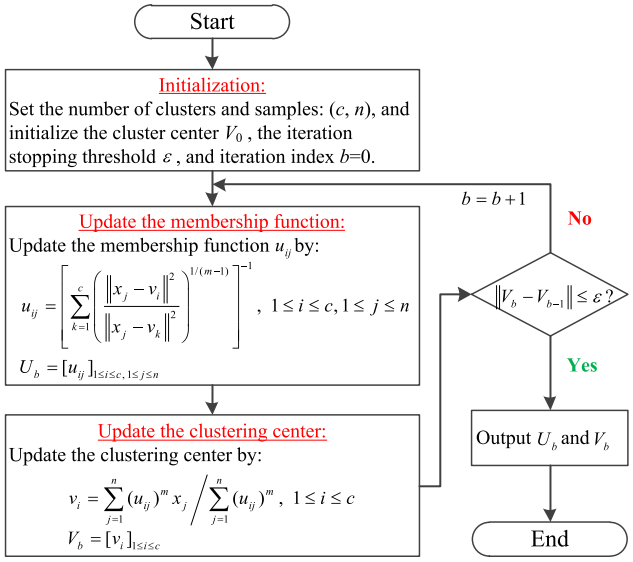



**FIGURE 12.** Flowchart of the FCM algorithm.

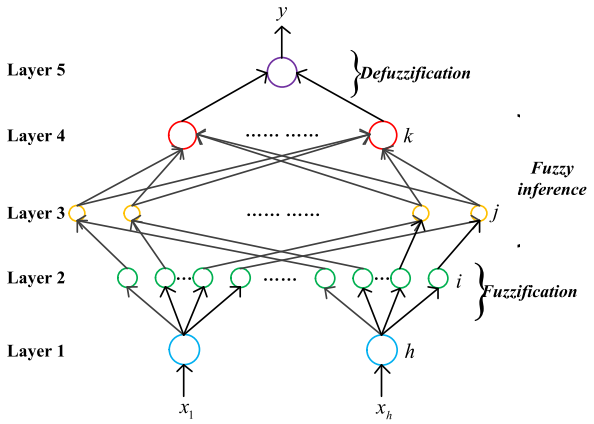



**FIGURE 13.** Typical structure of the FNN.

functions and fuzzy rules. Typical structure of FNN is shown in Fig. 13.

In PD recognition area, Abdel-Galil et al. proposed a novel neuro-fuzzy method for classification of different cavity sizes by combining the decision tree and FNN algorithms [97]. First, PD features were fed into the C4.5 algorithm to generate decision rules. Second, these rules were converted into a series of fuzzy IF-then rules. Finally, the parameters of membership functions were tuned by a back-propagation training algorithm. The authors verified that the proposed method can well identify the PD defects with different sizes. And more importantly, the whole process was interpretable.

#### 4) UNSUPERVISED METHODS

Up to now, most algorithms we have mentioned require both the data samples and their corresponding labels. In practice, however, it is very hard to obtain a well-labeled dataset. The unsupervised methods can learn valuable information such as features, types, or probability distribution directly from the raw data, without any manual guidance. Thus, it is now a very active research direction in the ML field.

The most popular unsupervised methods for PD diagnosis are clustering algorithms. For example, Morette et al. used the K-means clustering to distinguish noise from PD signals [32]. An improved version of K-means clustering namely X-means clustering algorithm was also employed to identify different PD types in power transformers [98]. The major principle of X-means clustering is to iteratively divide a data cluster into two, then calculates the corresponding BIC. Only when there is improvement in each loop, BIC value increases. Otherwise, the iteration steps should be stopped. Results demonstrated the superiority of X-means clustering over K-means clustering and FCM clustering in terms of efficiency and accuracy.

Another mainstream clustering strategy is the density-based clustering. In [63], Ren et al. introduced the DBSCAN to distinguish the corona discharge and creeping discharge. Testing results with unlabeled PD data showed that the accuracy can reach 89.3%. In [99], a novel density-based clustering algorithm was proposed by combining smoothed density clustering and density peak clustering methods. The presented algorithm performed well even when the clusters had very dissimilar densities and scatters.

To better cope with the online PD data (called stream data), two stream clustering algorithms namely Density Grids and DenStream were developed in [100]. There are two phases for stream clustering: the online phase and offline phase. At online phase, crucial information in stream data is maintained in terms of micro-clusters in a highly efficient way. Next, at offline phase, statistics of the micro-clusters are used to get the final clusters. The authors verified the effectiveness of the stream clustering method in processing the real-time PD data.

The SOFM is another kind of unsupervised method which is based on neural network. It usually contains two layers: the input layer and competition layer. The SOFM method learns to classify the input vectors according to the distribution as well as the spatial topological relation in the input space. In [33], the SOFM classifier was employed to identify the PD sources inside the electrical substation, an average accuracy of 94.9% for laboratory cases and 91.6% for field cases can be achieved.

#### 5) ENSEMBLE METHODS

EL is a very active and promising technique in ML field. The basic principle of EL is to combine several weak classifiers to form a strong classifier, which shows better performance than single classifier in terms of accuracy and generalization ability [108]. This working mechanism is just like people making decisions through voting. That is, different person thinks from different perspectives and levels, and a decision that integrates diverse information will be chosen as the final decision, which is often the best one in practice.

We give the diagram of EL method in Fig. 14. As can be seen, the following factors should be carefully considered to design an EL-based method: a) optimization of a single base classifier; b) diversity between different base classifiers; and c) the ensemble strategy.

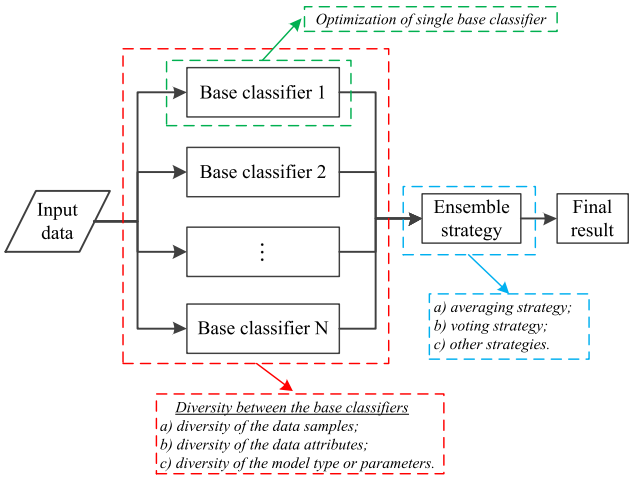



**FIGURE 14.** Block diagram of the EL method.

For the first factor, it is quite intuitive that one should choose the appropriate base classifier before any further steps. Some common metrics can help the designer to evaluate the performance of base classifiers, such as accuracy, error rate, precision, recall, ROC curve, etc.

The second essential factor is the diversity. Many scholars have proved that diversity between base classifiers is the key to promote generalization ability of EL method. As shown in Fig. 14, there are three ways to induce diversity. First, we can increase diversity from the perspective of data. Some dataset partition schemes including the hold-out method, the cross-validation method, and the bootstrap sampling method have been widely used in practice. Second, data attribute splitting is another strategy to introduce diversity. A famous case is the RF algorithm, which will be detailed in a later paragraph. The third diversity enhancement strategy is based on the varying of the types or parameters of the base models. For example, authors of [109] demonstrated that ensembles with different base models (called heterogeneous ensembles) showed better performance than with the same base models (called homogeneous ensembles).

The last significant factor is the ensemble strategy, which aims at combing the predictions of all base classifiers into a final output. The averaging rule and voting rule are two most widely adopted ensemble strategies. Let $g_m(\boldsymbol{x})$ be the output of the $m$th base classifier, where $m = 1, 2, \ldots, M$, then the averaging rule can be described as:

$$g(\boldsymbol{x}) = \frac{1}{M}\sum_{m=1}^{M} g_m(\boldsymbol{x}) \quad or \quad g(\boldsymbol{x}) = \sum_{m=1}^{M} a_m g_m(\boldsymbol{x}) \tag{7}$$

where $a_m$ is the weight of $g_m(\boldsymbol{x})$. Suppose $\bar{\mathcal{R}}(g)$ is the mean expected error of $M$ base classifiers, and $\mathcal{R}(g)$ is the expected error of the ensemble model $g(\boldsymbol{x})$, then one can easily prove that the following inequality holds:

$$\bar{\mathcal{R}}(g) \geq \mathcal{R}(g) \geq \frac{1}{M}\bar{\mathcal{R}}(g) \tag{8}$$

The voting rule essentially adopts the principle of ''minority obeys majority''. According to the specific mechanism, the voting rule can be further divided into the absolute majority voting, the relative majority voting, and the weighted voting. Their respective calculating process can be formulated as:

$$\underbrace{g(\boldsymbol{x}) = \begin{cases} y_i, & \sum_{m=1}^{M} g_m^i(\boldsymbol{x}) > \frac{1}{2}\sum_{m=1}^{M}\sum_{j=1}^{c} g_m^j(\boldsymbol{x}) \\ refuse\ to\ judge, & otherwise \end{cases}}_{\textbf{\textit{absolute majority voting}}} \tag{9}$$

$$\underbrace{g(\boldsymbol{x}) = y_i, \quad if \quad \sum_{m=1}^{M} g_m^i(\boldsymbol{x}) = \max_{j=1,2,\ldots,c} \sum_{m=1}^{M} g_m^j(\boldsymbol{x})}_{\textbf{\textit{relative majority voting}}} \tag{10}$$

$$\underbrace{g(\boldsymbol{x}) = y_i, \quad if \quad \sum_{m=1}^{M} a_m g_m^i(\boldsymbol{x}) = \max_{j=1,2,\ldots,c} \sum_{m=1}^{M} a_m g_m^j(\boldsymbol{x})}_{\textbf{\textit{weighted voting}}} \tag{11}$$

where $y_i (i = 1, 2, \ldots, c)$ is the class label, and $c$ is the number of classes. $g_m^i(\boldsymbol{x})$ denotes the output of the $m$th classifier regarding to $y_i$, and $a_m$ is the corresponding weight.

Based on the above-mentioned key factors, researchers have developed many kinds of EL methods, like the bagging-based methods, boosting-based methods, and the stacking-based methods. We demonstrate the diagrams of above three types of ensemble methods and list typical algorithms of each category in Fig. 15. Considering that our main purpose is to review their applications in PD recognition, the specific steps of each algorithm will not be detailed in this paper, interested readers can refer to the corresponding literatures.

Due to the excellent accuracy and generalization ability, the EL methods have been widely used in PD recognition. For example, Mas'ud et al. introduced the bagging technique (also called bootstrap aggregating) to identify the PD patterns of different fault geometries [15], [103]. They used six neural networks as the base classifiers and trained them by the data subsets obtained by bootstrap resampling. With dynamically weighted averaging strategy, results from all neural networks were combined to form the final decision. In testing cases, the proposed ENN produced an average recognition accuracy $\mu_E > 95\%$. Apart from the type recognition, the bagging methods have also been applied in other PD applications, like the source localization [110].

RF is another famous and powerful bagging-based method, which adopts the decision tree as the base classifier. Process of RF is illustrated in Fig. 16. Note that in RF algorithm, both sample perturbation and attribute perturbation are adopted, making it shows very competitive performance. Authors of [40], [45], [61], and [73] applied RF algorithm to identify PD sources for different high-voltage equipment, and one

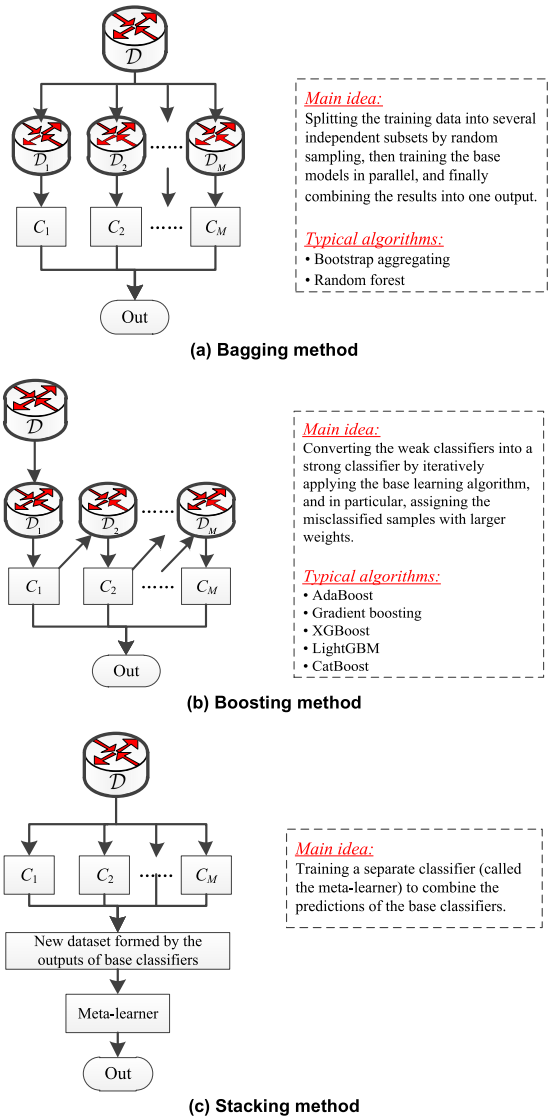



**FIGURE 15.** Diagrams of different EL methods.

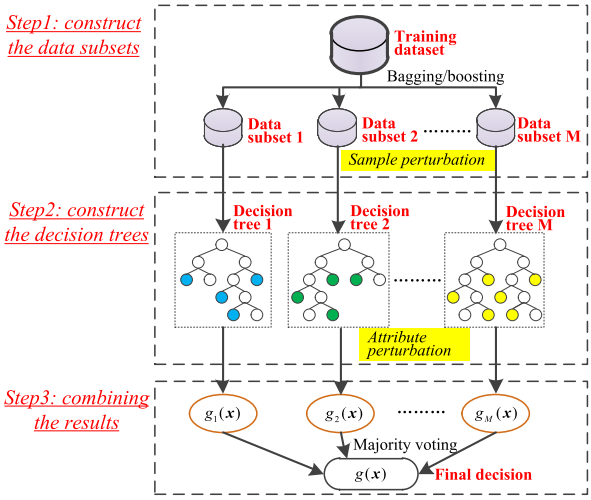



**FIGURE 16.** Flowchart of the RF algorithm.

can observe from Table 4 that all the methods adopting RF algorithm achieved high accuracy (at least 92%).

As illustrated in Fig. 15(b), boosting is another important ensemble technology. AdaBoost and XGBoost are two most widely adopted boosting methods. For instance, AdaBoost aims at converting several weak classifiers into a stronger classifier by iteratively assigning the misclassified samples with larger weights, so that its performance can be gradually enhanced. In [67], Mansour et al. evaluated the performances of five ML techniques in PD pulse sequence diagnosis, and they found that AdaBoost algorithm can achieve a lowest error (0.0282) among all methods. XGBoost is also a popular boosting method, which belongs to the category of GBDT. The objective function of XGBoost is defined as follows:

$$Obj = \sum_{i=1}^{N} \mathcal{L}(y_i, \hat{y}_i) + \sum_{m=1}^{M} \Omega(f_m) \tag{12}$$

where $y_i$ and $\hat{y}_i$ are respectively the true label and prediction of the $i$th sample, $\mathcal{L}(y_i, \hat{y}_i)$ is the loss function. $f_m$ is the function of the $m$th tree, $\Omega(f_m)$ denotes its complexity. In [102], the XGBoost algorithm was employed to identify the PD patterns of XLPE cables. By adjusting key parameters of XGBoost such as the type of base classifier, number of iterations, learning rate, etc., an overall accuracy rate of 96.93% was obtained. This was not only better than traditional ML algorithms like the decision tree (81.23%), SVM (85.06%), BPNN (87.36%), but also exceeded the RF method (89.66%).

Except for the above well-known EL methods, some other algorithms which adopt the ensemble ideas have also been applied to PD recognition. For example, the authors of [104] proposed a novel method namely the ensemble fuzzy-RBF neural networks, which integrated the advantages of fuzzy and EL methods. PD data from four typical insulation defects in GIS were employed to test the effectiveness of this method, and results showed that the classification rate can reach to 99.37±0.009. In [71], Yao et al. developed a novel ensemble framework considering both the combined feature sets and adaptive boost method. With these improvements, the overall recognition accuracy of UHF PD signals also reached a high level (above 91%).

### 6) OTHER METHODS

Except for the above mainstream algorithms, many other latest ML technologies have also been introduced to the type recognition of PD sources. Given the variety of classification principles and the awe-inspiring development of this field, we cannot include all of them in this paper. Instead, we only make a brief description of some representative literatures in Table 4.

## III. TYPE RECOGNITION BASED ON DEEP LEARNING ALGORITHM

Over the past decade, with the joint contribution of big data, massive computing power, and advanced learning algorithms,

the DL technologies have achieved remarkable success in almost every field. Compared with traditional ML algorithms, the most important advantages of DL are mainly reflected in automatic feature extraction ability and superior recognition ability. All these significant characteristics have made DL algorithms increasingly popular. In this section, we will make comprehensive review of the application of DL techniques in PD type recognition, including the DBN, DAN, CNN, RNN, GAN, GCN, DEL, etc. Representative literatures regarding to the above techniques are listed in Table 5.

### A. DBN-BASED METHODS

In the early stage, DBN is probably the most widely used DL algorithm. It was proposed by Hinton et al. in 2006, which consists of the generative part and discriminative part. The generative part is constructed by several layers called RBMs, while the discriminative part usually contains classification model such as softmax layer. In fact, we can also divide the training process of DBN into the following two steps: a) unsupervised training of the RBMs based on the CD method; b) supervised fine-tuning of the whole network. Structure of the RBMs and DBN are shown in Fig. 17.

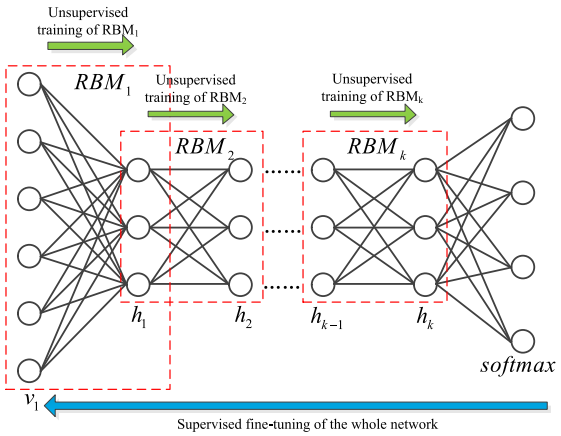



**FIGURE 17.** Structure of the RBMs and DBN.

As can be observed, RBM is an undirected probability graph model, which is consisted of two layers namely the visible layer and the hidden layer. The energy function of RBM can be formulated as:

$$E_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h}) = -\sum_{i=1}^{n_v} a_i v_i - \sum_{j=1}^{n_h} b_j h_j - \sum_{i=1}^{n_v}\sum_{j=1}^{n_h} w_{ij} v_i h_j \tag{13}$$

where $\boldsymbol{v} = \{v_i, i = 1, \ldots, n_v\}$ and $\boldsymbol{h} = \{h_j, j = 1, \ldots, n_h\}$ denote the visible unit and hidden unit, respectively, $a_i$ and $b_j$ are their corresponding biases. $w_{ij}$ is the weight between $v_i$ and $h_j$. Thus, learnable parameters of RBM can be expressed by $\boldsymbol{\theta} = [\boldsymbol{w}, \boldsymbol{a}, \boldsymbol{b}]$. When the state of visible unit $\boldsymbol{v}$ is known, the probability that hidden unit $\boldsymbol{h}$ is activated can be computed:

$$P_{\boldsymbol{\theta}}(h_j = 1|\boldsymbol{v}) = \sigma(b_j + \sum_{i=1}^{n_v} w_{ij} v_i) \tag{14}$$

Conversely, when the state of hidden unit $\boldsymbol{h}$ is known, the probability that visible unit $\boldsymbol{v}$ is activated can be computed:

$$P_{\boldsymbol{\theta}}(v_i = 1|\boldsymbol{h}) = \sigma(a_i + \sum_{j=1}^{n_h} w_{ji} h_j) \tag{15}$$

where $\sigma$ is the activation function. Because the units within the same visible layer or hidden layer are independent from each other, the above equations can be extended as:

$$P_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v}) = \prod_j P_{\boldsymbol{\theta}}(h_j|\boldsymbol{v}), \quad P_{\boldsymbol{\theta}}(\boldsymbol{v}|\boldsymbol{h}) = \prod_i P_{\boldsymbol{\theta}}(v_i|\boldsymbol{h}) \tag{16}$$

For each RBM, the training objective is to minimize the error between the reconstructed data and the real input data. To achieve this goal efficiently, Hinton et al. developed the CD algorithm in 2002. The CD algorithm first initializes parameter $\boldsymbol{\theta}$ randomly, and takes the training samples as the visible unit $\boldsymbol{v}^0$. Then, hidden unit $\boldsymbol{h}^0$ is calculated according to Equation (14). Afterwards, visible unit $\boldsymbol{v}^1$ is reconstructed by Equation (15). At last, calculates the hidden unit $\boldsymbol{h}^1$ again by Equation (14). Updating formulas of $\boldsymbol{\theta}$ can be written as follows:

$$\begin{aligned} \Delta w_{ij} &= \eta(< v_i^0 h_j^0 > - < v_i^1 h_j^1 >) \\ \Delta a_i &= \eta(< v_i^0 > - < v_i^1 >) \\ \Delta b_j &= \eta(< h_j^0 > - < h_j^1 >) \end{aligned} \tag{17}$$

where $\eta$ is the learning rate, $< >$ denotes the mathematical expectation. In practice, the mini-batch scheme is often used. Let $N$ be the number of a batch of training samples, then the training process can be rewritten as:

$$\begin{aligned} w_{ij,k} &= w_{ij,k-1} + \eta\left(\sum_{n=1}^{N}(v_{i,n}^{k-1} h_{j,n}^{k-1} - v_{i,n}^{k} h_{j,n}^{k})/N\right) \\ a_{i,k} &= a_{i,k-1} + \eta\left(\sum_{n=1}^{N}(v_{i,n}^{k-1} - v_{i,n}^{k})/N\right) \\ b_{j,k} &= b_{j,k-1} + \eta\left(\sum_{n=1}^{N}(h_{j,n}^{k-1} - h_{j,n}^{k})/N\right) \end{aligned} \tag{18}$$

Based on the above descriptions, now we can summarize the main steps of DBN as follows: a) sequentially train each RBM network in an unsupervised manner, that is, for each RBM, regard the output of the previous RBM as the visible layer, and update parameters $\boldsymbol{\theta}$ of current RBM according to Equation (18); b) fine-tuning the whole network (including all RBMs and the last classification layer) in a supervised manner.

Zhang et al. took the lead in applying DBN to the pattern recognition of PD data [112]. They used the PRPD spectrums to train the DBN model directly. Compared with SVM and BPNN, the overall accuracy of DBN were improved by 5.5% and 9.4%, respectively. Moreover, the recognition time were shortened by 38% and 49%, respectively. Jia et al. also adopted DBN to identify the typical insulation defects of power transformer in [111]. Jiang et al. introduced the ADAM

**TABLE 5. Summary of the DL methods for type recognition of PD signals.**

| Category | | Reference | Year | Sensor type | Data type | Major technology | Reported accuracy |
|---|---|---|---|---|---|---|---|
| DBN-based methods | | [111] | 2017 | IEC 60270 | Time-frequency spectrum | S-transform + 2D-PCA + DBN | At least 80% |
| | | [112] | 2016 | UHF | PRPD pattern | DBN | Overall accuracy of 96.4% |
| | | [113, 114] | 2020 | HFCT | Time-domain waveform | DBN + ADAM | At least 95% |
| | | [115] | 2020 | IEC 60270 | PRPD pattern | Four DBN-based approaches | Reached 99.8% |
| DAN-based methods | | [116] | 2020 | UHF | PRPD pattern | Stacked AE model | About 86.75% |
| | | [117] | 2019 | IEC 60270 + UHF | Time-domain waveform | SAE with a penalty term of KL divergence | Up to 99.7% |
| | | [118] | 2021 | HFCT | Time-frequency spectrum | a) SAE with a penalty term of transformed L1 norm; b) combining latent features of two parallel SAE models. | Up to 93.8% |
| | | [119] | 2020 | Coupling capacitor | Time-frequency spectrum | CWT + CAE | 100% for Architecture #1, 99.72% for Architecture #2. |
| | | [120] | 2019 | UHF | PRPS pattern | VAE + matching algorithm based on cosine distance | Up to 96.87% |
| | | [121] | 2019 | IEC 60270 | PRPD pattern | L2-VAE + Otsu threshold algorithm | 100% for dataset without unknown PD type, 98% for dataset with unknown PD type |
| | | [122] | 2019 | Unknown | 2D representation of PD activity | CVAE + combined data-driven and knowledge-driven model. | Up to 65% |
| CNN-based methods | Classical CNN model | [123] | 2020 | UHF | PRPD pattern | Cascaded CNN model | Overall accuracy of 99.64% |
| | | [124] | 2019 | HFCT | Time-domain waveform | Cascaded CNN models with different number of layers, convolution kernel sizes, activation functions, etc. | Average of 92.57% |
| | | [125] | 2021 | IEC 60270 | PRPD pattern | Cascaded CNN model for feature extraction, multiple FC layers for classification. | 99.6% for single-source PD, 96.7% for multi-source PD. |
| | | [126] | 2020 | Inductive sensor | PRPD pattern | AlexNet-liked CNN model | Reached 81.3% even at ±15% random white noise |
| | | [127] | 2020 | HFCT | 3D ( $q-n-\Delta t$ ) spectrum | Improved AlexNet model with Swish activation functions and dropout layers. | Up to 91.32% |
| | | [128] | 2019 | Acoustic | Time-frequency spectrum | BSS + cascaded CNN model | Reached 93.6% |
| | | [129] | 2022 | IEC 60270 | Pulse sequence analysis pattern | Pulse sequence analysis + cascaded CNN model | Up to 95.3% |
| | | [130] | 2018 | UHF | PRPS pattern | Cascaded CNN model + autoencoder | Up to 89.8% |
| | | [131] | 2019 | UHF | Time-domain waveform | Conditional variation auto-encoder + LCNN | Overall accuracy of 98.13% |
| | | [132] | 2019 | Acoustic | Time-frequency spectrum | Multiscale wavelet decomposition + mel-frequency cepstrum coefficients + cascaded CNN model | Average precision of 96.3% |
| | | [133] | 2021 | IEC 60270 | PRPD pattern | A customized CNN model based on the concept of residual neural networks. | More than 87% validation accuracy |
| | | [134] | 2021 | IEC 60270 | Marginal spectrum | VMD + Hilbert transform + ResNet18 | Average of 95% |
| | | [135] | 2018 | UHF | PRPD pattern | CNN with residual network structure | Average of 95.79% |
| | | [136] | 2019 | HFCT | Time-domain waveform | 1D-CNN | 97.38% for model-generated PD signals, 93.23% for real-world PD signals. |
| | | [137] | 2019 | UHF | Time-domain waveform | 1D-CNN | Up to 88.9% |
| | | [138] | 2021 | HFCT | Voltage amplitude-phase spectrum | 1D-CNN | At least 98% |
| | CNN-based transfer learning | [139] | 2021 | UHF | PRPD pattern + PRPS pattern | Model-based transfer learning, the base model is VGG-16. | Average precision of 91.5% |
| | | [140] | 2021 | IEC 60270 | PRPD pattern | DeepLoc CNN + model-based transfer learning | 98.40% for clean data, 90.80% for contaminated data. |
| | | [141] | 2021 | HFCT | PRPD pattern | Model-based transfer learning with ResNet50V2 and MobileNetV2 base models. | 96.2% for ResNet50V2_C3, 97.4% for MobileNet50V2_C3. |
| | | [142] | 2022 | UHF | PRPD pattern | Transfer learning with VGG, InceptionV3, and Resnet50 base models for feature extraction, SVM for classification. | Up to 98% |

**TABLE 5.** *(Continued.)* **Summary of the DL methods for type recognition of PD signals.**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | [143] | 2021 | UHF | Time-domain waveform | Feature-based transfer learning by domain adaptation, base model is residual CNN. | 98.9% for the experimental data, 89.96% for field data. |
| | | [144] | 2021 | UHF | Time-domain waveform | Improved 1D-CNN + domain adversarial transfer learning | 98.67% for source domain, >92% for target domain. |
| | Advanced CNN model | [145] | 2019 | UHF | Time-domain waveform | LCNN based on MobileNet-V2 | Overall recognition rate of 96.5% |
| | | [146] | 2020 | UHF | Time-domain waveform | LCNN based on MobileNet block, ShuffleNet block, EffNet block, respectively. | Up to 98.0% |
| | | [147] | 2021 | UHF | Time-domain waveform | GAN + MixNet DL model based on mixed depthwise convolution | Overall accuracy of 99.10% |
| | | [148] | 2020 | UHF | PRPD pattern | One-shot learning using a Siamese neural network. | Classification accuracy of 98.65% |
| | | [149] | 2023 | HFCT | PRPD pattern | MobileNetV2 + depth-width joint pruning | 98.28% at 23dB, 86.41% at 16dB |
| | | [150] | 2020 | IEC 60270 | Time-frequency spectrum | VMD-CWD + cross-layer feature fusion CNN | Up to 99.5% |
| | | [151] | 2022 | IEC 60270 | PRPD pattern | Multi-channel fusion of PRPD figures at different applied voltage frequencies, base model is cascaded CNN. | Almost 100% |
| | | [152] | 2021 | IEC 60270 | Time-domain waveform | 1D-CNN + PAM | Average accuracy of 96.7% |
| | | [153] | 2022 | Capacitive sensor | Time-domain waveform | 1D-CNN + Grad-CAM | Classification accuracy of 98.7% |
| | | [154] | 2023 | IEC 60270 | PRPD pattern | Cascaded CNN + SHAP | Average accuracy of 96.2% |
| | | [155] | 2022 | UHF | Time-domain waveform | 1D-CNN + DNAS | Reached 97.625% |
| RNN-based methods | | [156] | 2018 | UHF | PRPD pattern | LSTM network | Overall classification accuracy of 96.74% |
| | | [157] | 2021 | IEC 60270 | Sequence of PD events | A modified LSTM network | Single-class accuracy of 99%, multi-class accuracy of 43% |
| | | [158] | 2020 | Single layer coil | Time-domain waveform (VSB ENET Dataset) | Seasonal and trend decomposition using loess + feature engineering + LSTM | Average Precision of 79% |
| | | [159] | 2021 | Acoustic | Time-domain waveform | LMD + feature engineering + LSTM | Overall accuracy of 99.25% |
| | | [160] | 2020 | UHF | PRPD pattern | LSTM + self-attention neural network | Overall accuracy of 94% |
| GAN-based methods | | [161] | 2020 | HFCT | Time-domain waveform | DCGAN | Not found |
| | | [162] | 2022 | UHF | PRPD pattern | CGAN + ResNet | Up to 98.75% |
| | | [163] | 2022 | HFCT | Time-frequency spectrum | Improved GAN based on the WGAN-GP and CGAN. | Up to 99.1% |
| | | [164] | 2021 | HFCT | Time-frequency spectrum | SWT + AC-BEGAN | $F1_{mean}$ of 98.3% |
| | | [165] | 2022 | UHF | Time-domain waveform | Dual discriminator CGAN + automatic CNN | 99.15% accuracy on unbalanced samples |
| | | [166] | 2023 | UHF | Time-frequency spectrum | SWT + ACGAN + self - attention mechanism | Reached 95.75% for small and unbalanced datasets |
| GCN-based methods | | [167] | 2021 | HFCT | Time-frequency spectrum | GCN with SAGPool mechanism | Average accuracy of 98.95% for experimental data, 91.1% for field data. |
| | | [168] | 2022 | UHF | Time-domain waveform | GCN + ChannelSortPooling + capsule network | Average accuracy of 97.57% for experimental data, 82.07% for field data. |
| | | [169] | 2023 | HFCT | PRPD pattern | Knowledge distillation + GNN | Average accuracy is improved by 18%, and memory usage is also greatly reduced. |
| | | [170] | 2021 | UHF | PRPS pattern | ResNet in the feature space, PD-tailored KG and GCN in the knowledge space. | Accuracy of 99.58% on the experimental dataset, 95.67% on the field dataset. |
| DEL-based methods | | [171] | 2021 | UHF | PRPD pattern | CNN + averaging ensemble | Detects the unknown PRPD faults with 100% accuracy |
| | | [172] | 2021 | IEC 60270 | Time-domain waveform | Cable-specific adaption + neural network ensemble | Up to 98.65% for Ensemble 1, up to 97.36% for Ensemble 2. |
| | | [173] | 2022 | UHF + acoustic | PRPD pattern + time-domain waveform | Deep residual CNN + stacking-based ensemble learning | Overall accuracy of 97.5% |
| | | [174] | 2023 | UHF | Time-domain waveform | Autoencoder and 1D-CNN as the base-learners, wide and deep neural network as the | 94.0% accuracy for dataset A, 84.7% accuracy for dataset B. |

**TABLE 5.** *(Continued.)* **Summary of the DL methods for type recognition of PD signals.**

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | meta-learner. | |
| Other DL methods | [175] | 2019 | IEC 60270 | Time-domain waveform | VMD + Hilbert marginal spectrum + SAE + DNN | Overall accuracy of 98.5% |
| | [176] | 2019 | UHF | PRPD pattern | CNN + LSTM | Average accuracy of 93.25% |
| | [177] | 2021 | UHF | Time-domain waveform | CNN + LSTM | Overall accuracy of 97.9% |
| | [178] | 2020 | Single layer coil | Time-domain waveform (VSB ENET Dataset) | DWT + oversampling technology + LSTM | Average precision of 81% |
| | [179] | 2021 | Single layer coil | Time-domain waveform (VSB ENET Dataset) | CNN + LSTM with attention mechanism | 95.16% for no-PD, 94.44% for PD. |
| | [180] | 2021 | IEC 60270 | Periodic PRPD pattern (P-PRPD) | Damped oscillating AC voltage + CNN + LSTM | Overall accuracy of 92% |
| | [181] | 2018 | UHF | Time-domain waveform | Model building based on multi-column CNN and LSTM, model training based on TL approach inspired by manifold learning. | Overall accuracy of 98.20% |

training algorithm to the DBN [113], [114]. Testing results verified the effectiveness of the ADAM-DBN method over conventional DBN method and other ML methods such as SVM and BPNN. In [115], authors made a comparison of many ML algorithms in PD recognition. They found that when DBN was used as both feature extractor and classifier, the identification accuracy can reach to 99.8%, much higher than other traditional ML methods. In addition, the DBN was almost immune to noise.

### B. DAN-BASED METHODS

As another popular unsupervised DL methodology, the DAN aims at learning effective abstract features from unlabeled samples automatically. The AE usually consists of two parts namely the encoder and decoder. The encoder maps the high-dimensional input samples to the low-dimensional space, in which the abstract representation of input samples can be achieved. While the purpose of decoder is to reconstruct the input samples from those abstract features. Structure of AE is illustrated in Fig. 18, where $\boldsymbol{X} = \{\boldsymbol{x}_i \in \boldsymbol{R}^d\}_{i=1}^n$ denotes the input samples, $\hat{\boldsymbol{X}} = \{\hat{\boldsymbol{x}}_i \in \boldsymbol{R}^d\}_{i=1}^n$ denotes the output samples, $d$ and $n$ are the dimension and number of input samples, respectively.

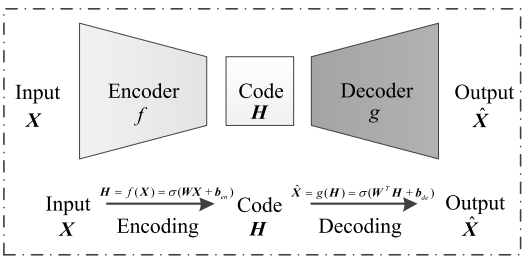



**FIGURE 18.** Structure of the AE.

Generally, the encoder and decoder are both composed of FC layers. Let $(\boldsymbol{W}, \boldsymbol{b}_{en})$ and $(\boldsymbol{W}^T, \boldsymbol{b}_{de})$ be the parameters of encoder and decoder, respectively, then the code vector $\boldsymbol{H}$ in latent space can be calculated by $f(\boldsymbol{X})$, and the output $\hat{\boldsymbol{X}}$ of AE is obtained by $g(\boldsymbol{H})$. The goal of AE is to minimize the difference between $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$, thus the objective function can be written as:

$$\underbrace{\arg\min_{\boldsymbol{W},\boldsymbol{b}} J(\boldsymbol{X},\hat{\boldsymbol{X}}) := \frac{1}{2}\sum_{i=1}^{n}\left\|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\right\|_2^2}_{Squared\ error\ loss\ function} \quad Or$$

$$\underbrace{\arg\min_{\boldsymbol{W},\boldsymbol{b}} J(\boldsymbol{X},\hat{\boldsymbol{X}}) := -\sum_{i=1}^{n}\left[\boldsymbol{x}_i \log(\hat{\boldsymbol{x}}_i) + (1-\boldsymbol{x}_i)\log(1-\hat{\boldsymbol{x}}_i)\right]}_{Cross-entropy\ loss\ function} \tag{19}$$

By stacking multiple AE models sequentially, the stacked AEs network can be obtained, which is a typical DAN model. Besides, there are many other improved versions of DAN models [182], such as the SAE, DAE, CAE, VAE, etc. Summary of the improved DAN models is given in Table 6.

In [116], the authors employed the AE model to identify UHF PD signals and on-site noise. In particular, the proposed AE model was trained by noise data only. Thus, if the input data was noise, the resulting reconstruction error should be less than a certain threshold. On the contrary, reconstruction error should be greater than the threshold. Testing results indicated that this model achieved an accuracy of 86.75%.

Duan et al. developed a DL model to recognize the current waveforms induced by different PD defects [117]. This model contained two parts: a SAE layer for automatic feature extraction and a softmax layer for classification. To introduce sparsity, authors added a penalty term namely KL divergence to the cost function, as shown in Table 6. Results showed that with well-selected hyper-parameters, the recognition rate can reach to 99.7%. Another SAE-based method was reported in [118], in which the transformed L1 norm was adopted. In addition, the authors designed two parallel SAE models for feature extraction of 1-D PD time-domain signals and 2-D

**TABLE 6. Summary of some improved DAN models.**

| Category | Purpose | Realization approach |
|---|---|---|
| SAE | Enhance the sparsity of the features in latent space. | Add a regularization term about activation degree to the loss function, such as the Kullback-Leibler divergence term. Corresponding loss function: or $J_{SAE}(\boldsymbol{W},\boldsymbol{b}) = J(\boldsymbol{X},\hat{\boldsymbol{X}}) + \beta\sum_{j=1}^{m} KL(\rho \,\|\, \hat{\rho}_j)$, where $a_j(\boldsymbol{x}_i)$ denotes the activation value of the $j$th hidden node regarding to $\boldsymbol{x}_i$, $\hat{\rho}_j$ is the average value of the $j$th hidden node. |
| DAE | Enhance the robustness of the features in latent space. | Introduce a degradation process to AE, and the noise-free samples are reconstructed with the noisy samples, so that the extracted abstract features are not susceptible to noise and have stronger robustness. |
| CAE | To better retain the spatial information of image samples. | Replace the FC layers in AE with convolutional layers. Specifically, in encoder, using downsampling layers + convolutional layers. In decoder, using upsampling layers + convolutional layers. |
| VAE | Force the hidden variable to meet specific distribution, thus the capability of data generation can be achieved. | **Step1:** encoding. For input samples $\boldsymbol{X}$, construct the expectation $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$ of the hidden variable $\boldsymbol{Z}$ by the encoding probabilistic model $Q(\boldsymbol{Z} \mid \boldsymbol{X})$. **Step2:** sampling. Randomly generating samples of $\boldsymbol{Z}$ according to $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$. **Step3:** decoding. Based on the samples of $\boldsymbol{Z}$, generating the reconstructed samples by the decoding probabilistic model $P(\boldsymbol{X} \mid \boldsymbol{Z})$. Corresponding loss function: $J_{VAE} = D_{KL}(Q(\boldsymbol{Z} \mid \boldsymbol{X}) \mid P(\boldsymbol{Z})) - E_{Q(\boldsymbol{Z}\mid\boldsymbol{X})}(\log(P(\boldsymbol{X} \mid \boldsymbol{Z})))$ |

PD time-frequency spectrums, respectively. Then, two kinds of latent features were combined to enrich fault information.

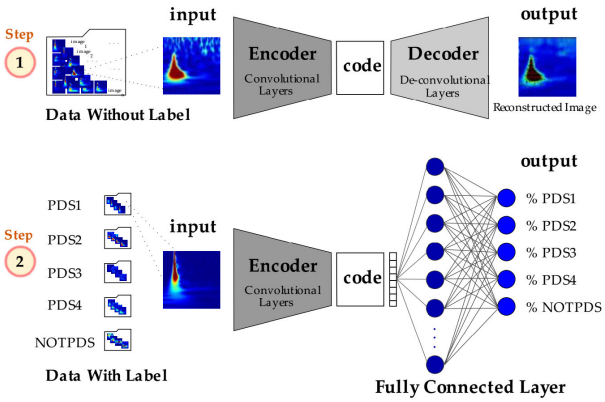



**FIGURE 19. A typical training process of CAE model for PD type recognition. Figure adopted from [119].**

In recent years, advantages of convolution operators in image processing have been fully verified. Therefore, the CAE was developed by integrating the convolution operators into AE, as illustrated in Table 6. In [119], the authors employed CAE to classify the PD data of MV switchgear. The complete training process is shown in Fig. 19. First, the PD signals were transformed into time-frequency images by using CWT. Then, the whole CAE model was trained by the unlabeled images. Finally, decoder part of the trained model was replaced by the classification part, and fine-tuning of the newly constructed model was accomplished by the labeled PD time-frequency images.

Unlike other AE models that generate latent feature vector $\boldsymbol{Z}$, VAE generates the distribution parameters, that is, the expectation $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$. As shown in Table 6, the cost function of VAE consists of two parts. The first term $D_{KL}(Q(\boldsymbol{Z}|\boldsymbol{X})|P(\boldsymbol{Z}))$ is used to force the generated vector follow a specific normal distribution, and the second term $-E_{Q(\boldsymbol{Z}|\boldsymbol{X})}(\log(P(\boldsymbol{X}|\boldsymbol{Z})))$ denotes reconstruction error. In [120] and [121], the VAE was adopted for PD recognition. For example, Dai et al. applied VAE to extract the deep features from PRPS images [120]. Compared with DBN eigenvalues, CNN eigenvalues, the extracted VAE feature vector showed better performance in PD type discrimination.

In [122], the CVAE model was used as a 2-D visualization tool to convert PD signals of practical hydrogenerator into the 2-D visualization latent space. Therefore, the experts can visually select the best training dataset by labelling those significant samples (i.e., samples nearby the conflict zones). Honestly, the idea behind this work is notably important, that is, to combine the expert knowledge with data-driven model. This research paradigm is an increasing research hotspot in the ML field, and we will discuss this issue in a later section.

### C. CNN-BASED METHODS

Compared with DBN and DAN, the CNN-based DL methods have gained much attention in recent years. First, CNN has relatively few parameters due to its characteristics like the weight sharing, local connection, pooling operation. Second, the nature of convolution determines that features learned by CNN are invariant to interference factors like the translation, distortion, rotation, and scaling. Third, the structure of CNN is flexible and easy to expand. According to different improvement ideas, researchers have proposed a series of high-performance CNN models, which can be divided into three categories: a) classical CNN models; b) CNN-based transfer learning models; and c) advanced CNN models. We depict the taxonomy of CNN-based methods in Fig. 20, and detailed discussion about these methods and their applications in PD type recognition will be given in the later parts of this subsection.

#### 1) CLASSICAL CNN MODELS

Classical CNN is a kind of deep neural network including structures like the convolutional layer, pooling layer, FC layer, etc. According to the network structure, we further divide the classical CNN models into the cascaded CNN models, the Inception-based CNN models, the ResNet-based CNN models, the DenseNet-based CNN models, and the 1D-CNN models.

Early CNN models adopt the cascaded network structure, which achieve automatic feature extraction and classification by stacking convolutional layers, pooling layers, FC layers

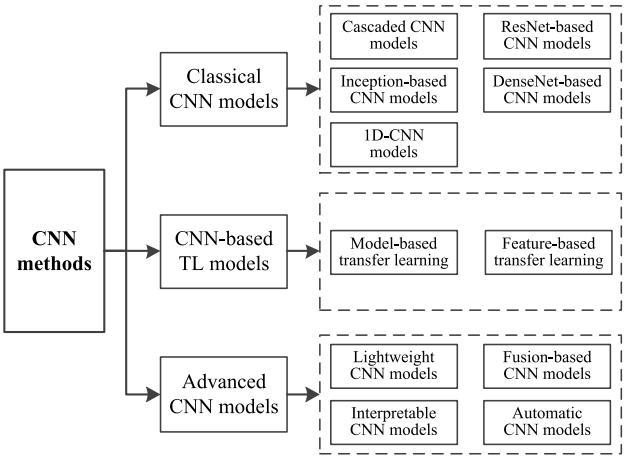



**FIGURE 20.** Taxonomy of the CNN methods.

and finally the classification layer (e.g., a softmax layer). Typical cascaded CNN model is shown in Fig. 21(a).

Later, researchers developed the Inception-based CNN model which consists of many so-called inception blocks. In inception block, convolution kernels of different sizes such as $1 \times 1$, $3 \times 3$, $5 \times 5$ are employed to extract information from different spatial sizes, and the resulting feature maps are concatenated along the channel dimension to form the final feature map. Fig. 21(b) shows the structure of the Inception-v1 block. Benefits from this kind of multi-branch structure, Inception-based CNN model can achieve better classification performance while reducing the number of model parameters.

Another effective way to promote the performance of deep CNN models is to introduce the residual blocks. The key innovation of this technique is to connect the input $x$ directly to the output $f(x) - x$ (also known as residual function) by crossing multiple intermediate learning layers, as illustrated in Fig. 21(c). Therefore, the final output of a residual block is formulated as $f(x)$. In ResNet, the input can be propagated forward faster through the cross-layer connections, which greatly alleviates the vanishing gradient problem.

The DenseNet is an extension of ResNet, in which all layers are directly connected. The major difference between them is that the outputs of DenseNet are concatenated, rather than added like in ResNet. Let $x_l$ denotes the output feature map of the $l$th bottleneck block, $x_0$ is the input, then $x_l$ can be expressed as:

$$x_l = [x_0, x_1, \ldots, x_{l-1}] \tag{20}$$

where $[\cdot]$ is the concatenating operator. In Fig. 21(d), we illustrate the typical structure of a dense block with three bottlenecks.

In recent years, the 1D-CNN have received much attention due to its ability to process 1-D signals directly. In 1-D convolution, one selects sequential segments from the data along the time axis and performs the same transformation on each segment.

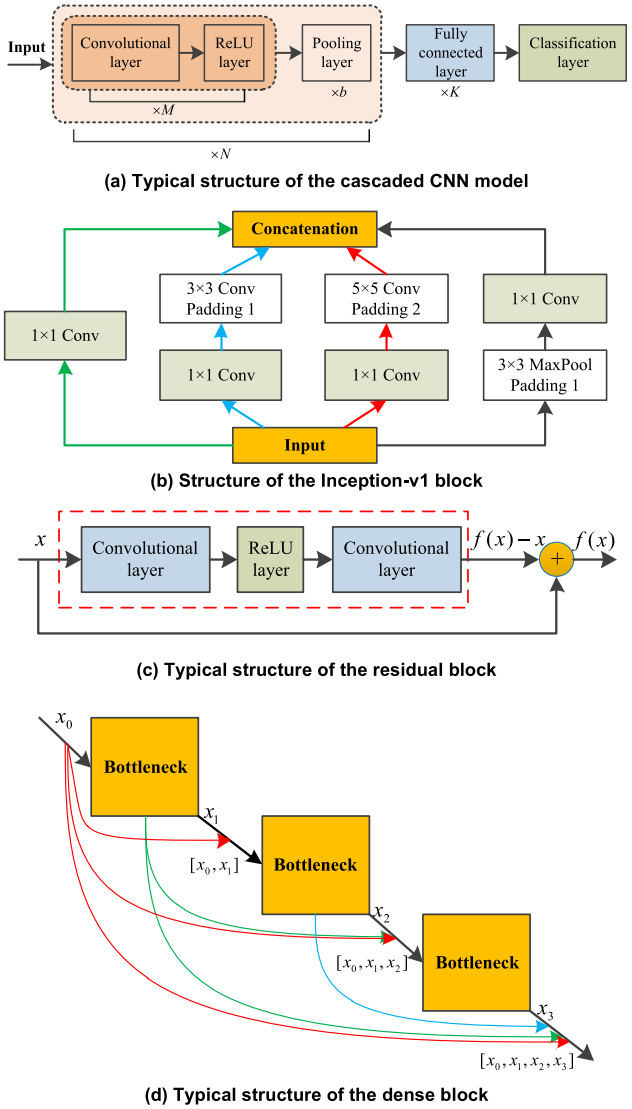



**FIGURE 21.** Structures of some classical CNN models.

The above classical CNN models have already been applied in the field of PD recognition, as we summarized in Table 5. Among them, authors of [123], [124], [125], [126], [127], [128], [129], [130], [131], and [132] adopted the cascaded CNN models. For example, in [123], Do et al. designed a cascaded CNN model for classifying PRPDs of the power transformer. In [126], an AlexNet-liked CNN model was used to identify the PRPD images collected from GIS, and 81.3% accuracy can be achieved even at ±15% random white noise. Besides, references [133], [134], and [135] adopted the ResNet-based CNN models for PD recognition. For instance, Borghei et al. built a customized ResNet model for discrimination of single and multi-source corona discharges in [133]. The results showed that the proposed ResNet model can achieve more than 87% validation accuracy only after 40 iterations. Gao et al. developed another CNN model for PD diagnosis based on

ResNet18 [134], and reported recognition accuracy of on-site PD data can reach to 95%.

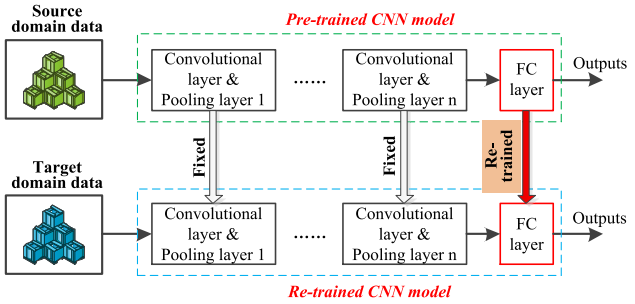
Considering that the original PD data measured by many monitoring systems are 1-D signals, 1D-CNN have become increasingly popular in the PD diagnosis field. For example, an end-to-end DL framework was presented based on 1D-CNN to identify the raw PD signals from power cables [136]. In that paper, the designed 1D-CNN model had only one 1-D convolutional layer, one pooling layer, and one FC layer. Therefore, compared to 2D-CNN model, the 1D-CNN model had a much smaller number of parameters. Even so, this model can learn features automatically from the noisy PD signals and realize the type recognition.

### 2) CNN-BASED TRANSFER LEARNING MODELS

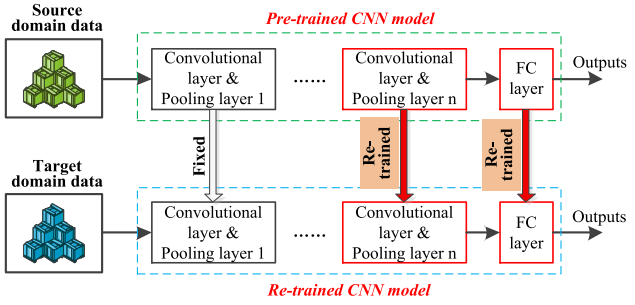
As a supervised DL algorithm, the high performance of CNN is based on the following two conditions: a) the number of training samples is sufficient and the fault type of each sample is accurately labeled; b) the dataset used to train the model (called the source-domain dataset) and the dataset used to test the model (called the target-domain dataset) follow the same distribution. Unfortunately, both the above conditions are difficult to satisfy in practice. TL is one of the most promising tools to solve the above problem [183], [184]. As for how to achieve the knowledge transfer from source domain to target domain in practice, most research works are focused on the model-based TL methods and feature-based TL methods. In PD diagnosis area, these two TL schemes have also been applied successfully.

The major technique adopted in model-based TL is called fine-tuning, which aims at retraining some learnable layers of the pre-trained DL model (already trained by large-scale source domain dataset) by using the target domain dataset. Concretely, there are two commonly used transfer strategies for model-based TL methods. In the first strategy, all learnable layers of the pre-trained DL model are fixed except the FC layers. Then, the FC layers should be re-trained based on the target domain dataset. In the second strategy, not only the FC layers, but also some other learnable layers should be re-trained. Fig. 22 illustrates the above two model-based TL strategies. Note that for both strategies, if the learning tasks of the source domain and target domain are different, the FC layers should be re-designed to adapt to the target domain learning task.

In [139], a TL method based on the VGG-16 network was proposed for PD recognition. At first, the parameters of the convolutional layers and the pooling layers of the pre-trained VGG-16 model were frozen. Then the remaining FC layers were re-trained by using the UHF PD spectrograms. Results showed that a satisfied accurate rate was obtained using only small amounts of training data. In [140] and [141], the model-based TL strategy was also adopted. Moreover, performances of different CNN models were compared when they were used as the base model for TL. For instance, four CNN architectures including the DeepLoc-Net, Squeeze-Net,




(a) TL strategy with only FC layers re-trained




(b) TL strategy with FC layers and other learnable layers re-trained

**FIGURE 22.** Two commonly used model-based TL strategies.

Inception-Net and VGG-Net were applied and compared in [140], and the authors found that the DeepLoc-Net with fixed scale binary PRPD images can get the best performance. In [142], the authors also introduced three CNN architectures as the base models for TL, which were the VGG, Inception-V3 and Resnet50. However, they replaced the FC layers with SVM classifier for PD classification. Therefore, advantages of CNN models in feature extraction and the SVM classifier in small data classification can be well combined. Although the model-based TL method is quite simple and effective, its limitations are also obvious. First, the similarity of the data distribution between source domain and target domain should be high enough, otherwise it may lead to negative transfer. Second, since different fine-tuning strategies can result in different effects, they should be adjusted according to the specific target data or tasks.

The other important scheme is the feature-based TL. The key of this scheme is to find a mapping function $g(\cdot)$, so that the source domain dataset $\boldsymbol{X}_S$ and target domain dataset $\boldsymbol{X}_T$ can have the similar distribution after being mapped to the new feature space by $g(\cdot)$. Then, the training and testing steps can be conducted in the new feature space, that is, using data $g(\boldsymbol{X}_S)$ to train the CNN model, and predicts the class label of $g(\boldsymbol{X}_T)$ by the trained model. We depict this process in Fig. 23.

Common practice to implement the feature-based TL can be categorized into the STM method and the AL method.

In STM method, the statistical criteria (e.g., mean value or high-order moment) are adopted to measure the discrepancy between source and target domains, and the domain-invariant features can be obtained by minimizing the distribution

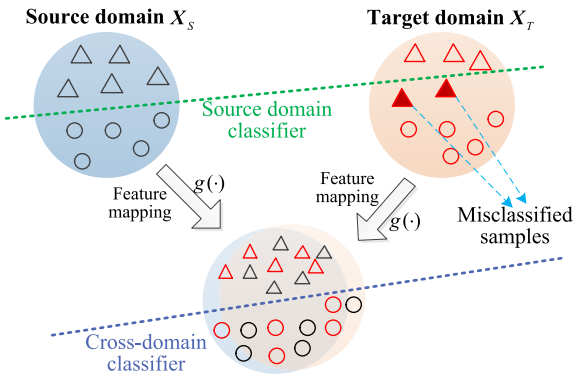



**FIGURE 23.** Illustration of the feature-based TL strategy.

discrepancy between the two domains. For example, a STM-based TL method was proposed in [143] for PD diagnosis of GIS. In this work, a novel loss function considering four parts was designed to train the CNN model, which are the source domain loss $\mathcal{L}_S$, target domain loss $\mathcal{L}_T$, MMD loss $\mathcal{L}_D(\boldsymbol{X}_S, \boldsymbol{X}_T)$, sliced Wasserstein distance loss $\mathcal{L}_C(\boldsymbol{H}_S, \boldsymbol{H}_T)$. Thus, the training objective can be formulated as [143]:

$$\min \mathcal{L}_S + \alpha\mathcal{L}_T + \beta\mathcal{L}_D(\boldsymbol{X}_S, \boldsymbol{X}_T) + \gamma\mathcal{L}_C(\boldsymbol{H}_S, \boldsymbol{H}_T) \quad (21)$$

where $\alpha$, $\beta$, $\gamma$ are the weight coefficients, $\boldsymbol{H}_S$ and $\boldsymbol{H}_T$ are the output category probability. By introducing additional regularization terms, the proposed CNN model can learn the domain-invariant and class-discriminative features. Results showed that by using the TL scheme, recognition accuracy of PD data was improved by 14.1% compared with using CNN directly.

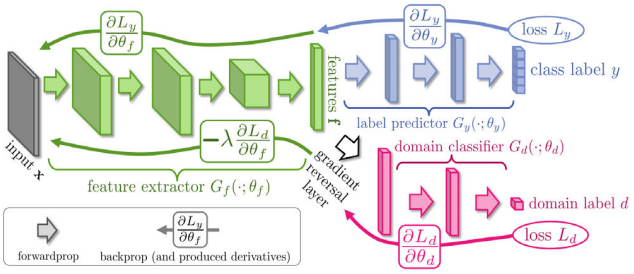



**FIGURE 24.** Structure of the DANN model. Figure adopted from [185].

As for the AL method, the feature alignment is achieved by a game-like learning mechanism namely the adversarial learning. Two famous AL-based TL methods are the DANN [185] and ADDA [186]. Take DANN method as an example, its network structure is demonstrated in Fig. 24. The training objective of the feature extractor of DANN is to minimize the loss of the label predictor, and simultaneously maximize the loss of the domain classifier. Obviously, the training process of this model is like playing a "minimax" game. Interested readers can refer to [185] for more details.

Likewise, the AL-based TL method has also been applied to the PD recognition. For example, Wang et al. developed a novel method for GIS PD diagnosis by combining the 1D-CNN and DANN [144]. In addition, during the adversarial training process, two domain classifiers were introduced to maximize their decision boundaries. Thus, the trained model can learn more class-discriminative representations of the raw PD data. The superiority of the proposed method over traditional 1D-CNN, 2D-CNN and some model-based TL methods were demonstrated by a series of testing results.

Considering that the problems such as lack of labelled on-site data and data distribution discrepancy are very common in PD fault diagnosis field, it can be predicted that the TL-based diagnostic methods will be a long-term research focus.

### 3) ADVANCED CNN MODELS

Besides the aforementioned methods, there are still many advanced CNN models been developed. Here, we divide them into four categories: a) lightweight CNN models; b) fusion-based CNN models; c) interpretable CNN models; and d) automatic CNN models.

In industrial applications, lightweight CNN methods are essential to reduce the hardware costs, improve diagnostic efficiency, and facilitate localized deployment of the models. A widely adopted lightweight CNN design idea is to reduce the number of the convolution kernels or the feature channels reasonably, or design more efficient convolutional operations. Some typical examples include the DSC, shuffle units, and SSC. Based on these novel modules, Wang et al. proposed a series of lightweight CNN models for GIS PD identification [145], [146], [147].

For example, they designed a MCNN based on the DSC modules in [145]. Unlike the traditional convolution operator, there are two parts namely the depthwise convolution and pointwise convolution in DSC, as shown in Fig. 25. Based on this new kind of convolution operator, the proposed MCNN model only took 7.3 seconds to finish the training, and the recognition accuracy can reach 96.5%.

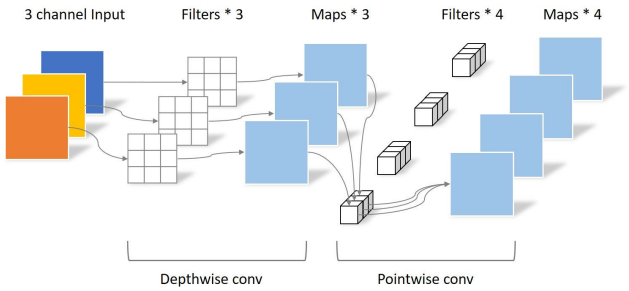



**FIGURE 25.** Structure of the DSC module.

In [149], another important technology namely the model pruning was adopted to build a lightweight CNN for PD diagnosis. Specifically, the MobileNetV2 was used as the base model, then the block-level pruning and the filter-level pruning were carried out along the depth and width directions, respectively. After the depth-width joint pruning, recognition accuracy was comparable to that of MobileNetV2, while

9.9 times of parameter compression and 2.3 times of inference acceleration were achieved.

The fusion CNN is also an effective variation, which aims at fusing multi-source features. For instance, an CNN model with cross-layer feature fusion scheme was proposed in [150]. Specifically, the outputs of several pooling layers were concatenated to form an augmented feature map. Benefit from this fusion scheme, the PD recognition accuracy was improved by 2.0%. Another fusion-based CNN model for PD diagnosis of the oil-paper insulation was reported in [151]. The main innovation of this paper lied in the channel-level fusion of the PD gray spectrums under different excitation frequencies (i.e., 50 Hz, 30 Hz, 5 Hz) to form a 3-channel feature map. Compared with traditional CNN, the accuracy of the multi-channel CNN method is increased by 1.2%.

With the widespread application of the ML algorithms in industrial fields, more and more researchers have noticed that the interpretability of the diagnostic results is almost equally significant as the diagnostic accuracy. Because a reasonable and convincing explanation of the results is the key for the relevant experts to take necessary measures. Fortunately, there are many techniques that can help researchers to understand how the ML model makes predictions, such as t-SNE, CAM, Grad-CAM, Shapley value, etc.

Recently, these techniques have been applied to the PD diagnosis field. For example, an interpretable 1D-CNN model for classification of PD waveforms from 3D-printed dielectric samples was reported in [153]. In this work, an interpretable attention model based on Grad-CAM was added. As a typical post-processing technique, Grad-CAM can intuitively show which regions of the PD waveform are most conducive to the type identification. One of the Grad-CAM importance colormaps is shown in Fig. 26. One can readily observe that the trained model focuses more on the pulse rising edge of the waveform when making a label prediction. Besides, the PAM [152] and Shapley values [154] have also been applied to interpret the results of PD diagnosis. These interpretable diagnostic results provide the key evidence for domain experts to confirm the PD fault and make subsequent maintenance strategies.

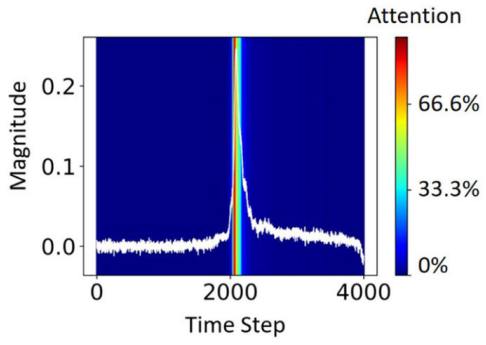



**FIGURE 26.** An example of the implementation of Grad-CAM in PD waveform. Figure adopted from [153].

Last, we focus on the automatic CNN technologies. We know that existing best-performing CNN models are all carefully designed by experts, which requires professional ML knowledge and a lot of trials and errors. Therefore, the AutoML has become another research hotspot in recent years. Obviously, a method that can automatically construct an optimized CNN architecture will be very desirable for the PD application. In [155], Jing et al. proposed an automatic search method of neural network architecture for PD recognition. First, a factorized hierarchical search space was adopted to design the CNN architecture. Then the Gumbel-softmax was used to replace the original softmax function for better randomness. Finally, the optimal CNN model was obtained by the DNAS strategy. Results demonstrated the superiority of the DNAS method over some traditional CNN models in term of recognition performance and generalization ability.

### D. RNN-BASED METHODS

Unlike the CNN models which are suitable for processing image data, the RNN models are designed for sequential data. The key characteristic of RNN is its memory ability by using neurons with self-feedback connections. Let vector $\boldsymbol{x}_t \in \mathbb{R}^M$ denotes the input of the network at time $t$, $\boldsymbol{h}_t \in \mathbb{R}^D$ denotes the state of the hidden layer, then $\boldsymbol{h}_t$ is not only related to the input $\boldsymbol{x}_t$ at time $t$, but also to the hidden layer state $\boldsymbol{h}_{t-1}$ at time $t-1$. Thus, the updating computation of $\boldsymbol{h}_t$ and output $\boldsymbol{y}_t$ can be formulated as:

$$\begin{aligned} \boldsymbol{h}_t &= \sigma_h(\boldsymbol{U}\boldsymbol{h}_{t-1} + \boldsymbol{W}\boldsymbol{x}_t + \boldsymbol{b}_h) \\ \boldsymbol{y}_t &= \sigma_y(\boldsymbol{V}\boldsymbol{h}_t + \boldsymbol{b}_y) \end{aligned} \tag{22}$$

where $\boldsymbol{U}, \boldsymbol{W}, \boldsymbol{V}, \boldsymbol{b}_h$ and $\boldsymbol{b}_y$ are the network parameters, $\sigma_h$ and $\sigma_y$ are the nonlinear activation functions.

The main problem of RNN is that it is difficult to model the dependencies between states with long time intervals due to the gradient explosion or vanishing gradient problems. This is known as the long-term dependence problem. A reasonable solution is to introduce the gating mechanism to control the speed of information accumulation, including selectively adding new information and forgetting previously accumulated information. LSTM network is probably one of the most famous gated RNNs. In LSTM, one internal state and three gates are employed [187]: a) internal state $\boldsymbol{c}_t \in \mathbb{R}^D$: linearly transmits information between internal states, and nonlinearly outputs information to the hidden state $\boldsymbol{h}_t \in \mathbb{R}^D$; b) forget gate $\boldsymbol{f}_t \in [0,1]^D$: controls the internal state $\boldsymbol{c}_{t-1}$ that how much information needs to be forgotten; c) input gate $\boldsymbol{i}_t \in [0,1]^D$: controls the temporary internal state $\tilde{\boldsymbol{c}}_t$ that how much information needs to be retained; d) output gate $\boldsymbol{o}_t \in [0,1]^D$: controls the internal state $\boldsymbol{c}_t$ that how much information needs to be outputted to the hidden state $\boldsymbol{h}_t$.

The structure diagram of a standard LSTM cell is shown in Fig. 27, and its calculation formulas are given

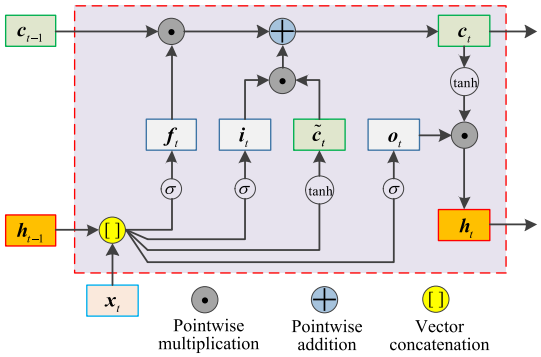



**FIGURE 27.** Structure diagram of a LSTM cell.

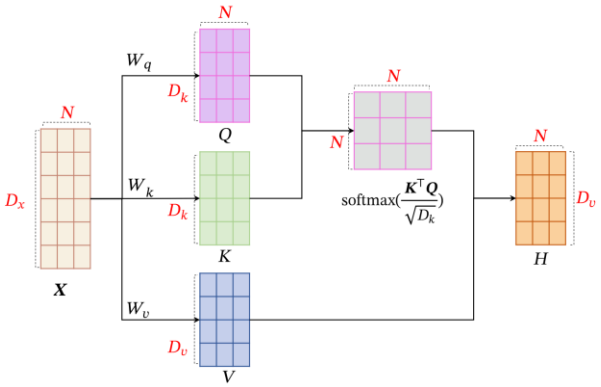



**FIGURE 28.** Computation process of the self-attention block. Figure adopted from [187].

as follows:

$$\begin{aligned}
\boldsymbol{i}_t &= \sigma(\boldsymbol{W}_i \boldsymbol{x}_t + \boldsymbol{U}_i \boldsymbol{h}_{t-1} + \boldsymbol{b}_i) \\
\boldsymbol{f}_t &= \sigma(\boldsymbol{W}_f \boldsymbol{x}_t + \boldsymbol{U}_f \boldsymbol{h}_{t-1} + \boldsymbol{b}_f) \\
\boldsymbol{o}_t &= \sigma(\boldsymbol{W}_o \boldsymbol{x}_t + \boldsymbol{U}_o \boldsymbol{h}_{t-1} + \boldsymbol{b}_o) \\
\tilde{\boldsymbol{c}}_t &= \tanh(\boldsymbol{W}_c \boldsymbol{x}_t + \boldsymbol{U}_c \boldsymbol{h}_{t-1} + \boldsymbol{b}_c) \\
\boldsymbol{c}_t &= \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \tilde{\boldsymbol{c}}_t \\
\boldsymbol{h}_t &= \boldsymbol{o}_t \odot \tanh(\boldsymbol{c}_t)
\end{aligned} \tag{23}$$

where $\{\boldsymbol{W}_i, \boldsymbol{W}_f, \boldsymbol{W}_o, \boldsymbol{W}_c, \boldsymbol{U}_i, \boldsymbol{U}_f, \boldsymbol{U}_o, \boldsymbol{U}_c, \boldsymbol{b}_i, \boldsymbol{b}_f, \boldsymbol{b}_o, \boldsymbol{b}_c\}$ are the network parameters, $\odot$ denotes the pointwise multiplication.

In [156], the LSTM network was employed to diagnose the GIS PD data. The authors constructed PD sequence data by combining PRPD signatures along different power cycles. Another improved LSTM network was proposed in [157] for superimposed PD patterns classification. To achieve multi-label classification, the authors replaced the last layer of a standard LSTM network from the softmax layer to a sigmoid layer. Thus, any output probabilities that exceed the preset threshold can be recognized as being present in the current data. In [158] and [159], similar technical route was adopted, that is, signal decomposition + feature engineering + LSTM. Take [159] as an example, PD signal was first decomposed into several components by the LMD algorithm, and only the residual component was retained. Then, it was further split into many small segments, and features were extracted from these segments to form the final feature matrix, which was accepted as the input of the LSTM network. With proper parameters setting, a 99.25% overall accuracy rate can be achieved. Last, a novel LSTM network based on the self-attention mechanism was reported in [160]. By introducing the self-attention block as shown in Fig. 28, the proposed DL model can learn the important relevance of inputs very well, thus improve the diagnostic accuracy.

### E. GAN-BASED METHODS

In DL applications, one harsh truth is that the small data, unbalanced data, low-quality data may all greatly deteriorate the accuracy of results. Thus, many GM technologies have been applied to expand the number of samples or enhance their qualities. The DBN and DAN are two typical GM examples. However, they have bottlenecks in the quality of complex sample generation.

In 2014, Goodfellow et al. proposed an epoch-making GM method GAN [188]. Inspired by the idea of two-person zero-sum game in the game theory, two neural network blocks namely the generator $G$ and discriminator $D$ are designed in GAN. The purpose of $G$ is to generate approximate samples to fool $D$, while the purpose of $D$ is to distinguish between the real samples and generated samples. Then, by adversarial training, the GAN can gradually improve the generating ability of $G$ and the discriminating ability of $D$. Finally, a Nash equilibrium can be achieved. We illustrate the structure of the GAN in Fig. 29.

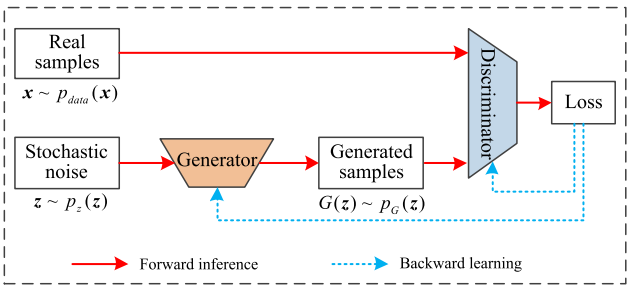



**FIGURE 29.** Illustration of the structure of GAN.

According to [188], the objective function of GAN is:

$$\begin{aligned}
\min_G \max_D V(D, G) = {} & E_{\boldsymbol{x} \sim p_{data}(\boldsymbol{x})}[\log D(\boldsymbol{x})] \\
& + E_{\boldsymbol{z} \sim p_z(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))]
\end{aligned} \tag{24}$$

where $D(\boldsymbol{x})$ denotes the probability of classifying $\boldsymbol{x}$ as real sample. Then, for the discriminator $D$, its training objective should be maximizing $D(\boldsymbol{x})$ while minimizing $D(G(\boldsymbol{z}))$. For the generator $G$, it should maximize $D(G(\boldsymbol{z}))$. After sufficient training iterations, the distribution of the generated samples $p_G(\boldsymbol{z})$ will approach the distribution of real samples $p_{data}(\boldsymbol{x})$. Recently, many DL models based on GAN are developed and applied to power system applications, such as the CGAN, DCGAN, WGAN, WGAN-GP, ACGAN [189].

In [161], the authors introduced the DCGAN to artificially generating PD data to complement the unbalanced sources. Compared with GAN, the main improvement of DCGAN is that CNN is used as the basic structure of the generator and discriminator to improve the quality of generated samples. To generate PD samples with specific type, the CGAN was introduced in [162]. In CGAN, the label information was supplemented as part of the inputs of the discriminator and generator, thus class-discriminative samples can be generated. Considering that each variant of GAN has its own pros and cons, it's a reasonable improvement idea to combine multiple GAN variants. For example, in [163], Fu et al. proposed a novel data augmentation method by taking advantages of the WGAN-GP and CGAN. First, the WGAN-GP adopted the Wasserstein distance and an additional gradient penalty to improve the convergence of training. Second, the CGAN could help to control the model to generate PD samples that had specific labels. Therefore, the final objective functions for the discriminator and generator can be formulated as:

$$L_D = \underbrace{E_{z\sim p_z(z)}[D(G(\boldsymbol{z}|\boldsymbol{c}))] - E_{\boldsymbol{x}\sim p_{data}(\boldsymbol{x})}[D(\boldsymbol{x}|\boldsymbol{c})]}_{Original\ Wasserstein\ loss} + \underbrace{\lambda E_{\hat{\boldsymbol{x}}\sim p(\hat{\boldsymbol{x}})}\left[||\nabla_{\hat{\boldsymbol{x}}} D(\hat{\boldsymbol{x}}|c)||_2 - 1\right]^2}_{Gradient\ penalty\ term} \tag{25}$$

$$L_G = -E_{z\sim p_z(z)}[D(G(\boldsymbol{z}|\boldsymbol{c}))] \tag{26}$$

where $\hat{\boldsymbol{x}} = \varepsilon \boldsymbol{x} + (1-\varepsilon)G(\boldsymbol{z})$, $\varepsilon \in U(0,1)$, $D(\boldsymbol{x}|c)$ denotes the output probability of discriminator for sample $\boldsymbol{x}$ with label $c$, $\lambda$ is the coefficient of the regularization term. After expanding the PD dataset by this method, the recognition rate was improved by 2.2% on average.

Another PD data enhancement scheme based on the AC-BEGAN was also reported in [164], the proposed model can stably generate high-quality samples for each PD type. Lately, Wang et al. explored other GAN-based data augmentation approaches to solve the small or unbalanced sample problem in practical PD diagnosis [165], [166], as outlined in Table 5. Readers can refer to the original papers for more details.

### F. GCN-BASED METHODS

By far, all our mentioned PD pattern recognition methods are designed for the so-called Euclidean structure data, such as 1-D PD waveform, 2-D or 3-D PRPD spectrogram, 2-D or 3-D time-frequency spectrum, etc. A common problem with these methods is that they only utilize the numerical information in regular space while ignore the topological and geometrical information. To cope with this disadvantage, the Euclidean structure data can be converted into the graph structured data first, and then the GSP technology can be applied to explore the correlation characteristics.

As a typical non-Euclidean structure data, the graph data consists of nodes and edges, which can be expressed as [190]:

$$\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E}, \boldsymbol{X}, \boldsymbol{A}), \quad \boldsymbol{V} = (v_1, v_1, \ldots, v_n), \quad \boldsymbol{e}_{ij} = (v_i, v_j) \in \boldsymbol{E} \tag{27}$$

where $\boldsymbol{V}$ and $\boldsymbol{E}$ respectively denote the nodes and edges of the graph, $\boldsymbol{X} \in R^{n\times d}$ is the node feature matrix and $\boldsymbol{A} \in R^{n\times n}$ is the adjacency matrix, $d$ is the feature dimension, $n$ is the number of nodes. Particularly, elements of adjacency matrix signify the presence of an edge between nodes $v_i$ and $v_j$:

$$A_{i,j} = \begin{cases} 1, & (v_i, v_j) \in \boldsymbol{E}\ and\ i \neq j \\ 0, & otherwise \end{cases} \tag{28}$$

Another widely used representation of graph data is the Laplacian matrix, which is calculated by:

$$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A} \tag{29}$$

where $\boldsymbol{L} \in R^{n\times n}$ is the Laplacian matrix, $\boldsymbol{D} = diag(d_1, \ldots, d_n)$ is the degree matrix, $d_i$ represents the number of edges with the $i$-th node as the endpoint. We give a graphical description of the calculation of Laplacian matrix in Fig. 30.

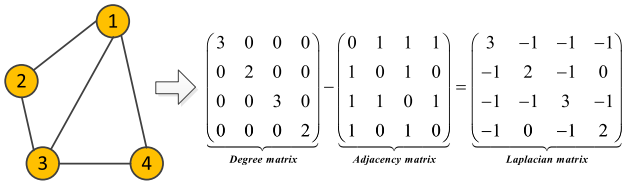



**FIGURE 30.** Illustration of a graph data and its Laplacian matrix.

Among many GSP technologies, the GCN has gained wide attention due to its excellent ability of feature extraction and end-to-end learning by using small data. According to [191], the graph convolution operation is formulated as:

$$\boldsymbol{Y} = \mathrm{ReLU}((\tilde{\boldsymbol{D}}^{-1/2}\tilde{A}\tilde{\boldsymbol{D}}^{-1/2})\boldsymbol{X}\boldsymbol{W}) \quad \tilde{\boldsymbol{D}}_{i,i} = \sum\nolimits_j \tilde{\boldsymbol{A}}_{i,j}, \quad \tilde{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{I}_n \tag{30}$$

where $\boldsymbol{Y} \in R^{n\times o}$ is the output, $o$ is the dimension of output feature. In addition, $\boldsymbol{W} \in R^{d\times o}$ denotes the learnable weight coefficient matrix, $\boldsymbol{I}_n$ is the identity matrix of order $n$.

In [167], Zhang et al. introduced the GSP technique to improve the PD diagnostic accuracy. First, the grayscale TF spectrums were transformed into the graph data by window-based scanning technology, as shown in Fig. 31. In this way, the graph data contained not only the numerical information of the PD spectrums, but also the relationships between the local regions. Furthermore, the authors constructed a GCN model by stacking graph convolution layers and SAGPool layers, which showed high accuracy rate with small data.

The authors of [168] and [169] combined GCN with other techniques to further improve the performance. For example, in [168], the GIS PD data were first converted

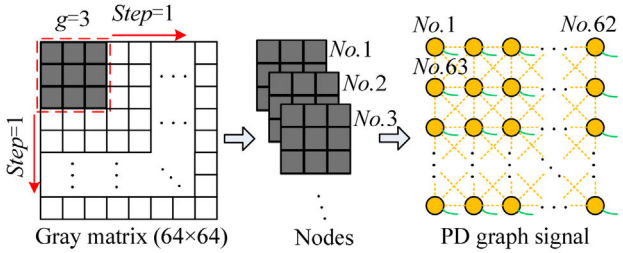



**FIGURE 31.** Graphical explanation of converting grayscale PD images into PD graph data. Figure adopted from [167].

into the graph data by convolution operation and k-nearest neighbors. Then a GCN network with ChannelSortPooling layer was designed for the graph representation of PD data. Finally, a capsule network was added to enhance the feature extraction. This combined scheme can achieve high precision without data generation, indicating its fantastic application prospect.

Another interesting research work was reported in [170]. In this article, a combined data-driven and knowledge-driven method was adopted. Specifically, the information extraction was conducted in both the feature space $\mathcal{F}$ and knowledge space $\mathcal{K}$. In feature space, a ResNet was employed to form the feature vector $F$ from the PRPS patterns. In knowledge space, a PD-tailored KG was established firstly based on the expert knowledge, as shown in Fig. 32. Then, the descriptive phrases in KG were transformed into semantic embeddings by a BERT model, and the semantic embeddings of all nodes were combined to form the embedding matrix $E$. A cluster-GCN was then adopted to learn the knowledge representation of each PD pattern from $E$, and output classification matrix $C$. At last, the extracted information from feature space (i.e., $F$) and knowledge space (i.e., $C$) was fused to obtain the final result. Because the expert knowledge is authoritative and explainable, the proposed scheme can be more reliable and robust compared to a separate data-driven method.

### G. DEL-BASED METHODS

As we discussed in Section II, the EL-based models can often achieve better performance than any individual models. In this subsection, the DEL methods and their applications in PD diagnosis are investigated.

The key of DEL method is to introduce the DL algorithms to traditional EL-based method, thus more powerful ability on feature extraction and classification can be expected [192].

In [171], a deep ensemble model for detection of unknown PRPD patterns was developed. The authors used multiple parallel CNN models, with each model outputting the class prediction probability of input sample. Then, an averaging ensemble strategy was adopted to get the ensemble output. The authors believed that if the maximum value of ensemble output was still less than a threshold, then much uncertainty may exist in the recognition result, thus the current sample should be judged as unknown category. Benefited from the ensemble scheme, the proposed method can detect unknown PRPD faults with almost 100% accuracy.

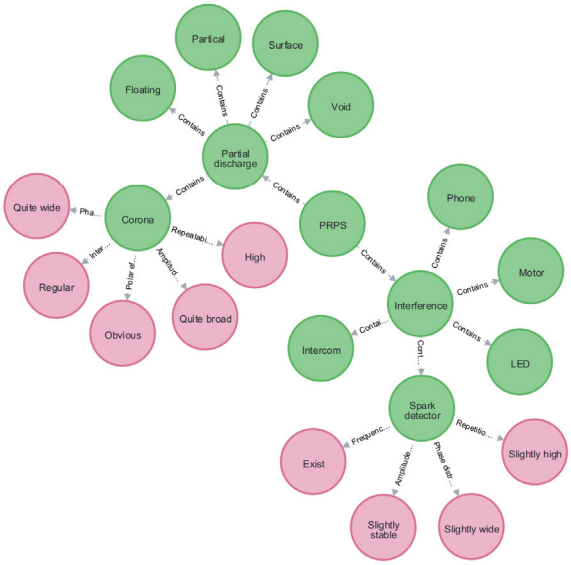



**FIGURE 32.** A PD-tailored knowledge graph established in [170].

Yeo et al. presented another DEL approach for medium voltage power cables PD identification [172]. This approach consisted of two major parts: a) cable-specific adaptation; b) neural network ensemble. The first task was achieved by re-training the classifier layer of the pre-trained DL models to better adapt to the MV cable application. The second task was accomplished by two ensemble schemes. Consequently, a high accuracy of 98.65% can be obtained by the ensemble NNs based on the CNN-BILSTM combination.

In both [173] and [174], the technical route based on DL model and stacking EL was adopted. For example, two 1D-CNN models and two autoencoder models were used as the base-learners, while a deep neural network was used as the meta-learner in [174]. To enhance the diversity between base classifiers, the authors divided the raw PD dataset into two groups, one had subtracted a mean value, while the other had not. After pre-processing steps like denoising, the two groups were respectively used as the input of the 1D-CNN model and the autoencoder model. The effectiveness of this method was validated by PD datasets collected from on-site overhead power lines.

### H. OTHER DL METHODS

Except the above representative methods, there are still some other DL-based schemes that have been applied to PD type recognition, most of which are the combination of two or more different DL algorithms. Considering that the basic DL models used in these methods have been reviewed in former sections, we only summarize them in Table 5.

Although we have tried our best to give a comprehensive survey about the applications of DL methods in PD diagnosis,

some literatures may still be omitted. Actually, that is why it is important to keep track of this hot research area.

## IV. SEVERITY ASSESSMENT

Another important application of pattern recognition methods in PD field is the severity assessment. A large number of theoretical and experimental studies have demonstrated that PD signals under different insulation degradation levels show discriminative characteristics. Therefore, advanced signal processing and ML technologies can be applied. In this section, we will focus on this topic.

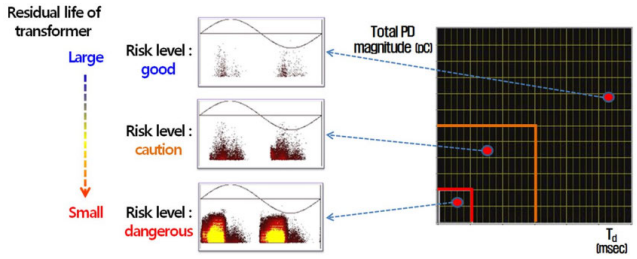



**FIGURE 33. A typical example of PD severity assessment for power transformer. Figure adopted from [193].**

First, we would like to briefly introduce the concept of PD severity assessment through Fig. 33 that adopted from [193]. As can be seen, with risk level became worse, the discharge intensity at rising edge of AC voltage increased significantly. At the same time, distribution of the PD features (total PD magnitude and discharge interval in this example) extracted at different severity stages also had obvious distinguishability. Thus, these discriminative features can be well identified by classification algorithms, which also meant that the PD level can be evaluated. Furthermore, people can even predict the remaining useful life of related equipment if the variation rules of PD features can be modeled accurately.

In essence, the PD severity assessment also belongs to the category of pattern recognition, thus relevant approaches can also be classified into two categories: traditional ML-based methods and DL-based methods. Considering that the mainstream algorithms regarding these two categories have been discussed in detail in Section II and Section III, we will no longer introduce the details of each cited literature in this section. Instead, we provide concise summaries about the two topics in Table 7 and Table 8, respectively. And we encourage the readers to refer to the original literature if they are interested in a specific method.

## V. DISCUSSION

### A. APPLICATION EFFECTS DISCUSSION

#### 1) TRADITIONAL ML-BASED METHODS FOR PD TYPE RECOGNITION

A common technical route of a traditional ML method often includes steps such as feature extraction, dimension reduction, and classification. In Section II, these topics are discussed in detail. Based on the results reported in the literatures and our own experiences, following perspectives can be summarized:

a) For traditional ML methods, feature extraction is widely considered to be the most significant step. In PD recognition, statistical or image-based features from the PRPD pattern are effective in most cases. However, if the power frequency voltage phase is unavailable (e.g. under DC voltage), features from TRPD pattern or TF spectrum should be given more attention. Furthermore, if possible, using multi-source feature information will be more advantageous than using single-source feature information only.

b) In practice, not all the extracted feature information is helpful for PD type recognition. On the contrary, too much irrelevant information not only interferes with the recognition tasks, but also greatly increases computing and storage costs. A proven and effective way to cope with this problem is to introduce the dimension reduction techniques. Both feature transformation algorithms and feature selection algorithms perform well in dimension reduction of PD signals. Despite all this, if an application requires retaining the interpretability and structure of the original data after dimension reduction, then the feature selection algorithm may be a better choice.

c) Classification is the last step of a traditional ML method to achieve the pattern recognition. According to the results of literatures listed in Table 4, one can observe that supervised learning algorithms such as ANN and SVM often achieve superior performances than unsupervised ones. However, if accurate class labels are difficult to obtain, semi-supervised or unsupervised learning algorithms will be more desirable. Considering the uncertainty of PD phenomenon itself and the signal propagation process, the fuzzy-based methods are also favorable in PD field. In particular, we believe that the FLI system will be a promising means for PD identification, because its inference rules often have clear physical meaning. Meanwhile, it does not require any training steps. Last, it can be observed from Table 4 that the average accuracy of ensemble algorithms is the highest among all methods. We have detailed the EL-based methods and their applications in PD recognition in Section II. And undoubtedly, we should pay closer attention to this rapidly developing technology.

In summary, traditional ML methods show advantages of low computational complexity, clear theoretical foundation, and good interpretability. However, with the increase of the data volume and diversity of data types (e.g., the unstructured data), traditional ML methods exhibit certain shortcomings in both recognition accuracy and generalization ability.

#### 2) DL-BASED METHODS FOR PD TYPE RECOGNITION

Unlike the traditional ML methods that rely on the artificially designed features, the DL method can automatically extract the feature information from the raw data. Moreover, the learning ability of DL models increases exponentially with the network depth. Therefore, various DL algorithms have been introduced to the PD diagnosis field in recent years.

**TABLE 7. Summary of the traditional ML-based methods for severity assessment of PD signals.**

| Reference | Year | Target device | Data type | Extracted PD signatures | Adopted ML algorithms | Main conclusions | Reported accuracy |
|---|---|---|---|---|---|---|---|
| [194] | 2014 | GIS | UHF PD data | Signal power | Not mentioned | a) Correlation between apparent charge and power of floating metal is logarithmic;<br>b) Small power of PD signal caused by epoxy resin void is correspond to large apparent charge. | Not mentioned |
| [195] | 2011 | GIS | IEC 60270 PD data, UHF PD data, acoustic PD data, ultraviolet image | Phase distribution characteristics and spectrum features of PD data | Not mentioned | a) As PD continues, indicators of PD data such as phase distribution and spectrum shape are constantly changing;<br>b) PD initiated by HV electrode protrusion in GIS can be divided into three levels, namely the petty discharge, medium discharge, and threatening discharge. | Not mentioned |
| [196] | 2015 | GIS | UHF PD data | Four statistical features, three-time interval of adjacent discharge features, equivalent cumulative discharge quantity, signal entropy | Two-level fuzzy comprehensive evaluation model | Four severity degrees of PD were defined: normality, attention, abnormity, and warning. | Average accuracy of 91.75% |
| [197] | 2017 | GIS | UHF PD data | A 16-dimension feature set extracted from PRPD pattern | SVM | The MRMR algorithm can help to choose the best valuable features to evaluate the PD severity. | Average accuracy of 89.5% |
| [198] | 2019 | GIS | UHF PD data | PD phase width, mean of maximum amplitude distribution, correlation coefficients, box dimension, information dimension | K-means clustering algorithm | a) The designed features can be used to represent the PD characteristics at different stages;<br>b) The PD stages of free metal particles in GIS can be divided as the development stage and near breakdown stage. | Not mentioned |
| [199] | 2019 | Oil-paper insulation in power transformer | IEC 60270 PD data | Cumulative PD energy | Not mentioned | a) Discharge energy can be used to quantify the discharge intensity;<br>b) Three degradation stages were divided: initiation, development, and danger. | Not mentioned |
| [200] | 2017 | Oil-pressboard insulation in converter transformer | IEC 60270 PD data | 24 wavelet moment invariants extracted separately from the PRPD pattern and equivalent time-frequency pattern | LS-SVM with PSO | a) According to the time-variation trends of several indicators, the PD process can be divided into five stages.<br>b) By using features from the PRPD pattern and ETF pattern, the LS-SVM classifier can be used to achieve the PD stage recognition. | 93.75% |
| [193] | 2019 | Power transformer | UHF PD data | Summation of the total PD magnitude during four cycles + discharge intervals | Not mentioned | The summation of total PD magnitude during four cycles and discharge intervals can well reflect deterioration of cast-resin transformer. | Not mentioned |
| [201] | 2021 | Oil-pressboard insulation in power transformer | IEC 60270 PD data | Average apparent discharge magnitude, maximum apparent discharge magnitude, discharge repetition rate, discharge power | Not mentioned | a) When the applied voltage $U$ exceeds $U_{th}$ and the maximum field strength of needle-bar electrode $E_{\max}$ exceeds $E_{th}$ , a new PD phenomenon called FDD can be defined;<br>b) According to the characteristics of FDD, five discharge stages can be defined. | Not mentioned |
| [202] | 2022 | Oil-paper insulation in power transformer | PD data based on HFCT, SiPM, ultrasonic sensor and UHF sensor | Multiple features including the pulse repetition rate, average discharge per cycle, statistical operators of PRPD pattern, average light radiation intensity, etc. | Not mentioned | a) By analyzing PD data characteristics from multiple detection devices, comprehensive connection between the discharge stage and measured PD data can be established;<br>b) According to the proposed method, PD stage of needle-plate electrode can be divided into the development stage, stagnation stage, and pre-breakdown stage. | Not mentioned |
| [203] | 2019 | Oil-pressboard insulation in power transformer | IEC 60270 PD data | Maximum discharge magnitude, average discharge magnitude, pulse repetition rate | Not mentioned | a) Based on surface morphology observed by scanning electron microscopy, the electrical aging process can be divided into the initial stage, developing stage, and erupting stage;<br>b) The average discharge magnitude and pulse repetition rate will decrease in the latter half of the developing stage, and they will decrease to a minimum before entering the erupting stage. | Not mentioned |
| [204] | 2022 | Oil-paper insulation in power | Microscopic images at different | 11 texture features based on the gray-level run-length matrix of | SVM | a) As degradation of oil-paper insulation, the textures of its surface images changes;<br>b) The multiple linear regression algorithm | Accuracy is higher than 90% |

**TABLE 7.** *(Continued.)* **Summary of the traditional ML-based methods for severity assessment of PD signals.**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | transformer | aging stages | microscopic images | | can be used to fit the relationship between texture features and PD severity degree. | |
| [205] | 2022 | Furnace transformer | IEC 60270 PD data | Apparent discharge, PD intensity | Fuzzy logic algorithm | a) A generalized indicator of the transformer insulation condition was defined for the first time based on the fuzzy rules;<br>b) The proposed method can achieve reliable state assessment of transformer insulation. | Not mentioned |
| [206] | 2018 | Gas insulated line | IEC 60270 PD data | Skewness and kurtosis values of the $n-Q$ spectrum and $n-\Delta t$ spectrum | Not mentioned | Statistical parameters can be used to describe the PD severity caused by metal particles on DC GIL insulator surface. | Not mentioned |
| [207] | 2017 | DC gas-insulated equipment | Five $SF_6$ decomposed components | Concentrations ratios of the $SF_6$ decomposed components | BPNN + SVM | Ratios such as $C(CO_2)/CT_1$, $C(CF_4)/C(SO_2)$, $C(CO_2)/C(SOF_2)$ and $C(CF_4)/C(CO_2)$ can be used to diagnose the PD Severity. | Up to 97.75% |
| [208] | 2019 | Solid insulated switchgear | IEC 60270 PD data | Nine statistical feature parameters | BPNN | Technical route based on statistical features and BPNN performs well in PD severity assessment of the solid insulated switchgear. | Not mentioned |
| [209] | 2021 | Ceramic insulator | Leakage current data + ultraviolet pulse data | Five GLCM features and three Tamura features extracted from the UV spectrograms | SVM | Three discharge severity levels were divided: corona discharge, local arc discharge, long arc discharge. | Overall accuracy of 90.6% |
| [210] | 2020 | Cables under HVDC | IEC 60270 PD data | Mean and standard deviation of five wavelet detail coefficients | SVM | Two ageing states namely the "virgin" and "aged" can be identified. | Up to 100% |
| [211] | 2017 | XLPE cables under HVDC | HFCT PD data | A feature matrix consists of statistical characteristics and time-domain features | CS theory | PD data from three defect types under two severity degrees can be well identified by the CS-based method. | Reached 95.16% |
| [212] | 2021 | HV polymeric insulation | UHF PD data | The power ratio for low frequencies and the equivalent bandwidth | Not mentioned | The proposed $PRL-\sigma_F$ map can provide robust representation of the PD tree growth under different stages. | Not mentioned |
| [213] | 2022 | HV equipment with polymer insulation | UHF PD data | 28 characteristic parameters including the wavelet coefficients, time-frequency domain parameters, etc. | BPNN | Growth process of electrical tree in polymer insulation can be divided into four stages: initial stage, growth stage, growth retardation stage and rapid growth stage; | Overall accuracy of 99.93% |

In Section III, we discussed different DL methods in detail. The DBN and DAN models are introduced to the PD field in an early stage. But the DBN model still adopts FC network structure and its training process is relatively complex, it is difficult to improve its performance by simply increasing the network depth. As a kind of generative model, the DAN is more frequently used to map the high-dimensional input data into a low-dimensional space, then features from this latent space can be employed for subsequent classification steps.

CNN is probably the most popular DL network currently. It has many advantages, including fewer parameters, flexible architecture, and strong robustness. As shown in Fig. 20, we divided the CNN methods into three categories and discussed them in detail. In general, the CNN-based method uses image data like the PRPD spectra or TF spectra as the input, and it can obtain a satisfactory recognition rate (above 95% in most reported results). Particularly, when training samples are rare or a distribution discrepancy existed between the source and target domains, the CNN-based TL algorithms can be used. Other advanced CNN techniques such as lightweight CNN models, fusion-based CNN models, interpretable CNN models, and automatic CNN models are also very effective and reliable tools to improve the recognition performance under different application scenarios.

Considering that many raw PD signals are saved in the form of 1-D waveform, the RNN models which are good at processing sequential data are also favorable in PD fault diagnosis. One of the most popular RNN models is the LSTM. Authors of [156], [157], [158], [159], and [160] all adopted LSTM models to realize the PD type identification. As shown in Table 5, most reported results are competitive.

Data imbalance is a common problem in many practical DL applications. The GAN is considered as a very effective technique to solve this problem by generating high-quality artificial samples. Many scholars also introduced GAN to the PD field, and the results demonstrated that PD data generated by GAN had high similarity with the real ones. Moreover, by supplementing label information to the inputs of GAN (e.g., CGAN), the model can generate PD data for specified class [162]. By adding extra penalty term to the objective function of GAN (e.g., WGAN-GP), the training process will become faster and more robust [163]. In summary, when the obtained PD data is insufficient or unbalanced, employment of GAN can well supplement the necessary samples, thus improves the overall performance of the adopted algorithm.

**TABLE 8. Summary of the DL-based methods for severity assessment of PD signals.**

| Reference | Year | Target device | Data type | Adopted DL algorithms | Main conclusions | Reported accuracy |
|---|---|---|---|---|---|---|
| [214] | 2019 | GIS | UHF PD data + acoustic PD data + equipment-related data | DL model based on LSTM and ensemble algorithm | a) PD data, technical impact factors, and equipment operation information can be combined to evaluate the severity assessment;<br>b) The proposed Bagging-LSTM model can provide more accurate prediction about the PD severity than traditional ML algorithms like the LSTM, BPNN, etc. | Overall accuracy of 95.8% |
| [215] | 2021 | GIS | Structured UHF PD data + unstructured textual data | VGG16 for PRPS data + global vectors for textual data + Bayesian formula for final failure probability | a) The global vectors model can effectively convert the textual PD data into numerical data for subsequent processing;<br>b) By using comprehensive fault information and combined DL model, the proposed method can provide more accurate and reliable severity evaluation results. | Precision: 41.7%; Recall: 93.8%; F1-Score: 57.7%; |
| [216] | 2017 | GIS | UHF PD data | SSAE | a) The proposed SSAE model can automatically extract features from a latent space, thus is suitable for PD severity assessment;<br>b) Based on the model, four PD stages namely normal state, attention state, serious state, and dangerous state can be identified. | Average accuracy of 92.2% |
| [217] | 2021 | Form-wound windings in electric aircraft propulsion | HFCT PD data | Autoencoder + LSTM-RNN | a) DL model based on autoencoder and LSTM-RNN was designed to derive one failure precursor from three kinds of PD features<br>b) The proposed DL model can make multi-time-step prediction based on the historical failure precursor data;<br>c) The proposed forewarning method can be deployed in the state-of-the-art embedded AI computing processor. | Maximum MAE was 3% |
| [218] | 2021 | Inverter-fed motor | Simulated PDIV data based on the Townsend theory | DBN | The DBN model can well fit the nonlinear relationship between the influence factors (including the varnished wire diameter, insulation layer thickness, relative dielectric constant, temperature) and the PDIV. | Max relative error was 5.9% |
| [219] | 2023 | XLPE cables | IEC 60270 PD data | CNN with transfer learning | a) According to the electrical tree growth, the PD stages of XLPE cables can be divided as the fast-growth phase, the slow-growth phase, and the breakdown phase;<br>b) Both CNN and CNN-based TL methods were adopted to identify the PD stages, and highest accuracy was obtained by using the Densenet201 with Nadam. | Highest accuracy of 99.78% |
| [220] | 2020 | Not mentioned | Visible PD images | A DL model with an encoder and a single-layer perceptron | a) Visible PD images contain abundant information of discharges at different stages;<br>b) With three kinds of statistical features, the proposed DL model can achieve higher accuracy than traditional ML algorithms in PD state recognition. | Up to 94% |

Different from the above DL methods that accept numeric data as the input, GCN methods accept the graph structured data, which consist of the nodes and edges. Thus, the first step of applying GCN is to convert the raw PD data into the PD graph data. Then, these unstructured data can be used to train a GCN model [167], [168], [169]. The key advantage of GCN is that it not only uses the numerical information, but also exploits the structural information between raw data. We also note that the GCN technique can be well combined with the PD-tailored KG shown in Fig. 32. By integrating the feature space information and the knowledge space information, the results show higher accuracy and better interpretability [170].

Last, the DEL algorithms and their applications in PD diagnosis are also investigated in Section III. As can be seen from Table 5, the combination of DL and EL techniques can result in very competitive results.

Although the DL-based PD type recognition can achieve very high accuracy, it still has the following disadvantages: a) it requires huge amount of high-quality training data and expensive computing and storage facilities; b) it is a slow and arduous process of trial-and-error testing to obtain a satisfied DL model; c) the trained DL model often shows significant performance degradation when applied to different scenarios such as new working conditions, or new equipment, etc.

### 3) METHODS FOR PD SEVERITY ASSESSMENT

In Section IV, we surveyed the papers related to PD severity assessment. Similarly, we divide the relevant approaches into the traditional ML-based methods and DL-based methods. Considering most concrete algorithms adopted in PD severity assessment have been discussed in previous sections, only some representative works were summarized in Table 7 and Table 8 to avoid excessive article length.

In essence, the key to identify different PD stages depends on the determination of feature parameters that can reflect the PD physical characteristics at various severity levels. And we should also note that for different insulation materials, their degradation mechanisms differ. For these reasons, it is not easy to determine feature indicators with broad applicability in PD severity identification. However, from the literatures listed in above two tables, we can also conclude that those statistical features extracted from PRPD or TF spectrograms perform well in this task. And, of course, if multiple data types or pattern recognition methods can be integrated (as

authors of [214] and [215] did), the diagnostic results can be more reliable.

### B. FUTURE PERSPECTIVES

As the increasing applications of advanced signal processing and machine learning technologies in UHF PD diagnosis, several important research issues remain to be addressed in the future. We conclude six open issues in this subsection.

#### 1) PD MECHANISMS UNDER NEW APPLICATION SCENARIOS

Research on the PD mechanism has always been the basis of PD detection and diagnosis. Therefore, the PD mechanism of new insulation materials under new types of voltages is an important scientific problem that needs to be studied urgently.

Here, we believe that the following aspects are particularly important: a) PD formation mechanism research in the new type of insulation materials (gas-type, liquid-type, solid-type, or their mixture) [221]; b) PD behaviors research under new types of voltages such as high-frequency pulse voltage [222], superposed inter-harmonic AC voltage [223], square pulses with ultra-fast $dv/dt$ [224], etc.; c) excitation mechanism and propagation characteristics study of PD electromagnetic waves [225], [226].

#### 2) NEW-TYPE UHF SENSORS

Note that for either online monitoring or offline monitoring, the performance of UHF PD sensor directly determines the accuracy and reliability of the diagnosis results. In particular, considering that more and more key power equipment are required to install the PD online monitoring systems before their operation, the design of new-type UHF sensor with high sensitivity and installation flexibility is also important.

Therefore, we point out three development directions of UHF PD sensors: a) optimize the antenna design to achieve the characteristics of small size, high gain, and broadband [227]; b) design of flexible or self-sensing UHF sensors to make the sensors conformal with power equipment body [228], [229]; c) design of integrated sensors that can perceive multiple kind of PD excited signals [230].

#### 3) ESTABLISHMENT OF COMMON DATABASE

As everyone knows, advanced ML algorithms (especially the DL algorithms) are data-hungry. In most existing studies, the authors used datasets collected purposely for their own research. However, we believe that high-quality open access datasets are more favorable for the PD field. This can not only attract more researchers to develop advanced diagnostic algorithms, but also provide benchmark for comparing the performances of different approaches.

Unfortunately, unlike many well-known open access datasets in the field of computer vision like the ImageNet, CIFAR-10 & CIFAR-100, MNIST, etc., very few public datasets are available in the PD field. To the best of our knowledge, there are only two relevant PD datasets can be found, as listed in Table 9.

**TABLE 9. Open access PD datasets.**

| Authors | Dataset name | Available website |
|---|---|---|
| Armando Rodrigo Mor *et al.* | PDflex Dataset | http://pdflex.ewi.tudelft.nl/examples/ |
| Addison Howard *et al.* | VSB ENET Dataset | https://kaggle.com/competitions/vsb-power-line-fault-detection |

#### 4) PATTERN RECOGNITION METHODS UNDER SMALL /UNBALANCED SAMPLES OR DIFFERENT WORKING CONDITIONS

The problems of small samples, unbalanced samples, varying working conditions are very common in the PD field.

First, insufficient training data may lead to the overfitting problem, resulting in poor generalization performance of the trained ML model. From the data perspective, the training samples can be supplemented by using data augmentation (e.g., performing translation, rotation, scaling, cropping operations on data) or data generation (e.g., by DAN or GAN) technologies. From the method perspective, transfer learning, zero-shot learning, one-shot learning, and few-shot learning algorithms are all worth considering.

Second, the unbalanced samples can result in the trained model being more inclined to classify the new data into categories with sufficient samples. One possible solution is to supplement the minority classes by using data augmentation or generation methods as described in previous paragraph. Another scheme is to adjust the weights of different class samples in the loss function, so that the model can focus more on the minority. Last, the EL methods can also be adopted, in which each sub-model focuses on a specific class.

Third, the variation of working conditions often causes the distribution discrepancy between training dataset and testing dataset. Therefore, the transfer learning methods discussed in Section III can be employed. In addition, the incremental learning [231] and lifelong learning methods are also potential solutions to this issue.

#### 5) COMBINED DATA-DRIVEN AND KNOWLEDGE-DRIVEN ML METHODS

The existing PD fault diagnosis methods can be generally divided into two categories: knowledge-driven methods and data-driven methods. The knowledge-driven methods are often built based on mechanism models or rule-based models. By contrast, in data-driven methods, people usually establish empirical models directly from the data samples.

To fully utilize the advantages of both schemes, combined data-driven and knowledge-driven method is highly expected. This novel ML paradigm has received increasing attention nowadays, and some impressive works about its application in PD diagnosis have been reported [122], [170]. Once again, we believe this research paradigm will become an important development direction.

#### 6) MULTI-SOURCE FUSION METHODS

As we mentioned above, considering the complexity of PD phenomenon itself, the diversity of power equipment,

and the uncertainty of surrounding environment, it almost unrealistic for a single detection method or diagnostic method to obtain reliable evaluation results on PD faults. Therefore, the multi-source fusion methods should be paid more attention.

Up to now, many scholars have introduced multi-source fusion techniques to the PD diagnosis field, and most of them focused on the feature-level fusion [150], [151] and decision-level fusion [68], [232]. Nevertheless, the following aspects still need to be considered: a) the data-level fusion scheme; b) the fusion scheme that cross different levels; c) the weighted fusion scheme in which sub-parts of different importance are given different weights; and d) fusion scheme that includes the expert knowledge.

## VI. CONCLUSION

As the second of our two-part reviews, pattern recognition problems of UHF PD signals are discussed in this paper. The main contents can be concluded as the following aspects.

First, we focus on the traditional ML-based PD type recognition issue. According to its common technical route, the feature extraction methods, dimension reduction methods, and classification methods are discussed in detail, and their representative applications in PD recognition are reviewed. In particular, we highlight some latest progress related to this topic, such as the feature transformation algorithms, feature selection algorithms, EL algorithms, etc.

Second, the DL-based methods and their applications in PD recognition are comprehensively investigated, including the DBN, DAN, CNN, RNN, GAN, GCN, DEL, etc. For each DL category, its basic ideas, typical models, recent research progresses, advantages and limitations are discussed in detail. Considering the high flexibility and excellent performance of CNN, we spend many pages to introduce the various CNN-based methods shown in Fig. 20 to help readers understand their values in PD diagnosis.

Third, approaches related to PD severity assessment are also surveyed. Since most methods used in this topic are similar with those in PD type recognition, we only list some representative works in Table 7 and Table 8.

Last, we thoroughly discuss the application effects of relevant algorithms in PD type recognition and severity assessment. And six significant open issues that remain to be addressed in the future are prospected.

## REFERENCES

[1] A. Subramaniam, A. Sahoo, S. S. Manohar, S. J. Raman, and S. K. Panda, "Switchgear condition assessment and lifecycle management: Standards, failure statistics, condition assessment, partial discharge analysis, maintenance approaches, and future trends," *IEEE Elect. Insul. Mag.*, vol. 37, no. 3, pp. 27–41, May 2021.

[2] Q. Khan, S. S. Refaat, H. Abu-Rub, and H. A. Toliyat, "Partial discharge detection and diagnosis in gas insulated switchgear: State of the art," *IEEE Elect. Insul. Mag.*, vol. 35, no. 4, pp. 16–33, Jul. 2019.

[3] M. Wu, H. Cao, J. Cao, H.-L. Nguyen, J. B. Gomes, and S. P. Krishnaswamy, "An overview of state-of-the-art partial discharge analysis techniques for condition monitoring," *IEEE Elect. Insul. Mag.*, vol. 31, no. 6, pp. 22–35, Nov. 2015.

[4] T. Mazhar, H. M. Irfan, I. Haq, I. Ullah, M. Ashraf, T. A. Shloul, Y. Y. Ghadi, and D. H. Elkamchouchi, "Analysis of challenges and solutions of IoT in smart grids using AI and machine learning techniques: A review," *Electronics*, vol. 12, no. 1, p. 242, Jan. 2023.

[5] N. C. Sahoo, M. M. A. Salama, and R. Bartnikas, "Trends in partial discharge pattern classification: A survey," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 12, no. 2, pp. 248–264, Apr. 2005.

[6] G. V. R. Xavier, H. S. Silva, E. G. da Costa, A. J. R. Serres, N. B. Carvalho, and A. S. R. Oliveira, "Detection, classification and location of sources of partial discharges using the radiometric method: Trends, challenges and open issues," *IEEE Access*, vol. 9, pp. 110787–110810, 2021.

[7] Z. G. Tang, M. Z. Tang, J. Z. Li, J. Y. Wang, C. Wu, and K. Wang, "Review on partial discharge pattern recognition of electrical equipment," *High Voltage Eng.*, vol. 43, no. 7, pp. 2263–2277, 2017.

[8] A. Mas'ud, R. Albarracín, J. Ardila-Rey, F. Muhammad-Sukki, H. Illias, N. Bani, and A. Munir, "Artificial neural network application for partial discharge recognition: Survey and future directions," *Energies*, vol. 9, no. 8, p. 574, Jul. 2016.

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

[10] S. Barrios, D. Buldain, M. P. Comech, I. Gilbert, and I. Orue, "Partial discharge classification using deep learning methods—Survey of recent progress," *Energies*, vol. 12, no. 13, p. 2485, Jun. 2019.

[11] S. Lu, H. Chai, A. Sahoo, and B. T. Phung, "Condition monitoring based on partial discharge diagnostics (don't short) using machine learning methods: A comprehensive state-of-the-art review," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 27, no. 6, pp. 1861–1888, Dec. 2020.

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

[13] 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.

[14] 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.

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

[16] 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.

[17] S. Song, Y. Qian, H. Wang, Y. Zang, G. Sheng, and X. Jiang, "Partial discharge pattern recognition based on 3D graphs of phase resolved pulse sequence," *Energies*, vol. 13, no. 16, p. 4103, Aug. 2020.

[18] 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 Del.*, vol. 34, no. 2, pp. 542–550, Apr. 2019.

[19] S. Sun, Y. Sun, G. Xu, L. Zhang, Y. Hu, and P. Liu, "Partial discharge pattern recognition of transformers based on the gray-level co-occurrence matrix of optimal parameters," *IEEE Access*, vol. 9, pp. 102422–102432, 2021.

[20] Z. Li, H. Wang, Y. Qian, R. Huang, and Q. Cui, "Pattern recognition of partial discharge in the presence of noise based on SURF," *Trans. China Electrotech. Soc.*, vol. 37, no. 3, pp. 775–785, 2022.

[21] J. Li, Q. Zhang, K. Wang, and R. Liao, "Partial discharge recognition reliability considering the influence of multi-factors based on the two-directional fuzzy-weighted two-dimensional principal component analysis algorithm," *Electr. Power Compon. Syst.*, vol. 44, no. 4, pp. 459–470, Feb. 2016.

[22] M. Rostaghi-Chalaki, K. Yousefpour, J. Klüss, M. Kurum, J. P. Donohoe, and C. Park, "Classification and comparison of AC and DC partial discharges by pulse waveform analysis," *Int. J. Electr. Power Energy Syst.*, vol. 125, Feb. 2021, Art. no. 106518.

[23] M.-S. Su, C.-C. Chia, C.-Y. Chen, and J.-F. Chen, "Classification of partial discharge events in GILBS using probabilistic neural networks and the fuzzy c-means clustering approach," *Int. J. Electr. Power Energy Syst.*, vol. 61, pp. 173–179, Oct. 2014.

[24] X. Peng, J. Wen, Z. Li, G. Yang, C. Zhou, A. Reid, D. M. Hepburn, M. D. Judd, and W. H. Siew, "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.

[25] H. Janani, S. Shahabi, and B. Kordi, "Separation and classification of concurrent partial discharge signals using statistical-based feature analysis," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 27, no. 6, pp. 1933–1941, Dec. 2020.

[26] J. Granado, C. Álvarez-Arroyo, A. Torralba, J. A. Rosendo-Macías, J. Chávez, and M. Burgos-Payán, "Time domain analysis of partial discharges envelope in medium voltage XLPE cables," *Electr. Power Syst. Res.*, vol. 125, pp. 220–227, Aug. 2015.

[27] J. Jiang, J. Chen, J. Li, X. Yang, Y. Bie, P. Ranjan, C. Zhang, and H. Schwarz, "Partial discharge detection and diagnosis of transformer bushing based on UHF method," *IEEE Sensors J.*, vol. 21, no. 15, pp. 16798–16806, Aug. 2021.

[28] M. Zhu, J. Xue, J. Zhang, Y. Li, J. Deng, H. Mu, G. Zhang, X. Shao, and X. Liu, "Classification and separation of partial discharge ultra-high-frequency signals in a 252 kV gas insulated substation by using cumulative energy technique," *IET Sci., Meas. Technol.*, vol. 10, no. 4, pp. 316–326, Jul. 2016.

[29] H. Janani, P. Jayasinghe, M. J. Jozani, and B. Kordi, "Statistical feature extraction and system identification algorithms for partial discharge signal classification using Laguerre polynomial expansion," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 27, no. 6, pp. 1924–1932, Dec. 2020.

[30] D. Evagorou, P. L. Lewin, V. Efthymiou, A. Kyprianou, G. E. Georghiou, A. Stavrou, and A. C. Metaxas, "Feature extraction of partial discharge signals using the wavelet packet transform and classification with a probabilistic neural network," *IET Sci., Meas. Technol.*, vol. 4, no. 3, pp. 177–192, May 2010.

[31] H. H. Sinaga, B. T. Phung, and T. R. Blackburn, "Recognition of single and multiple partial discharge sources in transformers based on ultra-high frequency signals," *IET Gener., Transmiss. Distrib.*, vol. 8, no. 1, pp. 160–169, Jan. 2014.

[32] N. Morette, L. C. Castro Heredia, T. Ditchi, A. R. Mor, and Y. Oussar, "Partial discharges and noise classification under HVDC using unsupervised and semi-supervised learning," *Int. J. Electr. Power Energy Syst.*, vol. 121, Oct. 2020, Art. no. 106129.

[33] D. K. Mishra, S. Dhara, C. Koley, N. K. Roy, and S. Chakravorti, "Self-organizing feature map based unsupervised technique for detection of partial discharge sources inside electrical substations," *Measurement*, vol. 147, Dec. 2019, Art. no. 106818.

[34] Y. Tian, S. Luo, B. Li, and W. Sun, "Optimal energy features of partial discharge signals in GIS extracted by Fisher linear discriminant," *Electr. Power*, vol. 52, no. 9, pp. 93–101, 2019.

[35] B. Han, C. Ma, J. Li, H. Wang, C. Liu, and T. Gao, "A feature parameters extraction method of PD UHF signal based on DTCWT and LLE algorithm," *Power Syst. Protection Control*, vol. 47, no. 20, pp. 65–72, 2019.

[36] R. Rostaminia, M. Saniei, M. Vakilian, S. S. Mortazavi, and V. P. Darabad, "An efficient partial discharge pattern recognition method using texture analysis for transformer defect models," *Int. Trans. Electr. Energy Syst.*, vol. 28, no. 7, p. e2558, Jul. 2018.

[37] J. Long, X. Wang, D. Dai, M. Tian, G. Zhu, and Y. Huang, "Feature extraction and classification of UHF PD signals based on improved S-transform," *High Voltage Eng.*, vol. 44, pp. 3649–3656, Jan. 2018.

[38] D. Dai, X. Wang, J. Long, M. Tian, G. Zhu, and J. Zhang, "Feature extraction of GIS partial discharge signal based on S-transform and singular value decomposition," *IET Sci., Meas. Technol.*, vol. 11, no. 2, pp. 186–193, 2017.

[39] X. Li, X. Wang, D. Xie, X. Wang, A. Yang, and M. Rong, "Time–frequency analysis of PD-induced UHF signal in GIS and feature extraction using invariant moments," *IET Sci., Meas. Technol.*, vol. 12, no. 2, pp. 169–175, Mar. 2018.

[40] 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.

[41] R. Liao, C. Guo, K. Wang, Z. Zuo, and A. Zhuang, "Adaptive optimal kernel time–frequency representation technique for partial discharge ultra-high-frequency signals classification," *Electr. Power Compon. Syst.*, vol. 43, no. 4, pp. 449–460, Feb. 2015.

[42] J. A. Ardila-Rey, M. Cerda-Luna, R. Rozas-Valderrama, B. A. de Castro, A. L. Andreoli, and B. Cevallos, "A new technique for separation of partial discharge sources and electromagnetic noise in radiofrequency measurements using energy ratios of different antennas," *High Voltage*, vol. 6, no. 3, pp. 525–530, Jun. 2021.

[43] D. Wotzka, W. Sikorski, and C. Szymczak, "Investigating the capability of PD-type recognition based on UHF signals recorded with different antennas using supervised machine learning," *Energies*, vol. 15, no. 9, p. 3167, Apr. 2022.

[44] X. Wang, X. Li, M. Rong, D. Xie, D. Ding, and Z. Wang, "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.

[45] S. Govindarajan, J. A. Ardila-Rey, K. Krithivasan, J. Subbaiah, N. Sannidhi, and M. Balasubramanian, "Development of hypergraph based improved random forest algorithm for partial discharge pattern classification," *IEEE Access*, vol. 9, pp. 96–109, 2021.

[46] V. Basharan, W. I. Maria Siluvairaj, and M. R. Velayutham, "Recognition of multiple partial discharge patterns by multi-class support vector machine using fractal image processing technique," *IET Sci., Meas. Technol.*, vol. 12, no. 8, pp. 1031–1038, Nov. 2018.

[47] J. Chen, C. Xu, and P. Li, "A feature extraction method for partial discharge pattern in GIS based on time-frequency analysis and fractal theory," *High Voltage Eng.*, vol. 47, no. 1, pp. 287–295, 2021.

[48] F. Gu, H. Chen, and M. Chao, "Application of improved Hilbert–Huang transform to partial discharge defect model recognition of power cables," *Appl. Sci.*, vol. 7, no. 10, p. 1021, Oct. 2017.

[49] F.-C. Gu, H.-C. Chen, and B.-Y. Chen, "A fractional Fourier transform-based approach for gas-insulated switchgear partial discharge recognition," *J. Electr. Eng. Technol.*, vol. 14, no. 5, pp. 2073–2084, Sep. 2019.

[50] M.-H. Wang, S.-D. 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.

[51] X. Zhang, S. Xiao, N. Shu, J. Tang, and W. Li, "GIS partial discharge pattern recognition based on the chaos theory," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 21, no. 2, pp. 783–790, Apr. 2014.

[52] I.-J. Seo, U. A. Khan, J.-S. Hwang, J.-G. Lee, and J.-Y. Koo, "Identification of insulation defects based on chaotic analysis of partial discharge in HVDC superconducting cable," *IEEE Trans. Appl. Supercond.*, vol. 25, no. 3, pp. 1–5, Jun. 2015.

[53] F.-C. Gu, H.-T. Yau, and H.-C. Chen, "Application of chaos synchronization technique and pattern clustering for diagnosis analysis of partial discharge in power cables," *IEEE Access*, vol. 7, pp. 76185–76193, 2019.

[54] 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.

[55] W. Wenbo, S. Lin, W. Bin, and Y. Min, "Partial discharge feature extraction based on synchrosqueezed windowed Fourier transform and multi-scale dispersion entropy," *Meas. Control*, vol. 53, nos. 7–8, pp. 1078–1087, Aug. 2020.

[56] Y. Xu, G. Sheng, F. Yang, X. Chen, Y. Qian, and X. Jiang, "Classification of partial discharge images within DC XLPE cables in contourlet domain," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 25, no. 2, pp. 486–493, Apr. 2018.

[57] 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.

[58] S. Luo, Y. Tian, B. Li, Y. Hu, and Q. Li, "Pattern recognition of ultra-high frequency partial discharge by using scale parameters-energy entropy characteristic pairs," *Electr. Power Eng. Technol.*, vol. 38, no. 4, pp. 152–158, 2019.

[59] S. Qian, Y. Peng, Z. Zhou, J. Chen, and M. Tang, "Partial discharge pattern recognition using hidden Markov models based on the entropy lifting wavelet coefficients," *Electr. Eng.*, vol. 21, no. 10, pp. 93–102, 2020.

[60] R. Jia, Y. Xie, H. Wu, J. Dang, and K. Dong, "Power transformer partial discharge fault diagnosis based on multidimensional feature region," *Math. Problems Eng.*, vol. 2016, pp. 1–11, Jan. 2016.

[61] S. S. Roy and S. Chatterjee, "Partial discharge detection framework employing spectral analysis of horizontal visibility graph," *IEEE Sensors J.*, vol. 21, no. 4, pp. 4819–4826, Feb. 2021.
[62] S. S. Roy and S. Chatterjee, "Complex network aided partial discharge signal recognition framework employing visibility graph," *IEEE Sensors Lett.*, vol. 4, no. 8, pp. 1–4, Aug. 2020.
[63] M. Ren, J. Zhou, and J. Miao, "Adopting spectral analysis in partial discharge fault diagnosis of GIS with a micro built-in optical sensor," *IEEE Trans. Power Del.*, vol. 36, no. 2, pp. 1237–1240, Apr. 2021.
[64] M. Ren, B. Song, T. Zhuang, and S. Yang, "Optical partial discharge diagnostic in SF6 gas insulated system via multi-spectral detection," *ISA Trans.*, vol. 75, pp. 247–257, Apr. 2018.
[65] Z. Cheng, F. Yang, B. Gao, P. Yu, Q. Yang, J. Tian, and X. Lu, "Partial discharge pattern recognition of XLPE cable based on vector quantization," *IEEE Trans. Magn.*, vol. 55, no. 6, pp. 1–4, Jun. 2019.
[66] J. Tang, D. Wang, L. Fan, R. Zhuo, and X. Zhang, "Feature parameters extraction of gis partial discharge signal with multifractal detrended fluctuation analysis," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 22, no. 5, pp. 3037–3045, Oct. 2015.
[67] D.-E.-A. Mansour, I. B. M. Taha, R. A. Farade, and N. I. B. A. Wahab, "Partial discharge diagnosis in GIS based on pulse sequence features and optimized machine learning classification techniques," *Electr. Power Syst. Res.*, vol. 211, Oct. 2022, Art. no. 108162.
[68] 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.
[69] Z. Pan, L. Liu, C. Qian, Z. Wang, and D. Yuan, "Pattern recognition of partial discharge based on fusion extreme learning machine," *Electr. Power Eng. Technol.*, vol. 38, no. 5, pp. 42–48, 2019.
[70] R. Yao, J. Li, M. Hui, and L. Bai, "Pattern recognition for partial discharge using adaptive boost classification model based on ensemble method," *Power Syst. Technol.*, vol. 46, no. 6, pp. 2410–2420, 2022.
[71] R. Yao, J. Li, M. Hui, L. Bai, and Q. Wu, "Pattern recognition for partial discharge using multi-feature combination adaptive boost classification model," *IEEE Access*, vol. 9, pp. 48873–48883, 2021.
[72] X. Peng, J. Li, G. Wang, Y. Wu, L. Li, Z. Li, A. A. Bhatti, C. Zhou, D. M. Hepburn, A. J. Reid, M. D. Judd, and W. H. Siew, "Random forest based optimal feature selection for partial discharge pattern recognition in HV cables," *IEEE Trans. Power Del.*, vol. 34, no. 4, pp. 1715–1724, Aug. 2019.
[73] W. Si, S. Li, H. Xiao, Q. Li, Y. Shi, and T. Zhang, "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.
[74] L. Luo, B. Han, J. Chen, G. Sheng, and X. Jiang, "Partial discharge detection and recognition in random matrix theory paradigm," *IEEE Access*, vol. 5, pp. 8205–8213, 2017.
[75] S. Govindarajan, V. Ragavan, A. El-Hag, K. Krithivasan, and J. Subbaiah, "Development of Hankel singular-hypergraph feature extraction technique for acoustic partial discharge pattern classification," *Energies*, vol. 14, no. 6, p. 1564, Mar. 2021.
[76] G. C. Montanari, R. Ghosh, L. Cirioni, G. Galvagno, and S. Mastroeni, "Partial discharge monitoring of medium voltage switchgears: Self-condition assessment using an embedded bushing sensor," *IEEE Trans. Power Del.*, vol. 37, no. 1, pp. 85–92, Feb. 2022.
[77] G. M. Mengounou, J. L. Jiosseu, and A. M. Imano, "Experimental investigation of partial discharges activities in palm kernel oil methyl ester and mineral oil by using a Rogowski coil," *Electr. Power Syst. Res.*, vol. 213, Dec. 2022, Art. no. 108708.
[78] Y. Li, X. Wang, Z. Liu, X. Liang, and S. Si, "The entropy algorithm and its variants in the fault diagnosis of rotating machinery: A review," *IEEE Access*, vol. 6, pp. 66723–66741, 2018.
[79] P. Ray, S. S. Reddy, and T. Banerjee, "Various dimension reduction techniques for high dimensional data analysis: A review," *Artif. Intell. Rev.*, vol. 54, no. 5, pp. 3473–3515, Jun. 2021.
[80] X. Huang, "Research and development of feature dimensionality reduction," *Comput. Sci.*, vol. 45, no. 6A, pp. 16–21, 2018.
[81] M. Sun, L. Yang, Z. Zhou, C. Gao, A. Zhang, and J. Li, "Pattern recognition for partial discharge of cable accessories based on SLLE," *Elect. Meas. Instrum.*, vol. 56, no. 22, pp. 25–30, 2019.
[82] R. Yao, J. Li, M. Hui, L. Bai, and Q. Wu, "Feature selection based on random forest for partial discharges characteristic set," *IEEE Access*, vol. 8, pp. 159151–159161, 2020.
[83] S. Song, Y. Qian, H. Wang, G. Sheng, and X. Jiang, "Improved algorithm for partial discharge pattern recognition based on histogram of oriented gradient attribute space," *Trans. China Electrotech. Soc.*, vol. 36, no. 10, pp. 2153–2160, 2021.
[84] S. Zhang, C. Li, K. Wang, J. Li, R. Liao, T. Zhou, and Y. Zhang, "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.
[85] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Comput. Elect. Eng.*, vol. 40, no. 1, pp. 16–28, Jan. 2014.
[86] B. H. Nguyen, B. Xue, and M. Zhang, "A survey on swarm intelligence approaches to feature selection in data mining," *Swarm Evol. Comput.*, vol. 54, May 2020, Art. no. 100663.
[87] H. Illias, G. Altamimi, N. Mokhtar, and H. Arof, "Classification of multiple partial discharge sources in dielectric insulation material using cepstrum analysis–artificial neural network," *IEEJ Trans. Electr. Electron. Eng.*, vol. 12, no. 3, pp. 357–364, May 2017.
[88] S. Polisetty, A. El-Hag, and S. Jayram, "Classification of common discharges in outdoor insulation using acoustic signals and artificial neural network," *High Voltage*, vol. 4, no. 4, pp. 333–338, Dec. 2019.
[89] Z. Li, L. Qian, and J. Li, "Type recognition of partial discharge in power transformer based on statistical characteristics and PNN," *Power Syst. Protection Control*, vol. 46, no. 13, pp. 55–60, 2018.
[90] J. Shu, L. Xu, Y. Lv, W. Duan, and S. Zhong, "Partial discharge pattern recognition of cables based on fuzzy adaptive resonance theory neural network," *Zhejiang Electr. Power*, vol. 40, no. 11, pp. 10–15, 2021.
[91] L. Wang, Y. Liu, and Y. Wang, "Partial discharge identification method in GIS based on EEMD energy moment and ISSA-SVM algorithm," *J. Electron. Meas. Instrum.*, vol. 36, no. 5, pp. 204–212, 2022.
[92] K. Sun, Y. Meng, and S. Dong, "FoSSA optimization-based SVM classifier for the recognition of partial discharge patterns in HV cables," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–12, Mar. 2022.
[93] Y. Zhou, Y. Liu, N. Wang, X. Han, and J. Li, "Partial discharge ultrasonic signals pattern recognition in transformer using BSO-SVM based on microfiber coupler sensor," *Measurement*, vol. 201, Sep. 2022, Art. no. 111737.
[94] L. A. Lumba, U. Khayam, and L. S. Lumba, "Application of fuzzy logic for partial discharge pattern recognition," in *Proc. Int. Conf. Elect. Eng. Inform.*, 2019, pp. 210–215.
[95] 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.
[96] D. Zhou, X. Zhang, Y. Zou, Y. Ni, and D. Wang, "Study on partial discharge pattern recognition for distribution cable based on T–F clustering and PRPD spectrum analysis," *J. Elect. Eng.*, vol. 17, no. 2, pp. 235–242, 2022.
[97] T. K. Abdel-Galil, R. M. Sharkawy, M. M. A. Salama, and R. Bartnikas, "Partial discharge pattern classification using the fuzzy decision tree approach," *IEEE Trans. Instrum. Meas.*, vol. 54, no. 6, pp. 2258–2263, Dec. 2005.
[98] Y. Xu, H. Xia, Z. Li, and M. Lv, "The partial discharge pattern recognition for the transformers using dynamic mode decomposition fractal features," *Proc. CSU-EPSA*, vol. 31, no. 12, pp. 35–43, 2019.
[99] L. C. C. Heredia and A. R. Mor, "Density-based clustering methods for unsupervised separation of partial discharge sources," *Int. J. Electr. Power Energy Syst.*, vol. 107, pp. 224–230, May 2019.
[100] K. Firuzi, M. Vakilian, B. T. Phung, and T. Blackburn, "Online monitoring of transformer through stream clustering of partial discharge signals," *IET Sci., Meas. Technol.*, vol. 13, no. 3, pp. 409–415, May 2019.
[101] P. Seri, R. Ghosh, and G. C. Montanari, "An unsupervised approach to partial discharge monitoring in rotating machines: Detection to diagnosis with reduced need of expert support," *IEEE Trans. Energy Convers.*, vol. 36, no. 3, pp. 2485–2492, Sep. 2021.
[102] W. Liu, H. Wang, Z. Shi, D. Li, X. Hu, and J. Li, "Research on partial discharge pattern recognition of XLPE cable based on improved XGBoost algorithm," *Electr. Meas. Instrum.*, vol. 59, no. 4, pp. 98–106, 2022.
[103] A. Abubakar Mas'ud, B. G. Stewart, and S. G. Mcmeekin, "Application of an ensemble neural network for classifying partial discharge patterns," *Electr. Power Syst. Res.*, vol. 110, pp. 154–162, May 2014.

[104] K. Zhou, S.-K. Oh, and J. Qiu, "Design of ensemble fuzzy-RBF neural networks based on feature extraction and multi-feature fusion for GIS partial discharge recognition and classification," *J. Electr. Eng. Technol.*, vol. 17, no. 1, pp. 513–532, Jan. 2022.

[105] R. Deng, Y. Zhu, X. Liu, and Y. Zhai, "Pattern recognition of unknown types in partial discharge signals based on variable predictive model and Tanimoto," *Trans. China Electrotech. Soc.*, vol. 35, no. 14, pp. 3105–3115, 2020.

[106] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Trans. Neural Netw.*, vol. 10, no. 5, pp. 988–999, Aug. 1999.

[107] S. Li, Z. Yuan, C. Wang, T. Chen, and Z. Guo, "Optimization of support vector machine parameters based on group intelligence algorithm," *CAAI Trans. Intell. Syst.*, vol. 13, no. 1, pp. 70–84, 2018.

[108] I. D. Mienye and Y. Sun, "A survey of ensemble learning: Concepts, algorithms, applications, and prospects," *IEEE Access*, vol. 10, pp. 99129–99149, 2022.

[109] W. You, Q. Li, N. Yang, K. Shen, W. Li, and Z. Wu, "Electricity theft detection based on multiple different learners fusion by stacking ensemble learning," *Autom. Electr. Power Syst.*, vol. 46, no. 24, pp. 178–186, 2022.

[110] E. T. Iorkyase, C. Tachtatzis, I. A. Glover, P. Lazaridis, D. Upton, B. Saeed, and R. C. Atkinson, "Improving RF-based partial discharge localization via machine learning ensemble method," *IEEE Trans. Power Del.*, vol. 34, no. 4, pp. 1478–1489, Aug. 2019.

[111] Y. Jia, Y. Zhu, Z. Lan, and L. Wang, "Pattern recognition on partial discharge signals of transformers based on S-transform and deep belief network," *Guangdong Electr. Power*, vol. 30, no. 1, pp. 108–115, 2017.

[112] X. Zhang, J. Tang, C. Pan, X. Zhang, M. Jin, and D. Yang, "Research of partial discharge recognition based on deep belief nets," *Power Syst. Technol.*, vol. 40, no. 10, pp. 3283–3289, 2016.

[113] G. Huang, Z. Li, Y. Xu, Y. Qian, G. Sheng, and X. Jiang, "Partial discharge pattern recognition of XLPE DC cable based on improved deep belief networks," *High Voltage Eng.*, vol. 46, no. 1, pp. 327–334, 2020.

[114] Z. Li, Y. Xu, and X. Jiang, "Pattern recognition of DC partial discharge on XLPE cable based on ADAM-DBN," *Energies*, vol. 13, no. 17, p. 4566, Sep. 2020.

[115] M. Karimi, M. Majidi, H. MirSaeedi, M. M. Arefi, and M. Oskuoee, "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.

[116] N. T. Thi, T.-D. Do, J.-R. Jung, H. Jo, and Y.-H. Kim, "Anomaly detection for partial discharge in gas-insulated switchgears using autoencoder," *IEEE Access*, vol. 8, pp. 152248–152257, 2020.

[117] L. Duan, J. Hu, G. Zhao, K. Chen, J. He, and S. X. Wang, "Identification of partial discharge defects based on deep learning method," *IEEE Trans. Power Del.*, vol. 34, no. 4, pp. 1557–1568, Aug. 2019.

[118] Y. Xu, H. Xia, S. Xie, and M. Lu, "The pattern recognition of multisource partial discharge in transformers based on parallel feature domain," *IET Sci., Meas. Technol.*, vol. 15, no. 2, pp. 163–173, Mar. 2021.

[119] S. Barrios, D. Buldain, M. P. Comech, and I. Gilbert, "Partial discharge identification in MV switchgear using scalogram representations and convolutional AutoEncoder," *IEEE Trans. Power Del.*, vol. 36, no. 6, pp. 3448–3455, Dec. 2021.

[120] J. Dai, Y. Teng, Z. Zhang, Z. Yu, G. Sheng, and X. Jiang, "Partial discharge data matching method for GIS case-based reasoning," *Energies*, vol. 12, no. 19, p. 3677, Sep. 2019.

[121] R. Deng, Y. Zhu, X. Liu, and Y. Zhai, "Pattern recognition of unknown signals for partial discharge based on L2-VAE," *Guangdong Electr. Power*, vol. 32, no. 9, pp. 69–77, 2019.

[122] R. Zemouri, M. Lévesque, N. Amyot, C. Hudon, O. Kokoko, and S. A. Tahan, "Deep convolutional variational autoencoder as a 2D-visualization tool for partial discharge source classification in hydrogenerators," *IEEE Access*, vol. 8, pp. 5438–5454, 2020.

[123] T.-D. Do, V.-N. Tuyet-Doan, Y.-S. Cho, J.-H. Sun, and Y.-H. Kim, "Convolutional-neural-network-based partial discharge diagnosis for power transformer using UHF sensor," *IEEE Access*, vol. 8, pp. 207377–207388, 2020.

[124] X. Peng, F. Yang, G. Wang, Y. Wu, L. Li, Z. Li, A. A. Bhatti, C. Zhou, D. M. Hepburn, A. J. Reid, M. D. Judd, and W. H. Siew, "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] S. Mantach, A. Ashraf, H. Janani, and B. Kordi, "A convolutional neural network-based model for multi-source and single-source partial discharge pattern classification using only single-source training set," *Energies*, vol. 14, no. 5, p. 1355, Mar. 2021.

[126] F.-C. Gu, "Identification of partial discharge defects in gas-insulated switchgears by using a deep learning method," *IEEE Access*, vol. 8, pp. 163894–163902, 2020.

[127] Y. Zhu, Y. Xu, X. Chen, G. Sheng, and X. Jiang, "Pattern recognition of partial discharges in DC XLPE cables based on convolutional neural network," *Trans. China Electrotech. Soc.*, vol. 35, no. 3, pp. 659–668, 2020.

[128] Z. Zhang, H. Yue, B. Wang, Y. Liu, and S. Luo, "Pattern recognition of partial discharge ultrasonic signal based on similar matrix BSS and deep learning CNN," *Power Syst. Technol.*, vol. 43, no. 6, pp. 1900–1906, 2019.

[129] C.-K. Chang, H.-H. Chang, and B. K. Boyanapalli, "Application of pulse sequence partial discharge based convolutional neural network in pattern recognition for underground cable joints," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 29, no. 3, pp. 1070–1078, Jun. 2022.

[130] 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.

[131] Y. Wang, J. Yan, Z. Yang, T. Liu, Y. Zhao, and J. Li, "Partial discharge pattern recognition of gas-insulated switchgear via a light-scale convolutional neural network," *Energies*, vol. 12, no. 24, p. 4674, Dec. 2019.

[132] Q. Che, H. Wen, X. Li, Z. Peng, and K. P. Chen, "Partial discharge recognition based on optical fiber distributed acoustic sensing and a convolutional neural network," *IEEE Access*, vol. 7, pp. 101758–101764, 2019.

[133] M. Borghei and M. Ghassemi, "A deep learning approach for discrimination of single- and multi-source corona discharges," *IEEE Trans. Plasma Sci.*, vol. 49, no. 9, pp. 2936–2945, Sep. 2021.

[134] A. Gao, Y. Zhu, Y. Zhang, and W. Cai, "Partial discharge pattern recognition of power transformers based on marginal spectrum images and deep residual network," *Power Syst. Technol.*, vol. 45, no. 6, pp. 2433–2442, 2021.

[135] Y. Jia, M. Deng, Y. Li, C. Ai, J. Yang, and C. Liu, "Research on GIS partial discharge pattern recognition based on deep residual network," *High Voltage App.*, vol. 54, no. 11, pp. 123–129, 2018.

[136] 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.

[137] X. Wan, H. Song, L. Luo, Z. Li, G. Sheng, and X. Jiang, "Application of convolutional neural networks in pattern recognition of partial discharge image," *Power Syst. Technol.*, vol. 43, no. 6, pp. 2219–2226, 2019.

[138] J. Wu, Y. Hu, C. Jiang, X. Zhang, K. Li, and Z. Zhao, "Classification of partial discharge signals in high-voltage cables based on convolutional neural network," *J. Wuhan Univ.*, vol. 67, no. 3, pp. 232–240, 2021.

[139] Z. Congcong, W. Gang, and G. Dong, "Partial discharge pattern recognition based on convolutional neural network," *Adv. Technol. Electr. Eng. Energy*, vol. 40, no. 3, pp. 72–80, 2021.

[140] W. J. K. Raymond, C. W. Xin, L. W. Kin, and H. A. Illias, "Noise invariant partial discharge classification based on convolutional neural network," *Measurement*, vol. 177, Jun. 2021, Art. no. 109220.

[141] J. Kim and K.-I. Kim, "Partial discharge online detection for long-term operational sustainability of on-site low voltage distribution network using CNN transfer learning," *Sustainability*, vol. 13, no. 9, p. 4692, Apr. 2021.

[142] Z. Tang, Z. Cao, and N. He, "Application of convolutional neural network transfer learning in partial discharge type diagnosis," *High Voltage App.*, vol. 58, no. 4, pp. 158–164, 2022.

[143] Y. Wang, J. Yan, Z. Yang, Q. Jing, Z. Qi, J. Wang, and Y. Geng, "A domain adaptive deep transfer learning method for gas-insulated switchgear partial discharge diagnosis," *IEEE Trans. Power Del.*, vol. 37, no. 4, pp. 2514–2523, Aug. 2022.

[144] Y. Wang, J. Yan, Z. Yang, J. Wang, and Y. Geng, "A novel 1DCNN and domain adversarial transfer strategy for small sample GIS partial discharge pattern recognition," *Meas. Sci. Technol.*, vol. 32, no. 12, Dec. 2021, Art. no. 125118.

[145] Y. Wang, J. Yan, Q. Sun, J. Li, and Z. Yang, "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, 2019.

[146] Y. Wang, J. Yan, Z. Yang, Y. Zhao, and T. Liu, "GIS partial discharge pattern recognition via lightweight convolutional neural network in the ubiquitous power Internet of Things context," *IET Sci., Meas. Technol.*, vol. 14, no. 8, pp. 864–871, Oct. 2020.

[147] Y. Wang, J. Yan, Z. Yang, Y. Zhao, and T. Liu, "Optimizing GIS partial discharge pattern recognition in the ubiquitous power Internet of Things context: A MixNet deep learning model," *Int. J. Electr. Power Energy Syst.*, vol. 125, Feb. 2021, Art. no. 106484.

[148] V.-N. Tuyet-Doan, T.-D. Do, N.-D. Tran-Thi, Y.-W. Youn, and Y.-H. Kim, "One-shot learning for partial discharge diagnosis using ultra-high-frequency sensor in gas-insulated switchgear," *Sensors*, vol. 20, no. 19, p. 5562, Sep. 2020.

[149] Y. Zhang and Y. Zhu, "A lightweight partial discharge diagnosis method of power equipment based on depth-width joint pruning," *Trans. China Electrotech. Soc.*, vol. 38, no. 7, pp. 1935–1945, 2023.

[150] A. Gao, Y. Zhu, W. Cai, and Y. Zhang, "Pattern recognition of partial discharge based on VMD-CWD spectrum and optimized CNN with cross-layer feature fusion," *IEEE Access*, vol. 8, pp. 151296–151306, 2020.

[151] J. Chen, Y. Zhou, Z. Bai, Y. Zhao, Y. Zhang, and L. Zhang, "Pattern recognition method of partial discharge in oil-paper insulation based on multi-channel convolutional neural network," *High Voltage Eng.*, vol. 48, no. 5, pp. 1705–1715, 2022.

[152] G. Michau, C.-C. Hsu, and O. Fink, "Interpretable detection of partial discharge in power lines with deep learning," *Sensors*, vol. 21, no. 6, p. 2154, Mar. 2021.

[153] S. Mantach, P. Gill, D. R. Oliver, A. Ashraf, and B. Kordi, "An interpretable CNN model for classification of partial discharge waveforms in 3D-printed dielectric samples with different void sizes," *Neural Comput. Appl.*, vol. 34, no. 14, pp. 11739–11750, Jul. 2022.

[154] R. Kitani and S. Iwata, "Verification of interpretability of phase-resolved partial discharge using a CNN with SHAP," *IEEE Access*, vol. 11, pp. 4752–4762, 2023.

[155] Q. Jing, J. Yan, Y. Wang, R. He, and L. Lu, "A novel differentiable neural network architecture automatic search method for GIS partial discharge pattern recognition," *Measurement*, vol. 195, May 2022, Art. no. 111154.

[156] 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.

[157] B. Adam and S. Tenbohlen, "Classification of superimposed partial discharge patterns," *Energies*, vol. 14, no. 8, p. 2144, Apr. 2021.

[158] M. Dong and J. Sun, "Partial discharge detection on aerial covered conductors using time-series decomposition and long short-term memory network," *Electr. Power Syst. Res.*, vol. 184, Jul. 2020, Art. no. 106318.

[159] J. Guo, Y. Zhao, Z. Wang, and L. Ding, "Partial discharge pattern recognition method of typical defects in basin insulator based on LMD and LSTM," *Southern Power Syst. Technol.*, vol. 15, no. 8, pp. 95–105, 2021.

[160] V.-N. Tuyet-Doan, T.-T. Nguyen, M.-T. Nguyen, J.-H. Lee, and Y.-H. Kim, "Self-attention network for partial-discharge diagnosis in gas-insulated switchgear," *Energies*, vol. 13, no. 8, p. 2102, Apr. 2020.

[161] J. A. Ardila-Rey, J. E. Ortiz, W. Creixell, F. Muhammad-Sukki, and N. A. Bani, "Artificial generation of partial discharge sources through an algorithm based on deep convolutional generative adversarial networks," *IEEE Access*, vol. 8, pp. 24561–24575, 2020.

[162] C. Xu, J. Chen, W. Liu, P. Li, and M. Zhu, "Pattern recognition of partial discharge PRPD spectrum in GIS based on deep residual network," *High Voltage Eng.*, vol. 48, no. 3, pp. 1113–1123, 2022.

[163] Y. Fu, K. Zhou, G. Zhu, Z. Wang, G. Wang, and Z. Wang, "A method for improving the recognition accuracy of cable termination partial discharge based on improved WGAN algorithm," *Power Syst. Technol.*, vol. 46, no. 5, pp. 2000–2008, 2022.

[164] Y. Zhu, Y. Zhang, W. Cai, and A. Gao, "Data augmentation and pattern recognition for multi-sources partial discharge based on boundary equilibrium generative adversarial network with auxiliary classifier," *Proc. CSEE*, vol. 41, no. 14, pp. 5044–5053, 2021.

[165] Y. Wang, J. Yan, Z. Yang, Q. Jing, J. Wang, and Y. Geng, "GAN and CNN for imbalanced partial discharge pattern recognition in GIS," *High Voltage*, vol. 7, no. 3, pp. 452–460, Jun. 2022.

[166] Q. Jing, J. Yan, Y. Wang, X. Ye, J. Wang, and Y. Geng, "A novel method for small and unbalanced sample pattern recognition of gas insulated switchgear partial discharge using an auxiliary classifier generative adversarial network," *High Voltage*, vol. 8, no. 2, pp. 368–379, Apr. 2023.

[167] Y. Zhang and Y. Zhu, "A partial discharge pattern recognition method combining graph signal and graph convolutional network," *Proc. CSEE*, vol. 41, no. 18, pp. 6472–6481, 2021.

[168] Y. Wang, J. Yan, X. Ye, Z. Qi, J. Wang, and Y. Geng, "GIS partial discharge pattern recognition via a novel capsule deep graph convolutional network," *IET Gener., Transmiss. Distrib.*, vol. 16, no. 14, pp. 2903–2912, Jul. 2022.

[169] Y. Zhang and Y. Zhu, "Incremental partial discharge recognition method combining knowledge distillation with graph neural network," *Trans. China Electrotech. Soc.*, vol. 38, no. 5, pp. 1390–1400, 2023.

[170] J. Tian, H. Song, G. Sheng, and X. Jiang, "Knowledge-driven recognition methodology of partial discharge patterns in GIS," *IEEE Trans. Power Del.*, vol. 37, no. 4, pp. 3335–3344, Aug. 2022.

[171] V.-N. Tuyet-Doan, H.-A. Pho, B. Lee, and Y.-H. Kim, "Deep ensemble model for unknown partial discharge diagnosis in gas-insulated switchgears using convolutional neural networks," *IEEE Access*, vol. 9, pp. 80524–80534, 2021.

[172] J. Yeo, H. Jin, A. R. Mor, C. Yuen, W. Tushar, T. K. Saha, and C. S. Ng, "Identification of partial discharge through cable-specific adaption and neural network ensemble," *IEEE Trans. Power Del.*, vol. 37, no. 3, pp. 1598–1607, Jun. 2022.

[173] Q. Jing, J. Yan, L. Lu, Y. Xu, and F. Yang, "A novel method for pattern recognition of GIS partial discharge via multi-information ensemble learning," *Entropy*, vol. 24, no. 7, p. 954, Jul. 2022.

[174] L. Klein, D. Seidl, J. Fulneč ek, L. Prokop, S. Mišák, and J. Dvorský, "Antenna contactless partial discharges detection in covered conductors using ensemble stacking neural networks," *Expert Syst. Appl.*, vol. 213, Mar. 2023, Art. no. 118910.

[175] J. Gao, Y. Zhu, Y. Zheng, K. Zhang, and S. Liu, "Pattern recognition of partial discharge based on Hilbert marginal spectrum and sparse autoencoder-deep neural networks," *Autom. Electr. Power Syst.*, vol. 43, no. 1, pp. 87–94, 2019.

[176] X. Zhou, X. Wu, P. Ding, X. Li, N. He, G. Zhang, and X. Zhang, "Research on transformer partial discharge UHF pattern recognition based on CNN-LSTM," *Energies*, vol. 13, no. 1, p. 61, Dec. 2019.

[177] T. Liu, J. Yan, Y. Wang, Y. Xu, and Y. Zhao, "GIS partial discharge pattern recognition based on a novel convolutional neural networks and long short-term memory," *Entropy*, vol. 23, no. 6, p. 774, Jun. 2021.

[178] N. Qu, Z. Li, J. Zuo, and J. Chen, "Fault detection on insulated overhead conductors based on DWT-LSTM and partial discharge," *IEEE Access*, vol. 8, pp. 87060–87070, 2020.

[179] Z. Li, N. Qu, X. Li, J. Zuo, and Y. Yin, "Partial discharge detection of insulated conductors based on CNN-LSTM of attention mechanisms," *J. Power Electron.*, vol. 21, no. 7, pp. 1030–1040, Jul. 2021.

[180] S. Jin, R. Zhang, and G. Du, "Partial discharge pattern recognition of cables considering attenuation of damped AC voltages," *High Voltage Eng.*, vol. 47, no. 7, pp. 2583–2590, 2021.

[181] G. Li, X. Wang, X. Li, A. Yang, and M. Rong, "Partial discharge recognition with a multi-resolution convolutional neural network," *Sensors*, vol. 18, no. 10, p. 3512, Oct. 2018.

[182] J. Lai, X. Wang, Q. Xiang, Y. Song, and W. Quan, "Review on autoencoder and its application," *J. Commun.*, vol. 42, no. 9, pp. 218–230, 2021.

[183] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021.

[184] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Jan. 2009.

[185] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.

[186] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2017, pp. 7167–7176.

[187] X. Qiu, *Neural Networks and Deep Learning*. Beijing, China: China Mach. Press, 2020.

[188] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 2672–2680.

[189] Z. Shao, C. Zhang, F. Chen, and Y. Xie, "A review on generative adversarial networks for power system applications," *Proc. CSEE*, vol. 43, no. 3, pp. 987–1004, 2023.

[190] B. Wu, X. Liang, S. Zhang, and R. Xu, "Advances and applications in graph neural network," *Chin. J. Comput.*, vol. 45, no. 1, pp. 35–68, 2022.

[191] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*.

[192] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan, "Ensemble deep learning: A review," *Eng. Appl. Artif. Intel.*, vol. 115, Jan. 2022, Art. no. 105151.

[193] G.-R. Han, J.-C. Jeon, and J.-M. Park, "Cast-resin transformers: Development of health assessment technique based on electromagnetic partial discharge detection," *Electr. Eng.*, vol. 101, no. 4, pp. 1211–1220, Dec. 2019.

[194] W. Gao, D. Ding, W. Liu, and X. Huang, "Investigation of the evaluation of the PD severity and verification of the sensitivity of partial-discharge detection using the UHF method in GIS," *IEEE Trans. Power Del.*, vol. 29, no. 1, pp. 38–47, Feb. 2014.

[195] B. Qi, C. Li, B. Geng, and Z. Hao, "Severity diagnosis and assessment of the partial discharge provoked by high-voltage electrode protrusion on GIS insulator surface," *IEEE Trans. Power Del.*, vol. 26, no. 4, pp. 2363–2369, Oct. 2011.

[196] 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., Transmiss. Distrib.*, vol. 9, no. 16, pp. 2493–2500, Dec. 2015.

[197] J. Tang, M. Jin, F. Zeng, S. Zhou, X. Zhang, Y. Yang, and Y. Ma, "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.

[198] N. H. He, W. Sha, and X. Li, "Evaluation on partial discharge severity of free metal particles defects in GIS," *Insulating Mater.*, vol. 52, no. 12, pp. 80–88, 2019.

[199] Y. Man, Y. Dong, W. Wang, Y. Xiao, T. Liu, and Q. Li, "Discharge severity evaluation method for typical defects of oil-paper insulation based on partial discharge energy," *High Voltage App.*, vol. 55, no. 8, pp. 223–229, 2019.

[200] S. Li, W. Si, and Q. Li, "Partition and recognition of partial discharge development stages in oil-pressboard insulation with needle-plate electrodes under combined AC–DC voltage stress," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 24, no. 3, pp. 1781–1793, Jun. 2017.

[201] Z. Xu, L. Yang, and Y. Wei, "Partial discharge characteristics of fast developing discharge in oil-pressboard insulation," *Proc. CSEE*, vol. 41, no. 21, pp. 7529–7540, 2021.

[202] W. Fan, Z. Zhang, C. Xia, M. Dong, X. Gao, and K. Wang, "Evolution law of partial discharge of oil-paper insulated needle-plate electrode based on multi-method measurement," *High Voltage Eng.*, vol. 48, no. 3, pp. 914–927, 2022.

[203] Y. Cui, P. Cao, L. Zhu, J. Deng, P. Peng, and S. Ji, "Electrical aging characteristic of oil immersed pressboard under partial discharge activity," *High Voltage Eng.*, vol. 45, no. 9, pp. 2945–2953, 2019.

[204] J. Cui, H. Dong, S. Li, and Y. Kang, "Aging state evaluation of insulating paper based on image feature recognition," *High Voltage Eng.*, vol. 48, no. 2, pp. 636–643, 2022.

[205] A. S. Karandaev, I. M. Yachikov, A. A. Radionov, I. V. Liubimov, N. N. Druzhinin, and E. A. Khramshina, "Fuzzy algorithms for diagnosis of furnace transformer insulation condition," *Energies*, vol. 15, no. 10, p. 3519, May 2022.

[206] J. Li, W. Gao, and W. Liu, "Evolution process and severity assessment of the partial discharge caused by immobilized metal particles on DC GIL insulator surface," *High Voltage App.*, vol. 54, no. 5, pp. 9–16, 2018.

[207] J. Tang, X. Yang, D. Yang, Q. Yao, Y. Miao, C. Zhang, and F. Zeng, "Using $SF_6$ decomposed component analysis for the diagnosis of partial discharge severity initiated by free metal particle defect," *Energies*, vol. 10, no. 8, p. 1119, Aug. 2017.

[208] W. Xu, Y. Nie, Y. Chen, and H. Cai, "Evaluation method of severity of partial discharge in solid insulated switchgear," *High Voltage App.*, vol. 55, no. 1, pp. 27–33, 2019.

[209] D. Ma, L. Jin, J. He, and K. Gao, "Classification of partial discharge severities of ceramic insulators based on texture analysis of UV pulses," *High Voltage*, vol. 6, no. 6, pp. 986–996, Dec. 2021.

[210] 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, Art. no. 106053.

[211] F. Yang, G. Sheng, Y. Xu, H. Hou, Y. Qian, and X. Jiang, "Partial discharge pattern recognition of XLPE cables at DC voltage based on the compressed sensing theory," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 24, no. 5, pp. 2977–2985, Oct. 2017.

[212] C. Madariaga, R. Schurch, J. Ardila-Rey, O. Muñoz, and S. Fingerhuth, "Partial discharge electrical tree growth identification by means of waveform source separation techniques," *IEEE Access*, vol. 9, pp. 64665–64675, 2021.

[213] L. Li, Y. Wang, J. Gao, N. Guo, and G. Han, "Pattern recognition of growth characteristics based on UHF PD signals of electrical tree," *IEEE Access*, vol. 10, pp. 49853–49861, 2022.

[214] H. Song, J. Dai, Z. Li, L. Luo, G. Sheng, and X. Jiang, "An assessment method of partial discharge severity for GIS in service," *Proc. CSEE*, vol. 39, no. 4, pp. 1231–1241, 2019.

[215] X. Meng, H. Song, J. Dai, L. Luo, G. Sheng, and X. Jiang, "Severity evaluation of UHF signals of partial discharge in GIS based on semantic analysis," *IEEE Trans. Power Del.*, vol. 37, no. 3, pp. 1456–1464, Jun. 2022.

[216] J. Tang, M. Jin, F. Zeng, X. Zhang, and R. Huang, "Assessment of PD severity in gas-insulated switchgear with an SSAE," *IET Sci., Meas. Technol.*, vol. 11, no. 4, pp. 423–430, Jul. 2017.

[217] Y. Wang, J. Wu, T. Han, K. Haran, and Y. Yin, "Insulation condition forewarning of form-wound winding for electric aircraft propulsion based on partial discharge and deep learning neural network," *High Voltage*, vol. 6, no. 2, pp. 302–313, Apr. 2021.

[218] P. Li, P. Wang, X. Zhang, J. Wang, and H. Guo, "Prediction of partial discharge inception voltage for inverter-fed motor based on deep belief network," *Chin. J. Sci. Instrum.*, vol. 42, no. 4, pp. 121–130, 2021.

[219] R. Sahoo and S. Karmakar, "Investigation of electrical tree growth characteristics and partial discharge pattern analysis using deep neural network," *Electr. Power Syst. Res.*, vol. 220, Jul. 2023, Art. no. 109287.

[220] Z. Yuan, Q. Ye, Y. Wang, and Z. Guo, "State recognition of surface discharges by visible images and machine learning," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–11, 2021.

[221] C. Li et al., "Insulating materials for realising carbon neutrality: Opportunities, remaining issues and challenges," *High Voltage*, vol. 7, no. 4, pp. 610–632, Aug. 2022.

[222] J. Jiang, B. Zhang, Z. Li, P. Ranjan, J. Chen, and C. Zhang, "Partial discharge features for power electronic transformers under high-frequency pulse voltage," *IEEE Trans. Plasma Sci.*, vol. 49, no. 2, pp. 845–853, Feb. 2021.

[223] X. Li, G. Wu, Y. Yang, Z. Wang, P. Xu, Y. Li, and B. Gao, "Partial discharge characteristics of oil-paper insulation for on-board traction transformers under superposed inter-harmonic AC voltages," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 27, no. 1, pp. 240–248, Feb. 2020.

[224] H. You, Z. Wei, B. Hu, Z. Zhao, R. Na, and J. Wang, "Partial discharge behaviors in power modules under square pulses with ultrafast dv/dt," *IEEE Trans. Power Electron.*, vol. 36, no. 3, pp. 2611–2620, Mar. 2021.

[225] J. Jiang, J. Chen, J. Li, X. Yang, R. Albarracín-Sánchez, P. Ranjan, and C. Zhang, "Propagation and localisation of partial discharge in transformer bushing based on ultra-high frequency technique," *High Voltage*, vol. 6, no. 4, pp. 684–692, Aug. 2021.

[226] J. Granado, A. Torralba, and C. Álvarez-Arroyo, "Effects of dispersion and multi-path propagation in partial discharges location," *IEEE Access*, vol. 8, pp. 219062–219070, 2020.

[227] P. Wang, S. Ma, S. Akram, P. Meng, J. Castellon, Z. Li, and G. C. Montanari, "Design of an effective antenna for partial discharge detection in insulation systems of inverter-fed motors," *IEEE Trans. Ind. Electron.*, vol. 69, no. 12, pp. 13727–13735, Dec. 2022.

[228] G. Zhang, J. Tian, X. Zhang, J. Liu, and C. Lu, "A flexible planarized biconical antenna for partial discharge detection in gas-insulated switchgear," *IEEE Antennas Wireless Propag. Lett.*, vol. 21, pp. 2432–2436, 2022.

[229] J. Cai, W. Yuan, X. Zhang, Y. Yang, J. Wu, and J. Li, "Ultra-high frequency self-sensing detection method for transformer partial discharge," *High Voltage Eng.*, vol. 47, no. 6, pp. 2041–2050, 2021.

[230] G. Zhang, C. Lu, H. Zhou, H. Tian, X. Yu, and X. Zhang, "Integrated ultrasonic and UHF sensing technology for partial discharge of power equipment," *High Voltage Eng.*, vol. 48, no. 12, pp. 5090–5101, 2022.

[231] P. Wang, H. Xiong, and H. He, "Bearing fault diagnosis under various conditions using an incremental learning-based multi-task shared classifier," *Knowl.-Based Syst.*, vol. 266, Apr. 2023, Art. no. 110395.
[232] K. Yin, Y. Wang, S. Liu, P. Li, Y. Xue, B. Li, and K. Dai, "GIS partial discharge pattern recognition based on multi-feature information fusion of PRPD image," *Symmetry*, vol. 14, no. 11, p. 2464, Nov. 2022.

**JIACHUAN LONG** (Member, IEEE) received the Ph.D. degree from Wuhan University, Wuhan, China, in 2017. He is currently an Associate Professor (academic backbone) with the School of Electronics and Information Engineering, Wuhan Donghu University, Wuhan. He is also an Expert of the Department of Science and Technology of Hubei Province, Wuhan Science and Technology Bureau. He has published more than 20 journal articles, including in IEEE TRANSACTIONS ON DIELECTRICS AND ELECTRICAL INSULATION, *IET Science, Measurement & Technology*, and Proceedings of the CSEE. His research interests include PD detection and diagnostic technology of high voltage equipment, signal processing, and pattern recognition.

He is a member of Hubei Provincial Institute of Communications and the Hubei Institute of Mechanical and Electrical Engineering. He was a recipient of the 2019 Wuhan Yellow Crane Talents Program and the 2022 Domestic Visiting Scholar Program for Young Backbone Teachers of Higher Education Institutions in Central and Western China. He is also a reviewer of many academic journals.

**LIJUAN XIE** received the M.S. degree from Xiangtan University, Xiangtan, China, in 2013. She is currently a Lecturer with the School of Electronics and Information Engineering, Wuhan Donghu University, Wuhan, China. Her research interests include intelligent signal processing and big data fault diagnosis technologies. She has participated in some scientific research projects, including the Natural Science Foundation of Hubei Province and Research Project of Hubei Provincial Department of Education.

**XIANPEI WANG** received the B.S. degree from North China Electric Power University, in 1984, and the M.S. and Ph.D. degrees from Wuhan University, Wuhan, China, in 1991 and 1999, respectively. Currently, he is a Professor with the Electronic Information School, Wuhan University. He is the author/coauthor of over 200 articles in international and domestic journals. His research interests include fault diagnosis of power equipment, intelligent monitoring technique for power systems, security of cyber physical power systems, and system reliability analysis.

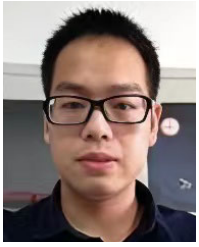
He won the Second Prize of 2014 Guangxi Science and Technology Progress Award.

**JUN ZHANG** (Senior Member, IEEE) received the M.S. degree from Wuhan University, Wuhan, China, in 2005. Since 2009, he has been the Laboratory Director of the Metrology Department, China Electric Power Research Institute. He has more than 14 years of experience in research and development of high-voltage measurement and testing technologies and related standardization works. He has published more than 20 articles and two monographs, authorized more than 20 patents, and presided or participated in the drafting of more than 20 national, industry, and enterprise standards.

He is a member of the IEEE PES China PSIM Committee, the National Electromagnetic Measurement Technology Committee High Voltage Measurement Subcommittee, and the National Electrical Instrumentation Standardization Technical Committee Secretary of the Instrumentation Professional Working Group.

**BING LU** (Member, IEEE) received the M.S. degree from Huazhong University of Science and Technology, Wuhan, China, in 2012. He is currently with China Electric Power Research Institute. He has published more than ten journal articles, including in the *High Voltage Apparatus* and *Wireless Communications and Mobile Computing*. His research interests include the calibration and testing technologies for metering equipment in power systems.

**CHUN WEI** received the M.S. degree from Wuhan University, Wuhan, China, in 2009. She is currently the Associate Dean of the School of Electronics and Information Engineering, Wuhan Donghu University, Wuhan. She has published more than ten journal articles and directed a number of scientific research projects. Her research interests include intelligent signal processing, big data processing, and other electronic technologies.



**DANGDANG DAI** received the Ph.D. degree from Wuhan University, Wuhan, China, in 2017. He is currently a Senior Engineer with State Grid Information & Communication Branch of Hubei Electric Power Company Ltd., Wuhan. He has published more than 15 journal articles, including in the *IET Science, Measurement & Technology* and *Spectroscopy and Spectral Analysis*. His research interests include PD detection and diagnostic technology of power equipment, digital signal processing, and cloud computing.

**GUOWEI ZHU** received the Ph.D. degree from Wuhan University, Wuhan, China, in 2017. He is currently a Senior Engineer with State Grid Information & Communication Branch of Hubei Electric Power Company Ltd., Wuhan. He has published more than 15 journal articles, including in IEEE TRANSACTIONS ON DIELECTRICS AND ELECTRICAL INSULATION and *Modern Physics Letters B*. His research interests include PD detection, network and information security technology, and vulnerability assessment of cyber physical systems.

**MENG TIAN** (Member, IEEE) received the B.S. and Ph.D. degrees from the Electronic Information School, Wuhan University, Wuhan, China, in 2011 and 2016, respectively. He was a Visiting Scholar with Southern Methodist University, in 2017 and 2018. He has presided more than ten scientific research projects, including the National Natural Science Foundation of China and Key Research and Development Program of Hubei Province. He has published more than 30 journal articles, including in IEEE TRANSACTIONS ON SMART GRID, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and *ISA Transactions*. His research interests include condition monitoring of power equipment, security of cyber physical power systems, cascading failures of multilayer networks, and image processing.